# SYSTEMS BIOLOGY AND OMICS APPROACHES TO UNDERSTAND COMPLEX DISEASES BIOLOGY

EDITED BY: Amit Kumar Yadav, Sanjay Kumar Banerjee, Kumardeep Chaudhary and Bhabatosh Das

PUBLISHED IN: Frontiers in Genetics, Frontiers in Neuroscience and Frontiers in Physiology

frontiers Research Topics

# frontiers



ISSN 1664-8714
ISBN 978-2-88976-078-7
DOI 10.3389/978-2-88976-078-7

## 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

# SYSTEMS BIOLOGY AND OMICS APPROACHES TO UNDERSTAND COMPLEX DISEASES BIOLOGY

Topic Editors:
**Amit Kumar Yadav,** Translational Health Science and Technology Institute (THSTI), India
**Sanjay Kumar Banerjee,** National Institute of Pharmaceutical Education and Research (Guwahati), India
**Kumardeep Chaudhary,** Council of Scientific and Industrial Research (CSIR), India
**Bhabatosh Das,** Translational Health Science and Technology Institute (THSTI), India

**Citation:** Yadav, A. K., Banerjee, S. K., Chaudhary, K., Das, B., eds. (2022). Systems Biology and Omics Approaches to Understand Complex Diseases Biology. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-88976-078-7

# Table of Contents

frontiers | Frontiers in Genetics

EDITORIAL
published: 12 April 2022
doi: 10.3389/fgene.2022.896818

# Editorial: Systems Biology and Omics Approaches for Understanding Complex Disease Biology

*Amit Kumar Yadav*[1]*, *Sanjay Kumar Banerjee*[2], *Bhabatosh Das*[1] and *Kumardeep Chaudhary*[3†]

[1]Translational Health Science and Technology Institute, NCR Biotech Science Cluster, Faridabad, India, [2]Department of Biotechnology, National Institute of Pharmaceutical Education and Research (NIPER), Guwahati, India, [3]Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, United States

**Keywords: multi-omics, systems biology, transcriptomics, proteomics, metabolomics, network biology, disease biology, machine learning**

## Editorial on the Research Topic

### Systems Biology and Omics Approaches to Understand Complex Disease Biology

High-throughput omics technologies have seamlessly galvanized the fields of big data and systems biology (Karczewski and Snyder, 2018). The amalgamation of omics techniques (genomics, transcriptomics, proteomics, metabolomics, and lipidomics etc.) and computational methods have enhanced our understanding of diseases in exquisite molecular detail (Adela et al., 2019; Aggarwal et al., 2020). Since computational methods help to unlock the potential of big-data (Shilo et al., 2020; Subramanian et al., 2020; Tolani et al., 2021), we solicited articles that applied systems biology approaches to complex diseases. The hosted topic received an excellent response and 13 manuscripts were accepted after careful editing.

Few studies harnessed the power of publicly available transcriptomic datasets. Khatri et al. studied 19 transcriptomics datasets to understand pancreatic ductal adenocarcinoma (PDAC). They constructed a support vector machine (SVM) classification model to predict a 9-gene biomarker panel of secretory proteins capable of predicting disease outcomes and patient risk stratification. Kowalski et al. evaluated the expression of epigenetics-related genes after valproic acid, carbamazepine, or phenytoin exposure in fetal development. Using weighted gene co-expression network analysis (WGCNA) on transcriptome data, they identified genes that correlated with Fetal Valproate Syndrome, and Fetal Hydantoin Syndrome.

Some studies applied proteomics or metabolomics analysis to study complex diseases. Using quantitative proteomics (iTRAQ), Das et al. studied the slow recovery in zebrafish males compared to females, following hypoxic-ischemic insult. The analysis exposed a sex-based difference in the neuronal cell recovery, with the increased levels of H3K9me3 in males confirmed through ChIP-qPCR. This can be used to develop novel targets for gender-specific therapeutic strategy. Another proteomics study by Zhou et al. used tandem mass tag (TMT) proteomics to understand the connection between "Heart failure with preserved ejection fraction" (HFpEF) and hypertension (HTN). The functional and network analysis revealed seven common differentially expressed proteins in HFpEF and HTN, for which molecular docking studies were performed to identify therapeutic targets. Sardar et al. integrated proteomics and clinical data to identify biomarkers of COVID-19 progression using artificial intelligence. Using feature selection and cross-validation on normalized protein expression data, a LogitBoost model was developed. They also identified 18 potential proteins for drug repurposing, when understanding of COVID-19 disease was in its early stages (Chatterjee et al., 2020). The prominent clinical abnormalities also included cardiovascular functions (Shen et al., 2020), which was also studied recently (Rizvi et al., 2022). The authors

OPEN ACCESS

***Correspondence:**
*Amit Kumar Yadav*
*amit.yadav@thsti.res.in*

***†Present addresss:***
*Kumardeep Chaudhary, CSIR-Institute of Genomics and Integrative Biology, New Delhi, India*

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

***Received:*** *15 March 2022*
***Accepted:*** *28 March 2022*
***Published:*** *12 April 2022*

***Citation:***
*Yadav AK, Banerjee SK, Das B and Chaudhary K (2022) Editorial: Systems Biology and Omics Approaches for Understanding Complex Disease Biology. Front. Genet. 13:896818. doi: 10.3389/fgene.2022.896818*

developed CovidPrognosis webserver for predicting patient survival, for assisting in rapid patient triaging. Wang et al. studied the serum metabolome of 99 patients with acute ischemic stroke to identify biomarkers and its heterogeneous subtypes. Using PCA and OPLS-DA analysis, the authors identified 18 metabolites including oleic acid, linoleic acid, arachidonic acid, which could differentiate between stroke patients and healthy individuals. The authors also identified differences in ischemic stroke subtypes to explore pathophysiological mechanisms.

Integrating multiple omics data was another popular theme for some articles. Zhao et al. presented an interesting study on the pathogenesis of polycystic ovary syndrome (PCOS)- the most common, endocrine and metabolic disease in women of reproductive age. Limited studies have been performed with multiomics analyses of granulosa cells (GCs) considering epigenetics as a regulatory factor. The authors systematically investigated the differences in the mRNA-miRNA-lncRNA transcriptome and genome-wide DNA methylation modification profiles and their regulatory networks. The data revealed that all differentially expressed genes were associated with steroid biosynthesis and glycolysis/gluconeogenesis pathways. Diabetic retinopathy (DR) requires early diagnostic markers and effective treatment strategies. Another interesting integrative computational approach by Kumari et al. was devised to capture differentially expressed genes that were also the targets of miRNA, as well as depicted atypical methylation patterns. The authors identified hub genes and network modules from the PPIs of the early and late disease genes. They also identified the pathways related to oxidoreductase activity, extracellular matrix binding, immune response, cell adhesion, PI3K-Akt signaling pathway, ECM receptor interaction and leukocyte migration. They reported 7 hub genes and 9 early genes as potential candidates for prognostic, diagnostic, or therapeutic application. A fascinating approach to integrate metabolism with the regulatory-metabolic network using transcriptomics data was demonstrated by Sun et al. to understand tumour heterogeneity in hepatocellular carcinoma (HCC). The authors studied disease perturbation in regulation and metabolism using unified mechanistic modeling approach, which used transcriptomics data with regulatory-metabolic network model to understand HCC stratification. They identified transcription factors and target genes impacting tumorigenesis and integrated this information with constraint-based models identifying five important genes associated with cancer growth. Non-negative matrix factorization was used for stratification of differential genes from TCGA samples to understand HCC pathways and find potential targets. In another excellent multiomics approach, the pathogenesis of carotid atherosclerosis (CAS) (a cause of stroke) was studied by Ji et al. with respect to the interactions between gut microbiome and metabolome. Authors attempted an integrated analysis of the transcriptome (from GEO) with in-house generated metabolome and microbiome data for in-depth understanding of the "microbiota–metabolite–gene" axis in the pathogenesis of CAS. Interestingly, the study identified α-N-Phenylacetyl-L-glutamine as an increased metabolite in CAS patients. *FABP4* was the most upregulated gene and was positively associated with *Acidaminococcus*, an anaerobic bacteria living in the human gut. The authors integrated and overlaid different omics data to understand CAS pathogenesis. However, the study could have benefitted more from generating transcriptome data from the same patients as the microbiome and metabolome data.

Biswas et al. developed a sophisticated analysis platform-ADOPHIN, to allow the analysis of pan-omics data in context of the Human interactome. They developed a meta-interactome network with protein-protein interactions (PPIs), regulatory interactions between miRNAs and their respective target genes, transcription factors and their targets. The authors discovered topologically important nodes (TINs) with regulatory networks between various biomolecules (proteins, transcription factors, or miRNAs), linked to signaling and metabolic pathways. The genes, proteins or miRNA from multi-omics data are mapped onto the compiled interactome to capture the biological context-specific interactions, as demonstrated by authors in cervical, breast and ovarian cancers. Such meta-interactome mining approaches with cross-pathway links and connectivity analysis, provide a user-friendly method to explore multi-omics data.

Though excellent studies in their own right, some of the studies may require, and even benefit from independent validation. Furthermore, the power of such integrated analysis can increase with more data types, beyond methylation and transcription/gene expression (Hasin et al., 2017; Yan et al., 2018). This can help in triaging more functional interconnections and discovery of relevant candidates for further research.

Apart from exceptional articles, the topic also had two excellent reviews. The review by Zamith-Miranda et al. appraises the biogenesis, composition and functions of fungal extracellular vesicles using multi-omics studies. Shedding of extracellular vesicles is a conserved process across all three kingdoms of life. The mechanisms and sites of fungal extracellular vesicle formation, their nucleic acid content and importance in virulence, pathogenicity and antimicrobial resistance are discussed. The review concludes with the current knowledge gaps in the extracellular vesicles biology and their future. An excellent review on the current landscape of Single-Cell Transcriptomics (scRNA-seq) data acquisition and bioinformatics analysis is presented by Adil et al. scRNA-seq has emerged as an instrumental technique to decipher cellular heterogeneity in complex diseases. However, the volume, granularity and sparsity of data poses outstanding challenges-in data generation and downstream analysis. An overview of scRNA-seq profiling techniques and biophysical cell-isolation methods is followed by the widely-used tools for sequencing read-alignment and mRNA expression quantification. The bottlenecks and current software solutions reviewing the methods for normalization, batch-effect removal, imputation, dimensionality reduction, subtype/cluster identification are also covered. Finally, the review discusses the multiple evolving strategies to integrate multi-omics datasets at the single-cell level.

The topic has covered multiple omics with different computational methods, network analysis and modeling to study a diverse array of biological problems in complex diseases. We hope this interesting assortment of article collection invigorates the readers towards novel applications of multiomics for deeper dissection of disease biology.

## AUTHOR CONTRIBUTIONS

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

## FUNDING

AKY is supported by DBT-Big Data Initiative grant (BT/PR16456/BID/7/624/2016) and Translational Research Program (TRP) at THSTI funded by DBT. AKY and BD are also supported by THSTI intramural grant (2021-2023). SKB is supported by ICMR grant (No: 5/7/1747/CH/Adhoc/RBMCH-2021).

## ACKNOWLEDGMENTS

The guest editors acknowledge the peer reviewers who made this special issue possible with their timely, insightful, and critical comments for improving the manuscripts in this collection.

## REFERENCES

Adela, R., Reddy, P. N. C., Ghosh, T. S., Aggarwal, S., Yadav, A. K., Das, B., et al. (2019). Serum Protein Signature of Coronary Artery Disease in Type 2 Diabetes Mellitus. *J. Transl Med.* 17, 17. doi:10.1186/s12967-018-1755-5

Aggarwal, S., Banerjee, S. K., Talukdar, N. C., and Yadav, A. K. (2020). Post-translational Modification Crosstalk and Hotspots in Sirtuin Interactors Implicated in Cardiovascular Diseases. *Front. Genet.* 11, 356. doi:10.3389/fgene.2020.00356

Chatterjee, P., Nagi, N., Agarwal, A., Das, B., Banerjee, S., Sarkar, S., et al. (2020). The 2019 Novel Coronavirus Disease (COVID-19) Pandemic: A Review of the Current Evidence. *Indian J. Med. Res.* 151, 147–159. doi:10.4103/ijmr.ijmr_519_20

Hasin, Y., Seldin, M., and Lusis, A. (2017). Multi-omics Approaches to Disease. *Genome Biol.* 18, 83. doi:10.1186/s13059-017-1215-1

Karczewski, K. J., and Snyder, M. P. (2018). Integrative Omics for Health and Disease. *Nat. Rev. Genet.* 19, 299–310. doi:10.1038/nrg.2018.4

Rizvi, Z. A., Dalal, R., Sadhu, S., Binayke, A., Dandotiya, J., Kumar, Y., et al. (2022). Golden Syrian Hamster as a Model to Study Cardiovascular Complications Associated with SARS-CoV-2 Infection. *Elife* 11, e73522. doi:10.7554/eLife.73522

Shen, B., Yi, X., Sun, Y., Bi, X., Du, J., Zhang, C., et al. (2020). Proteomic and Metabolomic Characterization of COVID-19 Patient Sera. *Cell* 182, 59–72. doi:10.1016/j.cell.2020.05.032

Shilo, S., Rossman, H., and Segal, E. (2020). Axes of a Revolution: Challenges and Promises of Big Data in Healthcare. *Nat. Med.* 26, 29–38. doi:10.1038/s41591-019-0727-5

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

Tolani, P., Gupta, S., Yadav, K., Aggarwal, S., and Yadav, A. K. (2021). Big Data, Integrative Omics and Network Biology. *Adv. Protein Chem. Struct. Biol.* 127, 127–160. doi:10.1016/bs.apcsb.2021.03.006

Yan, J., Risacher, S. L., Shen, L., and Saykin, A. J. (2018). Network Approaches to Systems Biology Analysis of Complex Disease: Integrative Methods for Multi-Omics Data. *Brief Bioinform* 19, 1370–1381. doi:10.1093/bib/bbx066

**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: 10 September 2020
doi: 10.3389/fgene.2020.572284

# A Transcriptomics-Based Meta-Analysis Combined With Machine Learning Identifies a Secretory Biomarker Panel for Diagnosis of Pancreatic Adenocarcinoma

*Indu Khatri*[1,2] *and Manoj K. Bhasin*[1,3]*

[1] Division of IMBIO, Department of Medicine, Beth Israel Lahey Health, Harvard Medical School, Boston, MA, United States, [2] Department of Immunology and Leiden Computational Biology Center, Leiden University Medical Center, Leiden, Netherlands, [3] Department of Pediatrics and Biomedical Informatics, Children's Healthcare of Atlanta, Emory School of Medicine, Atlanta, GA, United States

OPEN ACCESS

**Edited by:**
*Amit Kumar Yadav, Translational Health Science and Technology Institute (THSTI), India*

**Reviewed by:**
*Deepak Sharma, Indian Institute of Technology Roorkee, India*
*Oksana Sorokina, The University of Edinburgh, United Kingdom*

***Correspondence:**
*Manoj K. Bhasin*
*manoj.bhasin@emory.edu*

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

**Received:** *13 June 2020*
**Accepted:** *12 August 2020*
**Published:** *10 September 2020*

**Citation:**
*Khatri I and Bhasin MK (2020) A Transcriptomics-Based Meta-Analysis Combined With Machine Learning Identifies a Secretory Biomarker Panel for Diagnosis of Pancreatic Adenocarcinoma. Front. Genet. 11:572284. doi: 10.3389/fgene.2020.572284*

Pancreatic ductal adenocarcinoma (PDAC) is generally incurable due to the late diagnosis and absence of markers that are concordant with expression in several sample sources (i.e., tissue, blood, plasma) and platforms (i.e., Microarray, sequencing). We optimized meta-analysis of 19 PDAC (tissue and blood) transcriptome studies from multiple platforms. The key biomarkers for PDAC diagnosis with secretory potential were identified and validated in different cohorts. Machine learning approach i.e., support vector machine supported by leave-one-out cross-validation was used to build and test the classifier. We identified a 9-gene panel (IFI27, ITGB5, CTSD, EFNA4, GGH, PLBD1, HTATIP2, IL1R2, CTSA) that achieved ∼0.92 average sensitivity and ∼0.90 average specificity in distinguishing PDAC from healthy samples in five training sets using cross-validation. These markers were also validated in proteomics and single-cell transcriptomics studies suggesting their prognostic role in the diagnosis of PDAC. Our 9-gene classifier can not only clearly discriminate between better and poor survivors but can also precisely discriminate PDAC from chronic pancreatitis (AUC = 0.95), early stages of progression [Stage I and II (AUC = 0.82), IPMA and IPMN (AUC = 1), and IPMC (AUC = 0.81)]. The 9-gene marker outperformed the previously known markers in blood studies particularly (AUC = 0.84). The discrimination of PDAC from early precursor lesions in non-malignant tissue (AUC > 0.81) and peripheral blood (AUC > 0.80) may assist in an early diagnosis of PDAC in blood samples and thus will also facilitate risk stratification upon validation in clinical trials.

**Keywords: biomarker, pancreatic cancer, secretory, transcriptome, validation**

**Abbreviations:** AUC, area under the curve; CA 19-9, carbohydrate antigen 19-9; CDF, chip definition file; CP, chronic pancreatitis; DE, differentially expressed; GEO, gene expression omnibus; GGH, γ-glutamyl hydrolase; FDR, false discovery rate; HPA, human protein atlas; IPMA, intraductal papillary-mucinous adenoma; IPMC, intraductal papillary-mucinous carcinoma; IPMN, intraductal papillary mucinous neoplasm; LOOCV, leave-one-out cross-validation; noTM, no transmembrane segments; PanIN, pancreatic intraepithelial neoplasia; PC, pancreatic cancer; PDAC, pancreatic ductal adenocarcinoma; ROC, receiver operating characteristic; SVM, support vector machines; TCGA, tissue cancer genome atlas.

## INTRODUCTION

Pancreatic ductal adenocarcinoma (PDAC) is the most common type of pancreatic cancer (PC), which is one of the fatal cancers in the world with 5-year survival rate of <5% due to the lack of early diagnosis (Fesinmeyer et al., 2005). One of the challenges associated with an early diagnosis is distinguishing PDAC from other non-malignant benign gastrointestinal diseases such as chronic pancreatitis (CP) due to the histopathological and imaging limitations (Brand and Matamoros, 1998). Although imaging techniques such as endoscopic ultrasound and FDG-PET have improved the sensitivity of PDAC detection but have failed to distinguish PC from focal mass-forming pancreatitis in >50% cases. Dismal prognosis of PC yields from asymptomatic early stages, speedy metastatic progression, lack of effective treatment protocols, early loco regional recurrence, and absence of clinically useful biomarker(s) that can detect PC in its precursor form(s) (Ballehaninna and Chamberlain, 2012). Studies have indicated a promising 70% 5-year survival for cases where incidental detections happened for stage I pancreatic tumors that were still confined to pancreas (Frena, 2001; Schneider and Schulze, 2003). Therefore, it only seems rational to aggressively screen for early detection of PDAC. CA19-9 is the most common and the only FDA approved blood-based biomarker for diagnosis, prognosis, and management of PC but it has several limitations such as poor specificity, lack of expression in the Lewis negative phenotype, and higher false positive elevation in the presence of obstructive jaundice (Ballehaninna and Chamberlain, 2012). A large number of carbohydrate antigens, cytokeratin, glycoprotein, and Mucinic markers and hepatocarcinoma–intestine–pancreas protein, and PC-associated protein markers have been discovered as a putative biomarkers for management of PC (Ballehaninna and Chamberlain, 2013). However, none of these have demonstrated superiority to CA19-9 in the validation cohorts. Previously, our group discovered a novel five-genes-based tissue biomarker for the diagnosis of PDAC using innovative meta-analysis approach on multiple transcriptome studies. This biomarker panel could distinguish PDAC from healthy controls with 94% sensitivity and 89% specificity and was also able to distinguish PDAC from CP, other cancers, and non-tumor from PDAC precursors at tissue level (Bhasin et al., 2016). The relevance of tissue-based diagnostic markers remains unclear owing to the limitations of obtaining biopsy samples. Additionally, most current studies are based on small sample sizes with limited power to identify robust biomarkers. Provided the erratic nature of PC, the major unmet requirement is to have reliable blood-based biomarkers for early diagnosis of PDAC.

The crucial requisite for better PDAC diagnosis has driven a large number of genome-level studies defining the molecular landscape of PDAC to identify early diagnosis biomarkers and potential therapeutic targets. Despite many genomics studies, we do not have a reliable biomarker that is able to surpass the sensitivity and specificity of CA19-9. The independent studies suffer from inherent statistical limitations where the datasets derived from different batches, techniques and platforms and analytic methods result in the lack of concordance (Ramasamy et al., 2008). The published gene signatures of individual microarray studies are not concordant with comparative analysis and meta-analysis studies when standard approaches are used due to variability in analytical strategies (Ramasamy et al., 2008).

In our work, we have included all the available gene expression datasets for PDAC versus healthy subjects from GEO[1] and ArrayExpress database[2] measured via microarray or sequencing platforms. We have included the datasets derived from blood and tissue sources excluding cell lines in our analysis, which was not included previously. The cell lines were excluded for they do not depict normal cell morphology and do not maintain markers and functions seen *in vivo*.

The approach of combining multiple studies has previously been stated to reveal biological insight by increasing the reproducibility and sensitivity which is generally not evident in the independent original datasets (Wang et al., 2004). Using the uniform pre-processing, normalization and batch correction approaches in the meta-analysis can assist in eliminating false-positive results. Therefore, we used multiple datasets in combinations and further divided them in training, testing and validation sets to identify and validate the markers with secretory signal peptides. We hypothesize that proteins with secretory potential will be secreted out of the tissue into the blood and these markers can be used as prognostic markers in a non-invasive manner. There were no previous studies on identification of marker genes that could be used with least-invasive methods. Also, a set of multiple genes targeting different pathways and biological processes are more reliable and sensitive than single gene-based marker for complex diseases like cancer (Ramasamy et al., 2008). We also corroborated the protein expression of our markers in proteomics datasets obtained from human protein atlas (HPA)[3].

## MATERIALS AND METHODS

### Dataset Identification

The publicly available microarray repositories i.e., ArrayExpress (see text footnote 2) and GEO (see text footnote 1) were searched for gene expression studies of human pancreatic specimens. The selected datasets were divided into five training sets and fourteen independent validation sets for initial development and validation of biomarkers. To avoid the representation of the datasets only from tissues the few blood studies available were divided across all training and validation phase of this study.

Each training dataset (GSE18670, E-MEXP-950, GSE32676, GSE74629, and GSE49641) included a minimum of four samples of normal pancreas and a minimum of four samples of PDAC. In training set we included minimum two datasets with source pancreatic tissue and peripheral blood. This was done to identify a predictor based on genes that are detectable in both pancreatic tissue and blood. Datasets GSE18670 (Set1: 6 normal, 5 PDAC), GSE32676 (Set6: 6 normal, 24 PDAC) and E-MEXP-950 (Set3: 10 normal, 12 PDAC) was derived from pancreatic tissue, whereas

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

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

[3]https://www.proteinatlas.org/

GSE74629 (Set4: 14 normal, 32 PDAC) and GSE49641 (Set5: 18 normal, 18 PDAC) contain transcriptome profile of peripheral blood PDAC patients.

Further, 14 validation sets were also divided into three groups, one "Test sets" (**Table 1A**); second "Validation Sets" (**Table 1A**) and third "Prospective Validation Sets" (**Table 1B**). Five Tissue studies were included: one from microdissected tissue samples (Set6: 6 normal, 6 PDAC) and four from whole tissues (Set7: 45 normal, 40 PDAC; Set8: 6 normal, 6 PDAC; Set9: 8 normal and 12 PDAC and Set10: 15 normal, 33 PDAC). One blood study from peripheral blood was also validated using the biomarker (E-Set11: 14 normal, 12 PDAC).

For Phase I Validation we selected five datasets from different platforms from whole tissues and blood platelets, including comparison of normal versus PDAC samples similar to training and test sets. Four whole tissue datasets (V1: 61 normal, 69 PDAC; V2: 20 normal, 36 PDAC; V3: 9 normal, 45 PDAC; and V4: 12 normal, 118 tumor) and one dataset from blood with samples from blood platelets (V5: 50 normal, 33 PDAC) were included.

In Prospective Validation, the performance of the developed PDAC biomarker panel was tested on four additional independent datasets i.e.,: (i) PDAC versus normal (pancreatic) tissue from TCGA database (PV1: 4 normal, 150 PDAC), (ii) PDAC versus normal pancreatic tissues in early stages [PV2: 61 normal, 69 PDAC (Stage I and II)], (iii) PDAC versus CP (PV3: 9 pancreatitis, 9 PDAC), and (iv) PDAC precursor lesions (IPMA, IPMC, and IPMN) with associated invasive carcinoma [PV4: 6 normal, 15 PDAC precursors (5 IPMA, 5 IPMC, 5 IPMN)] versus normal pancreas tissues (**Table 1B**). Three datasets utilized oligonucleotide- based microarray platforms (two versions of Affymetrix GeneChips and Gene St 1.0 microarrays in one dataset) whereas the cancer genome atlas (TCGA) data is the sequencing data obtained using RNA-sequencing technology.

## Quality Control and Outlier Analysis

Stringent quality control and outlier analysis was performed on all datasets used for training and validation to remove low quality arrays from the analysis. The technical quality of arrays was determined on the basis of background values, percent present calls and scaling factors using various bioconductor packages (Wilson and Miller, 2005; Kauffmann et al., 2009). The arrays with high quality were subjected to outlier analysis using array intensity distribution, principal component analysis, array-to-array correlation and unsupervised clustering. The samples that were identified to be of low quality or identified as outliers were eliminated from the analysis.

## Mapping of Platform Specific Identifiers to Universal Identifier

To facilitate the collation of the differentially expressed (DE) genes identified by analysis of individual datasets, the platform specific identifiers associated with each dataset were annotated to corresponding universal gene symbol identifiers. Gene symbols were used in subsequent analyses including comparative analysis of different datasets as well as predictor development. Briefly Affymetrix data was annotated using the custom CDF from brainarray[4]. Affymetrix probe set IDs that could not be mapped to an Entrez gene identifiers were removed from the gene lists. For Agilent- 028004, HumanHT-12 V4.0 and Gene St 1.0 studies the raw matrix was directly retrieved from the GEO interactive web tool, GEO2R[5], which were further processed and normalized. The normalized and annotated genes for TCGA was obtained from Broad GDAC Firehose database[6]. We have removed 29 non-PDAC samples from TCGA during validation as our classifier was trained using PDAC samples (Peran et al., 2018).

## Pre-processing and Normalization of Microarray Datasets

Potential bias introduced by the range of methodologies used in the original microarray studies, including various experimental platforms and analytic methods, was controlled by applying a uniform normalization, preprocessing and statistical analysis strategy to each dataset. Raw microarray dataset were normalized using vooma (Law, 2013) algorithm which estimates the mean-variance relationship and use the relationship to compute appropriate gene expression level weights. Similarly, RNA-sequencing datasets were normalized using voom algorithm (Law et al., 2014). The normalized datasets were used for performing meta-analysis as well as predictor development.

## Differential Gene Expression Analysis for Generating Meta-Signature

To generate PDAC meta-signature, we performed differential expression analysis on individual datasets from training sets by comparing normal versus cancer samples. To identify DE genes, a linear model was implemented using the linear model microarray analysis software package (LIMMA) (Ritchie et al., 2015). LIMMA estimates the differences between normal and cancer samples by fitting a linear model and using an empirical Bayes method to moderate standard errors of the estimated log-fold changes for expression values from each probe set. In LIMMA, all genes were ranked by t-statistics using a pooled variance, a technique particularly suited to small numbers of samples per phenotype. The DE probes were identified on the basis of absolute fold change and Benjamini and Hochberg corrected *P*-value (Benjamini and Hochberg, 1995). The genes with multiple test corrected *P*-value $< 0.05$ were considered as DE. Comparative analyses were performed to identify those genes that are significantly DE across multiple PDAC datasets. Genes that are concordantly over or under expressed in three PDAC datasets (two tissues and one blood study) were included in PDAC meta-signature.

## Secretory Gene Set Identification

To identify a non-invasive predictor based on genes with secretory potential, we selected genes that had signal peptide for secretory proteins with no transmembrane segments (noTM).

[4]http://brainarray.mbni.med.umich.edu

[5]www.ncbi.nlm.nih.gov/geo/geo2r/

[6]http://gdac.broadinstitute.org

The Biomart package in R (Durinck et al., 2005) with quering the gene symbols to SignalP database facilitated the analysis. The Ensembl Biomart database enables users to retrieve a vast diversity of annotation data for specific organisms. After loading the library, one can connect to either public BioMart databases (Ensembl, COSMIC, Uniprot, HGNC, Gramene, Wormbase and dbSNP mapped to Ensembl) or local installations of these. One set of functions can be used to annotate identifiers such as Affymetrix, RefSeq and Entrez-Gene, with information such as gene symbol, chromosomal coordinates, OMIM and Gene Ontology or vice-versa.

## Training and Independent Validation of PDAC Classifier Using Support Vector Machine

The upregulated secretory genes DE from PDAC meta-signature was used for training of PDAC classifier. Classifier was generated by implementing the random forest (RF) using caret{R} and support vector machines (SVM) approach using e1071{R}. Polynomial kernel was used to develop the classifier. RF and SVM was first tuned using 10-fold cross-validation at different costs and the best cost and gamma functions were later used to perform classification on testing and validation sets. Classifiers were trained using normalized, preprocessed gene expression values from each of the five training datasets independently. To independently validate our model in each dataset, performance of classifiers in the training sets was evaluated using internal LOOCV. We assessed the classifier of five to ten genes selected from the set of upregulated genes to identify the biomarker panel that works best in both tissue and blood-based studies. The complete sets of possible combinations of five to ten genes were drawn from the upregulated genes and the accuracy of each classifier was assessed. The performance of classifiers was measured using threshold-dependent (e.g., sensitivity, specificity, accuracy) and threshold-independent ROC analysis. In ROC analysis, the AUC provides a single measure of overall prediction accuracy. The biomarker panel with the highest performance in the training sets (both tissue and blood-based studies) was chosen for assessment of predictive power in six independent test datasets using threshold-dependent and -independent measures i.e., AUC. SVM outperformed the RF models in the training datasets.

## Survival Analysis

To determine the association of key genes with survival in PC, we performed survival analysis using the TCGA database[7]. The survival analysis was performed on PDAC mRNA of 150 patients [excluding samples related to normal tissues and non-PDAC tissues (Peran et al., 2018)]. Survival analysis was performed on the basis of individual mRNA expression using the Kaplan-Meier (K-M) approach (Kaplan and Meier, 1958). The normalized expression data for each gene was divided into high and low median groups. The survival analysis was

[7]https://cancergenome.nih.gov/

**TABLE 1A |** Datasets used for development and validation of secretory genes based PDAC classifier.

| Groups | Dataset | Normal | Tumor | Sample type | Platform | Accession |
|---|---|---|---|---|---|---|
| Training Sets | Set 1 | 6 | 5 | Enriched | U133 Plus 2.0 | E-GEOD-18670 |
| | Set 2 | 6 | 24 | Whole Tissue | U133 Plus 2.0 | E-GEOD-32676 |
| | Set 3 | 10 | 12 | Microdissected | U133A | E-MEXP-950 |
| | Set 4 | 14 | 32 | Peripheral Blood | HumanHT-12 V4.0 | GSE74629 |
| | Set 5 | 18 | 18 | Peripheral Blood | Gene St 1.0 | GSE49641 |
| Test sets | Set 6 | 6 | 6 | Microdissected | U133A | E-MEXP-1121 |
| | Set 7 | 45 | 40 | Whole Tissue | Gene St 1.0 | GSE28735 |
| | Set 8 | 6 | 6 | Whole Tissue | Gene St 1.0 | GSE41368 |
| | Set 9 | 8 | 12 | Whole Tissue | U133 Plus 2.0 | E-GEOD-71989 |
| | Set 10 | 15 | 33 | Whole Tissue | U133 Plus 2.0 | E-GEOD-16515 |
| | Set 11 | 14 | 12 | Peripheral Blood | U133 Plus 2.0 | E-GEOD-15932 |
| Validation Sets | V1 | 61 | 69 | Whole Tissue | Gene St 1.0 | E-GEOD-62452 |
| | V2 | 20 | 36 | Whole Tissue | U133 Plus 2.0 | E-GEOD-15471 |
| | V3 | 9 | 45 | Whole Tissue | Agilent-028004 | GSE60979 |
| | V4 | 12 | 118 | Whole Tissue | U219 | GSE62165 |
| | V5 | 50 | 33 | Blood Platelet | HiSeq-2500 | GSE68086 |

**TABLE 1B |** Datasets used for prospective validation of secretory genes based PDAC classifier.

| Group | Dataset | Group | Pancreatic tumor | Sample type | Platform | Accession |
|---|---|---|---|---|---|---|
| Prospective Validation Sets | PV1 | 4 Normal | 150 PDAC | Tissue | RNA-Seq | TCGA |
| | PV2 | 61 Normal | 69 PDAC (Stage I and II) | Whole Tissue | Gene St 1.0 | E-GEOD-62452 |
| | PV3 | 9 (Pancreatitis) | 9 (PDAC) | Whole Tissue | U95Av2 | E-EMBL-6 |
| | PV4 | 7 (Normal) | 15 (IPMA, IPMC, IPMN) | Microdissected | U133 Plus 2.0 | GSE19650 |

performed using K-M analysis from survival package in R. The results of the survival analysis were visualized using K-M survival curves with log rank testing. The results were considered significant if the *P*-values from the log rank test were below 0.05. The effects of mRNA on the event were calculated using univariate Cox proportional hazard model without any adjustments.

## Pathways Analysis

The biological pathways for the genes was performed using ToppFun software of ToppGene suite (Chen et al., 2009). ToppGene is a one-stop portal for gene list enrichment analysis and candidate gene prioritization based on functional annotations and protein interactions network. ToppFun detects functional enrichment of the provided gene list based on transcriptome, proteome, regulome (TFBS and miRNA), ontologies (GO, Pathway), phenotype (human disease and mouse phenotype), pharmacome (Drug-Gene associations), literature co-citation, and other features. The biological pathways with FDR $< 0.05$ were considered significantly affected.

# RESULTS

## PDAC Differential Expression Analysis and Meta-Signature Development

To develop a gene based minimally invasive biomarker for differentiating PDAC from normal/pancreatitis, we identified 19 microarray and RNA sequencing studies containing PDAC and normal samples. These datasets based on their origin i.e., blood or tissue were divided into training sets, independent test sets, validation sets and prospective validation sets (**Figure 1**; Overview of meta-analysis strategy). For classifier training, we performed meta-analysis on 3-tissue and 2-blood-based PDAC studies to identify meta-signature that are DE in blood and tissue during PC. To account for the differences in microarray/sequencing platform used in studies, we processed and normalized studies according to their platforms and selected the genes that are common across multiple studies. The number of DE secretory genes ranged from 480 to 810 genes, totaling 2,010 significantly DE genes in the five training datasets. We identified 74 genes (35 downregulated and 39 upregulated) with concordant directionality in at least two of the three tissue datasets and one of the two blood datasets (**Figure 2A**, shown in red color and **Supplementary Table S1**).

The 74 genes showed consistent expression across the PDAC samples in the selected five datasets (3 tissue and 2 blood datasets) as compared to the normal pancreas (**Figure 2B**), with the extent of over-expression or under-expression denoted by red or green shading, respectively. Pathway analysis of these 74 common PDAC genes depicted significant enrichment (*P*-value $< 0.05$) in multiple extracellular matrix-associated pathways (e.g., Ensemble of genes encoding extracellular matrix and extracellular matrix-associated proteins, remodeling of the extracellular matrix, structural ECM glycoproteins, Cell adhesion molecules) (**Supplementary Figure S1**). These pathways play important roles in the adhesion of cells that is a key process in progression of PDAC.

## Variables Selection and Class Prediction Analysis in Training Sets

The 39 upregulated genes from the 74 common genes were selected for predictor development. We have specifically targeted upregulated genes for their therapeutics and diagnostic applications. We plotted boxplots of these 39 genes across all the five training sets and removed the genes with opposite direction in any of these five sets. The 27 concordantly upregulated genes (**Supplementary Table S2**) were selected after the boxplot analysis. These combined gene set clearly discriminated between PDAC and normal pancreas samples in all the datasets of training set, as depicted in the heatmap for 27 genes (**Supplementary Figure S2A**) and principal component analysis (PCA) plots (**Supplementary Figure S2B**). The predictors based on 5 to 10 genes were developed and assessed by LOOCV implementing with a polynomial kernel based SVM classifier. All the possible combination of five to ten genes were tested from 27 upregulated genes. The classifiers containing the selected 9 genes i.e., IFI27, ITGB5, CTSD, EFNA4, GGH, PLBD1, HTATIP2, IL1R2, and CTSA performed with highest accuracy. These 9 genes were upregulated in PDAC as compared to the normal pancreas in all the five training sets (**Figures 2C,D**).

We performed LOOCV cross-validation analysis of the 9-gene PDAC classifier across the five training datasets to determine its predictive performance. For each of the five training datasets individually, sensitivity ranged from 0.83 to 1.0 and specificity 0.71 to 1.00 for the predictor (**Supplementary Figure S3A**, **Table 2**). Comparison of the 9-gene PDAC classifier performance in tissues (Set1-Set3) and blood datasets (Set 4 and Set 5) showed approximately 0.94 sensitivity and 0.97 specificity for the tissue datasets, in contrast to 0.88 sensitivity and 0.80 specificity for the blood datasets (**Supplementary Figure S3B**, **Table 2**). AUC for the three tissue datasets ranged from 0.89 to 1.00 with median = 0.96 (**Supplementary Figure S3B**) and for two blood datasets from 0.92 to 0.96 with median = 0.94 (**Table 2**, **Supplementary Figure S3C** and **Figure 2E**), demonstrated threshold independent performance). The average gene expression plots with all the samples combined from the five training sets (**Supplementary Figure S4A**) and the PCA plots of training sets (**Supplementary Figure S4B**) from 9 genes supported the discriminatory power of the marker combinations in identification of PDAC subjects from normal.

## Biological Significance of Selected Genes

CTSA and CTSD are involved in extracellular matrix associated proteins; IFI27 and IL1R2 in cytokine signaling in immune system; ITGB5 and HTATIP2 in apoptotic pathway and EFNA4, GGH and PLBD1 are involved in Ephrin signaling, fluoropyrimidine activity and glycerophospholipid biosynthesis, respectively. The genes selected based on the presence of signal

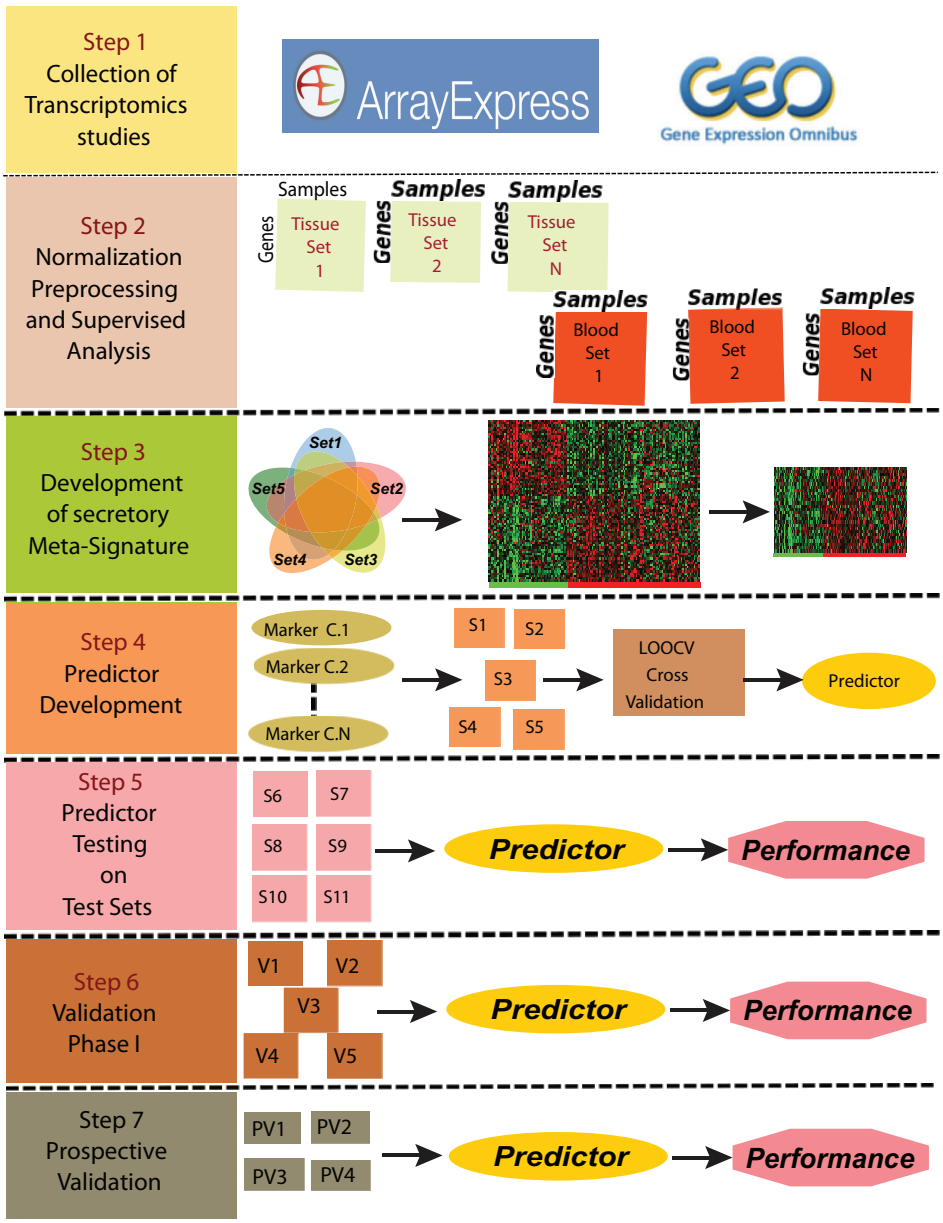


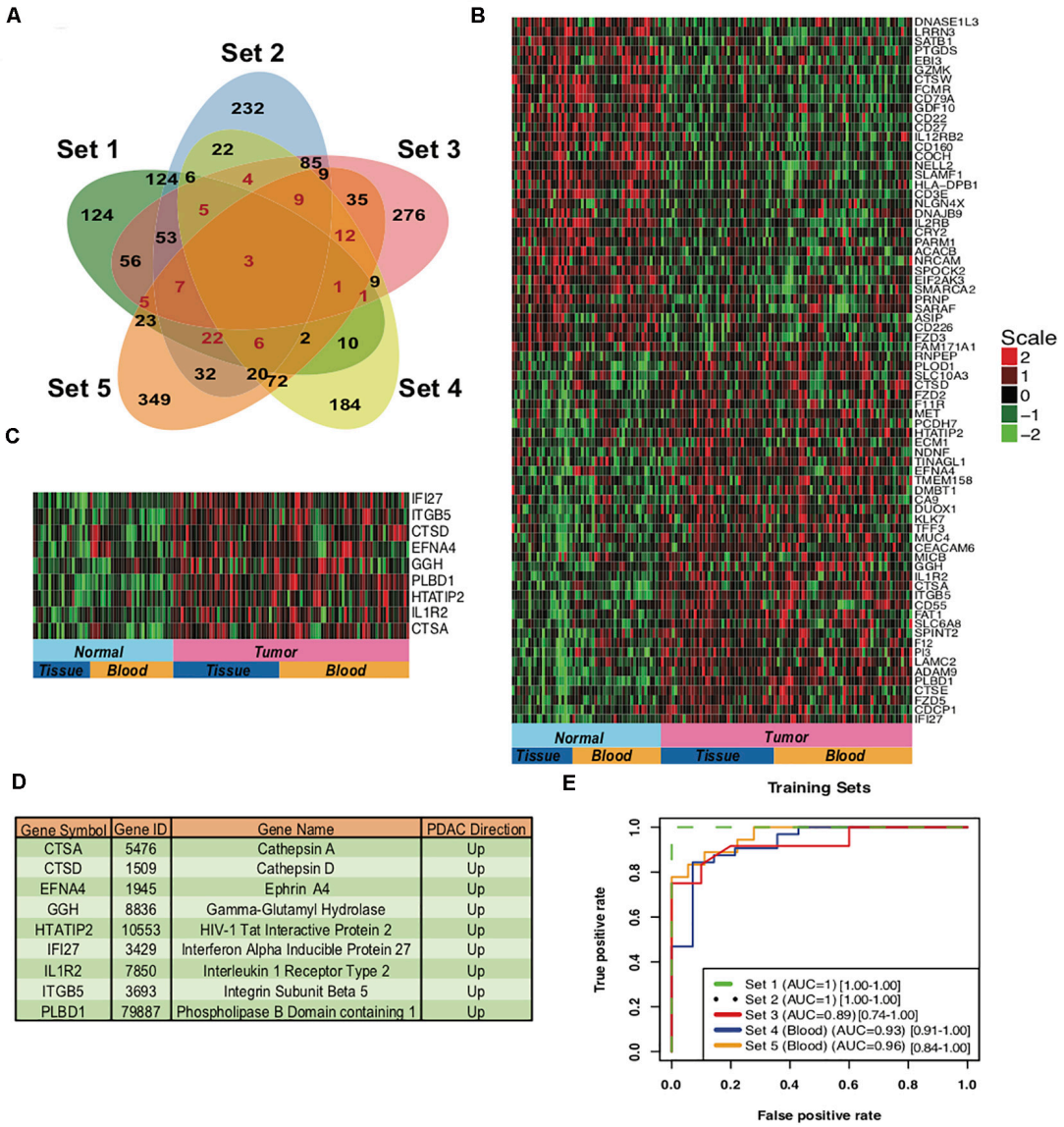

**FIGURE 1 |** Overview of the meta-analysis approach for development and validation of PDAC biomarker panel. Predictor was developed using the data from Set1-Set5 (S1-S5 in Step 4) and was further tested on Set5-Set10 and validated on V1-V5 and PV1-PV4 datasets.

peptide for secretion are supposed to be secretory; however, the signal peptide is also present in several membrane proteins (Uhlen et al., 2015). In the selected classifier genes, CTSD, EFNA4 and IL1R2 are predicted to be secretory proteins whereas CTSA, GGH, PLBD1, IFI27, ITGB5 and HTATIP2 are predicted to be intracellular or membrane bound proteins in HPA. Furthermore, CTSA and PLBD1 are also localized in Lysosomes and GGH is secretory protein as per UniProtKB[8] predictions. Since our 9 gene markers could be detected with a detectable expression in both tissues and blood samples from PDAC patients, we further validated the performance of these genes for PDAC Diagnosis.

## Independent Performance of Classifier in Differentiating PDAC From Normal

The biomarker set designed above was further tested in six independent sets with five tissue and one blood based PDAC studies. The classifier genes depicted an upregulation pattern in most of independent validation sets **Supplementary Figure S5**. The boxplot revealed higher expression of all the 9 genes, averaged over test sets, in the tumor samples as compared to the

[8]www.uniprot.org




| Gene Symbol | Gene ID | Gene Name | PDAC Direction |
|---|---|---|---|
| CTSA | 5476 | Cathepsin A | Up |
| CTSD | 1509 | Cathepsin D | Up |
| EFNA4 | 1945 | Ephrin A4 | Up |
| GGH | 8836 | Gamma-Glutamyl Hydrolase | Up |
| HTATIP2 | 10553 | HIV-1 Tat Interactive Protein 2 | Up |
| IFI27 | 3429 | Interferon Alpha Inducible Protein 27 | Up |
| IL1R2 | 7850 | Interleukin 1 Receptor Type 2 | Up |
| ITGB5 | 3693 | Integrin Subunit Beta 5 | Up |
| PLBD1 | 79887 | Phospholipase B Domain containing 1 | Up |

**FIGURE 2 |** Meta-signature of genes that are consistently DE in multiple datasets and candidate PDAC diagnostic biomarker panel. **(A)** Venn diagram of the five training datasets for the DE genes. 74 genes (marked in red) with concordant directionality are common to at least 2 of the 3 tissue datasets (Set 1 to Set 3) and one of the 2 blood datasets (Set 4 and Set 5). **(B)** Heatmap of the 74 meta-signature genes DE in PDAC from five training datasets. Red = upregulated, Green = downregulated. **(C)** Heatmap of the 9-upregulated marker genes in training sets for PDAC biomarker panel. **(D)** Description of the genes from the 9-gene based PDAC biomarker panels. **(E)** AUC plot [CI: 95%] for 9-gene PDAC classifier across the five training sets using leave one out cross-validation (LOOCV). Set1 and Set 2 are matched normal samples i.e., obtained from same individual. Set 3 normal samples are not matched, Normal samples are obtained from the patients undergoing surgery with other pancreatic diseases. Set 4 and Set 5 are blood sourced studies therefore the normal subjects were matched for gender, age and habits.

healthy (**Figure 3A**). For each of the six datasets individually, sensitivity ranged from 0.75 to 1.00 and specificity from 0.71 to 1.00 for the predictor (**Figure 3B**, **Table 2**). Comparison of the 9-gene PDAC classifier performance in tissue and blood showed an average 0.94 sensitivity and 0.97 specificity for the tissue datasets, in contrast to 0.75 sensitivity and 0.71 specificity for the blood dataset. AUC for the five tissue datasets ranged from 0.94 to 1.00 and for one blood datasets AUC was 0.80 (**Figure 3C**, **Table 2**).

## High Accuracy of Our 9-Gene PDAC Classifier in Predicting PDAC in 5 Independent Validation Sets

In five validation sets, the 9-gene PDAC classifier accurately predicted the class of PDAC compared to normal with maximum AUC of 1.00 in the independent validation tissue (V2) set that contained 20 normal and 36 PDAC samples. More than 0.95 AUC was observed in three independent validation tissue sets (V2, V3 and V4) that contained 36, 45 and 118 PDAC

**TABLE 2 |** The performance matrix of the 9-gene PDAC classifier on the training, testing, validation and prospective validation sets.

| Groups | Datasets | Accuracy | Sensitivity | Specificity | AUC |
|---|---|---|---|---|---|
| Training Sets | Set 1 | 1.00 | 1.00 | 1.00 | 1.00 |
| | Set 2 | 1.00 | 1.00 | 1.00 | 1.00 |
| | Set 3 | 0.87 | 0.83 | 0.90 | 0.89 |
| | Set 4 | 0.82 | 0.93 | 0.71 | 0.93 |
| | Set 5 | 0.86 | 0.83 | 0.89 | 0.97 |
| Test Sets | Set 6 | 1.00 | 1.00 | 1.00 | 1.00 |
| | Set 7 | 0.92 | 0.90 | 0.93 | 0.94 |
| | Set 8 | 1.00 | 1.00 | 1.00 | 1.00 |
| | Set 9 | 0.95 | 0.91 | 1.00 | 1.00 |
| | Set 10 | 0.96 | 0.93 | 1.00 | 0.94 |
| | Set 11 | 0.73 | 0.75 | 0.71 | 0.80 |
| Validation Sets | V1 | 0.79 | 0.76 | 0.83 | 0.83 |
| | V2 | 0.98 | 0.97 | 1.00 | 1.00 |
| | V3 | 0.94 | 1.00 | 0.89 | 0.98 |
| | V4 | 0.95 | 1.00 | 0.91 | 0.99 |
| | V5 | 0.83 | 0.84 | 0.82 | 0.89 |
| Prospective Validation Sets | PV1 | 0.82 | 0.94 | 0.72 | 0.93 |
| | PV2 | 0.74 | 0.74 | 0.75 | 0.82 |
| | PV3 | 0.83 | 0.78 | 0.89 | 0.95 |
| | PV4-IPMA | 1.00 | 1.00 | 1.00 | 1.00 |
| | PV4-IPMC | 0.84 | 0.83 | 0.86 | 0.81 |
| | PV4-IPMN | 1.00 | 1.00 | 1.00 | 1.00 |

and 20, 9 and 12 normal pancreas samples, respectively, (**Figure 4A** and **Table 1B**). The boxplot revealed higher expression of all the 9 genes, averaged over validation sets, in the tumor samples as compared to the healthy samples (**Figure 4B**). In a tissue dataset (V1) containing 61 normal and 69 tumor samples a specificity of 0.83 and sensitivity of 0.76 was determined. In 50 normal and 33 PDAC blood platelet sample (V5) 0.84 sensitivity, 0.82 specificity and 0.88 AUC was achieved. The prediction of the PDAC class in comparison to normal was accurate with a sensitivity ranging 0.76–1.00 and specificity ranging between 0.82 and 1.00 (**Figure 4C** panel II, **Table 2**). **Supplementary Figure S6** presents the heatmap of the nine genes in individual validation datasets and the PCA plots depicting the discrimination of PDAC from normal samples.

## Cross-Platform Performance of Classifier on TCGA Pancreatic Samples

We further estimated the cross-platform performance of classifiers on the most widely used PC sample resource namely TCGA. TCGA dataset contains 150 PDAC samples and 4 normal samples and gene expression pattern analysis is not in consistence with other studies (**Supplementary Figure S7C**). The cross-platform validation of classifier on TCGA data also achieved high sensitivity (0.94) and specificity (0.72) indicating the stability of the classifier in handling the cross-platform variation in absolute gene expression signal (**Figure 5** PV1). The classifier achieved an excellent AUC of 0.93 (**Table 2**). The lower specificity of TCGA datasets might be due to the limited number of normal samples in the dataset. Heatmap of the 9 genes and PCA plots depicts the discrimination of two classes with the nine genes in the TCGA samples (**Supplementary Figure S7** PV1).

The markers did not show concordance in the TCGA dataset; however, the significance of these genes in the survival analysis can be very well established using the TCGA database. The samples were partitioned at median for selected nine-genes and survival analysis was performed on two clusters (**Supplementary Figure S8**). The results showed the combined survival of genes was able to clearly discriminate between better and poor survivors (*P*-value significance of 0.05 and hazard Ratio of 0.85), indicating their prognostic role in PDAC. High CTSD, EFNA4, HTATIP2, IFI27, ITGB5 and PLBD1 expression is associated with shortened survival time. Also, the survival analysis of these genes with a hazard ratio of $>1$ at significant *P*-value indicates their prognostic importance.

## Performance of Classifier in Identifying Early Stage PDAC

As it is well established in literature that lack of established strategies for early detection of PDAC result in poor prognosis and mortality, we therefore tested performance of our classifiers on stage I and II PDAC. The predictor could distinguish stage I and II PDACs from normals with 0.74 sensitivity and 0.75 specificity and an AUC 0.82 (**Figure 5** PV2, **Table 2**). Heatmap of the nine genes and PCA plots depicts the discrimination of two classes with the nine genes in early stages PDAC samples (**Supplementary Figure S7** PV2).

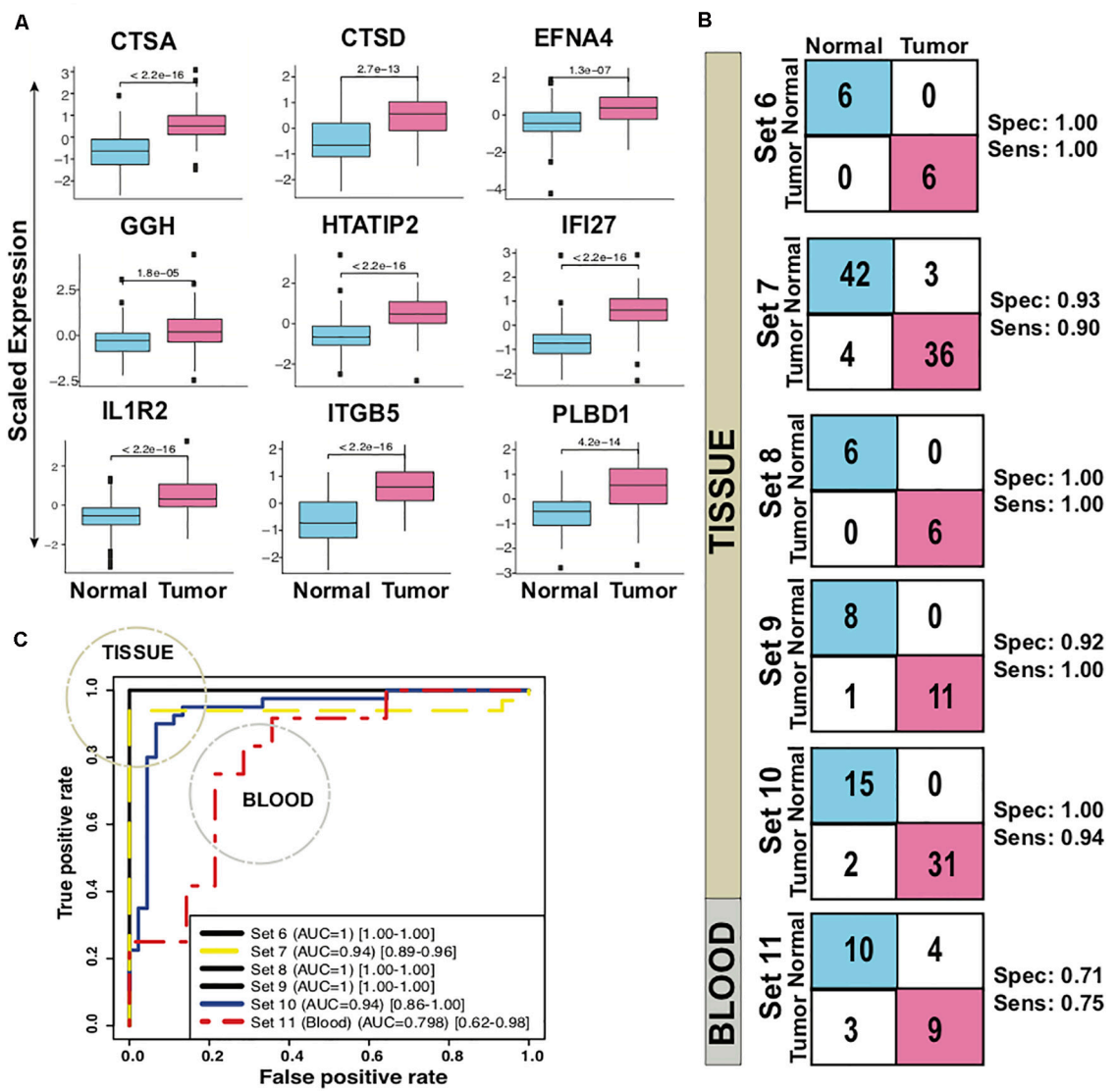



**FIGURE 3 |** Performance of 9-gene PDAC Classifier on test sets using leave one out cross-validation (LOOCV). **(A)** The boxplot of the averaged expression of the genes across all the six test datasets. The *P*-values as calculated by *t*-test between the groups are on the individual genes. **(B)** Diagnostic performance of the 9-gene PDAC classifier on the six test sets of PDAC vs. normal pancreas. Sensitivity (Sens.) and specificity (Spec.) indicated besides each set. **(C)** AUC plot for 9-gene [CI: 0.95–0.99] PDAC classifier across the six test datasets.

## Performance of Classifier in Discriminating PDAC From Pancreatitis

Identification of CP and discriminating it from PDAC is a key challenge. As our 9-gene PDAC classifier accurately established the differences between PDAC and CP, it is important to include further validation steps for the biomarker panel. The array U95Av2 have the recorded signal intensity values for all the genes except PLBD1, hence only 8 genes were tested as a classifier for the discrimination of CP from PDAC. We tested the biomarker on the PV3 dataset wherein there were nine samples each for CP and PDAC. The classifier genes on PV3 dataset depicted significantly altered expression pattern between PDAC from CP (**Supplementary Figure S7** PV3). The classifier achieved a specificity of 0.89 and sensitivity of 0.78 with an overall accuracy of 0.83 and an AUC of 0.95 in discriminating PDAC from CP (**Figure 5** PV3, **Table 2**).

## Classifier Discriminated Pre-cancerous Lesions From Normal Pancreas With Good Accuracy

To estimate the ability of the biomarker panel in discriminating precancerous lesions from a normal pancreas, we tested its performance on independent dataset containing normal main pancreatic duct epithelial cells microdissected by lasers and neoplastic epithelial cells from potential PDAC precursor lesions i.e., IPMA, IPMC and IPMN [15]. Classifier genes

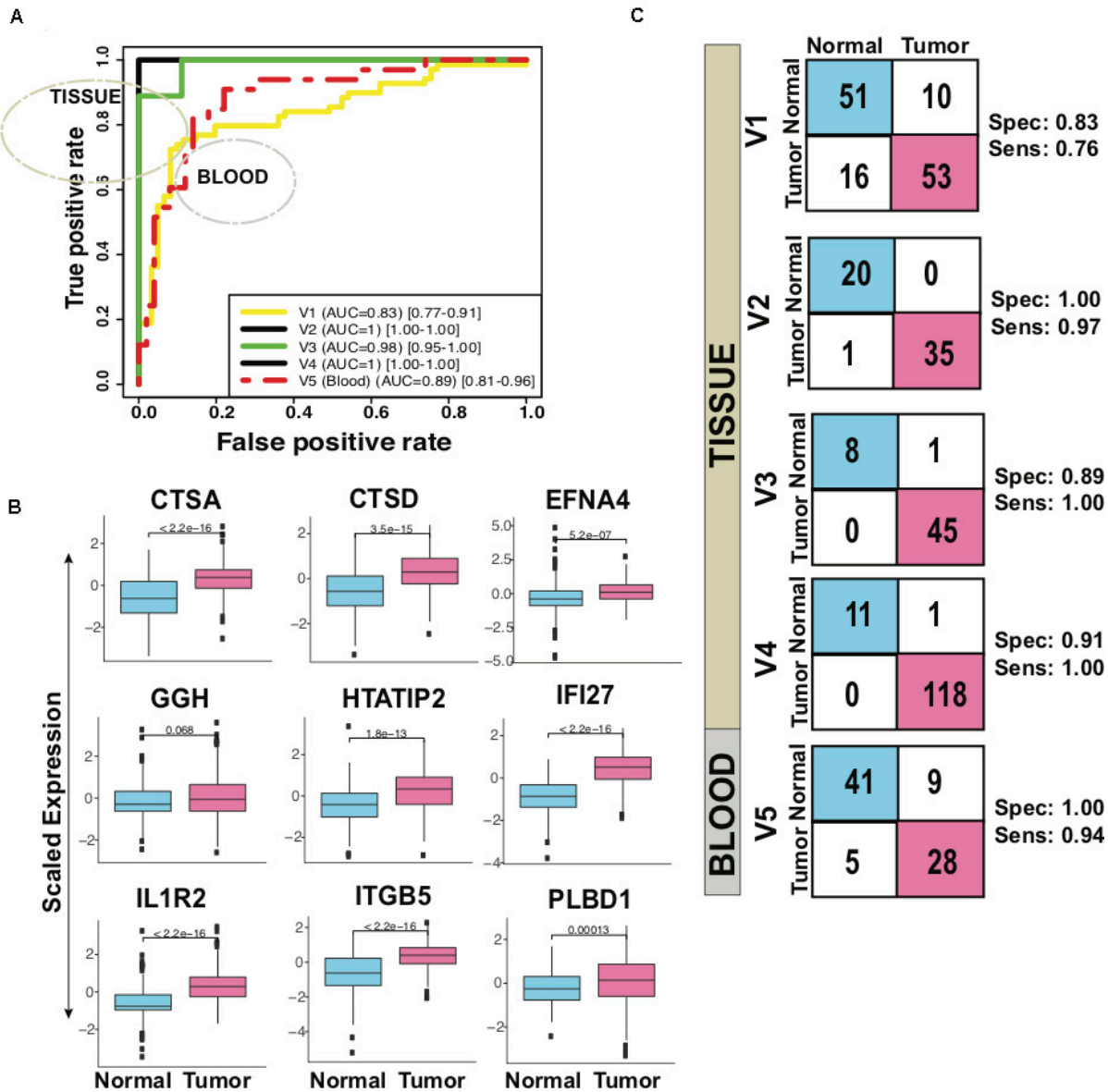



**FIGURE 4 |** Performance of 9-gene PDAC Classifier on validation sets using leave one out cross-validation (LOOCV). **(A)** The boxplot of the averaged expression of the genes across all the five validation datasets. The *P*-values as calculated by *t*-test between the groups are mentioned on the individual genes. **(B)** Diagnostic performance of the 9-gene PDAC classifier on the five validation sets of PDAC vs. normal pancreas. Sensitivity (Sens.) and specificity (Spec.) indicated besides each set. **(C)** AUC plot [CI: 0.95–0.99] for 9-gene PDAC classifier across the five validation datasets.

were consistently overexpressed in the PDAC samples, GGH was under-expressed in IPMA samples whereas it was overexpressed across the other PDAC precursors, IPMC and IPMN (**Supplementary Figure S9**). The 9-gene PDAC classifier separated all potential PDAC precursor (IPMA, IPMC, IPMN) samples from the normal pancreatic duct samples except for one normal sample and one IPMC sample (**Figure 5** PV4). The biomarker panel differed IPMA and IPMN from normal pancreas with 1.00 sensitivity and 1.00 specificity, achieving an AUC of 1.00 (**Figure 5** PV4). The predictor separated IPMC from healthy pancreas with 0.83 sensitivity and 0.86 specificity, achieving an AUC of 0.81 (**Table 2**).

## Classifier Performed Better Than Previously Known Markers

To estimate the performance of our current marker as compared to the previously established markers we compared the performance of our marker with each study [Bhasin et al. (2016), Balasenthil et al. (2017), Kisiel et al. (2015), and Immunovia (Mellby et al., 2018)]. We used polynomial

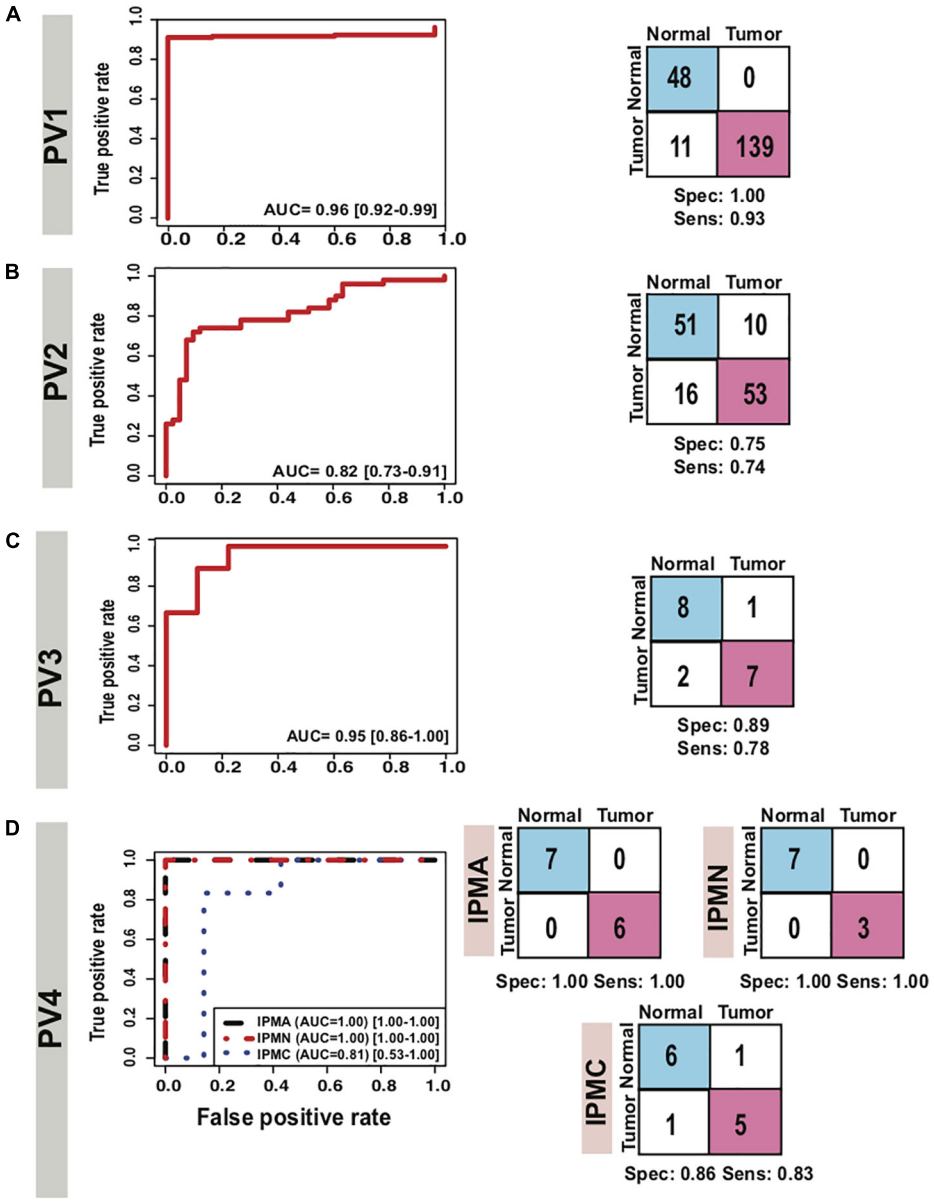


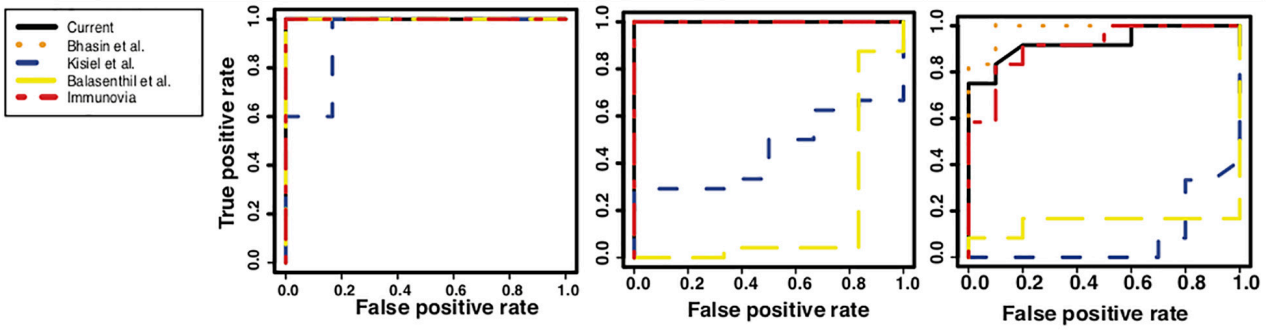

**FIGURE 5 |** Performance of 9-gene PDAC Classifier on prospective validation sets using leave one out cross-validation (LOOCV). AUC plot [CI: 0.95–0.99] for 9-gene PDAC classifier and the diagnostic performance of **(A)** the classifier for PV1 dataset, **(B)** the classifier for PV2 dataset. **(C)** The classifier for IPMA, IPMC and IPMN subjects in PV4 dataset and **(D)** the classifier for PV3 dataset.

kernel for each set of markers and selected best model to record the performance on all the training, test and validation datasets (**Supplementary Figure S10** and **Supplementary Table S3**). We found that all the methods performed well in tissue biopsies samples whereas when applied to the blood studies the performance of our marker set is the best (**Figure 6**). Our set of markers has performed well in tissues as well as blood studies and will be an ideal minimally invasive biomarker for studying in future studies and clinical trials.

## Validation of the Markers in Single-Cell Transcriptomics Studies

Furthermore, as the markers are derived from bulk sequencing protocols it is important to know if the markers discovery is not influenced by different cell-types in normal and cancerous pancreas. Therefore, we used single-cell RNA-sequencing data published by Peng et al. (2019) suggesting heterogeneity in PDAC tumor to plot expression of our markers on different cell-types. Using standard Seurat single-cell analysis methodology (Butler et al., 2018; Stuart et al., 2019), we identified that our

| TISSUE | Set 1 | | | | Set 2 | | | | Set 3 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Acc | Sens | Spec | AUC | Acc | Sens | Spec | AUC | Acc | Sens | Spec | AUC |
| Current | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 0.87 | 0.83 | 0.90 | 0.89 |
| Bhasin et al | 0.91 | 1.00 | 0.83 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 0.95 | 1.00 | 0.90 | 0.98 |
| Balasenthil et al | 1.00 | 1.00 | 1.00 | 1.00 | 0.50 | 1.00 | 0.00 | 0.57 | 0.47 | 0.83 | 0.1 | 0.16 |
| Kisiel et al | 0.63 | 0.60 | 0.67 | 0.73 | 0.50 | 1.00 | 0.00 | 0.07 | 0.20 | 0.41 | 0.00 | 0.15 |
| Immunovia | 0.80 | 0.60 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 0.78 | 0.67 | 0.90 | 0.92 |




| BLOOD | Set 4 | | | | Set 5 | | | | Set 11 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Acc | Sens | Spec | AUC | Acc | Sens | Spec | AUC | Acc | Sens | Spec | AUC |
| Current | 0.82 | 0.93 | 0.71 | 0.93 | 0.89 | 0.89 | 0.89 | 0.93 | 0.73 | 0.75 | 0.71 | 0.80 |
| Bhasin et al | 0.49 | 0.97 | 0.00 | 0.12 | 0.50 | 0.45 | 0.56 | 0.53 | 0.47 | 0.00 | 0.93 | 0.22 |
| Balasenthil et al | 0.50 | 1.00 | 0.00 | 0.01 | 0.78 | 0.78 | 0.78 | 0.81 | 0.71 | 0.58 | 0.85 | 0.65 |
| Kisiel et al | 0.50 | 1.00 | 0.00 | 0.35 | 0.47 | 0.44 | 0.50 | 0.49 | 0.51 | 0.16 | 0.85 | 0.23 |
| Immunovia | 0.72 | 0.88 | 0.57 | 0.77 | 0.59 | 0.56 | 0.62 | 0.64 | 0.29 | 0.09 | 0.50 | 0.18 |

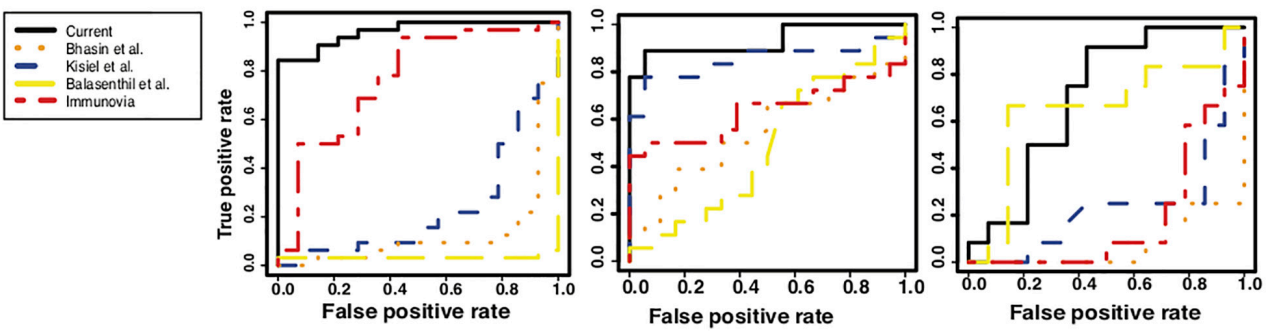



**FIGURE 6 |** Comparative performance of 9-gene PDAC Classifier with different previously established biomarkers. AUC plot [CI: 0.95–0.99] for 9-gene PDAC classifier across the three tissue and three blood datasets. The boxes colored in mustard color have greater than 0.80 AUC.

markers are not associated with any cell-types and are expressed across major cell types in pancreatic cancer (**Supplementary Figure S11**). All our markers depicted upregulation in various tumor microenvironment cells including immune cells and endothelial cells.

## Validation of Markers in Blood-Based Proteomics Study

The nine-gene markers in the classifier were discovered and validated from the transcriptomics studies, hence the validation of their expression at the protein level is necessary. Therefore, we confirmed the expression of the nine genes at the protein level in publicly available proteomics studies and HPA. The immunolabeling of the proteins of the respective genes in HPA (**Supplementary Figure S12**) suggest higher staining of the proteins in tumors as compared to the normal samples except IFI27 where the expression of the protein cannot be detected. To further validate the protein expression of our markers we searched for the corresponding proteins in multiple pancreatic cancer proteomics studies (Chen et al., 2005; Crnogorac-Jurcevic et al., 2005; Cui et al., 2009; McKinney et al., 2011; Kosanam et al.,

2013; Wang et al., 2013; Iuga et al., 2014). CTSD, a cathepsin family protein, and Ephrin and Interferon gamma family markers are found to be highly expressed in multiple proteomics studies (Chen et al., 2005; Cui et al., 2009; McKinney et al., 2011).

## DISCUSSION

We applied a data mining approach to a large number of publicly available transcriptome datasets derived from pancreatic cancer and healthy blood and tissues, followed by class prediction analysis using machine learning and validation of the classifier in the independent datasets to discover candidate PDAC biomarkers (Harsha et al., 2009; Ranganathan et al., 2009). We explored the genes with secretory peptide DE in the PDAC as compared to normal pancreas/blood, for the first time to investigate an accurate secretory/non-invasive biomarker panel for the PDAC diagnosis. We report here a 9-gene PDAC classifier that differentiates PDAC as well as the precursor lesions from the normal with high accuracy. This 9-gene PDAC classifier was validated independently in 12 different blood and tissue studies. The 9-gene PDAC classifier encodes proteins with secretory potential in pancreas and few other tissues.

Approximately 2500 candidate biomarkers have been associated with PDAC previously while some of these biomarkers are in various evaluation stages only CA19-9 is approved by FDA (Koprowski et al., 1979, 1981; Hyöty et al., 1992). However, accuracy of CA19-9 is not accurate enough for screening, especially for an early detection of PDAC. Presently, the extensive validation of diagnostic or predictive gene/protein expression biomarkers for accurate discrimination between healthy patients, benign, premalignant and malignant disease are still lacking. Therefore, we aimed to identify a biomarker panel with greater sensitivity and specificity and identified a 9-gene marker panel that performs with high accuracy in discriminating PDAC with normal pancreas across multiple platforms, using either whole/microdissected tissue or peripheral blood.

To determine whether the genes in our classifier reflect key pathophysiological pathways associated with the development of PDAC, we reviewed available information for the role of these genes. Most of our 9-gene classifier genes have been linked to tumorigenesis, indicating a causal role in the development and progression of PDAC. HTATIP2 is involved in apoptosis function in liver metastasis related genes (Shi et al., 2009), gastric cancer (Xu et al., 2010) and pancreatic cancer (Ouyang et al., 2014). IFI27, functioning in immune system, has been suggested as a marker of epithelial proliferation and cancer (Grutzmann et al., 2003; López-Casas and López-Fernández, 2010). ITGB5 involved in integrin signaling have been found to be upregulated in several analysis studies (Van den Broeck et al., 2012). The Integrin and ephrin pathways have been proposed to play a crucial role in pancreatic carcinogenesis and progression, including *ITGB1*, a paralog of *ITGB5*, and EPHA2 as most important regulators (Van den Broeck et al., 2012). EPHA2 belongs to ephrin receptor subfamily and is involved in developmental events, especially in the nervous system and in erythropoiesis. To this family belongs one of our genes EFNA4 which activates another ephrin receptor EPHA5. IL1R2 was identified as possible candidate gene in PDAC that can lead to defects of the apoptosis pathway (Rückert et al., 2010). Moreover, Il1, the ligand of IL1R2, is secreted by the pancreatic cells (Arlt et al., 2002) and has an important function in inflammation and proliferation that can also trigger the apoptosis (Dupraz et al., 2000; Ruckdeschel et al., 2002; Yoshida et al., 2004). CTSD have been shown to be upregulated in the PDAC cancer (Iacobuzio-Donahue et al., 2003). AGR2, a surface antigen, has been shown to promote the progression of PDAC cells through regulation of Cathepsins B and D genes (Dumartin et al., 2011). CTSA was identified as one of the 76 deregulated genes in a study aiming for the development of early diagnostic markers as well as potential novel therapeutic targets for both familial and sporadic PDAC (Crnogorac-Jurcevic et al., 2013). PLBD1 has been found to be upregulated in various studies with five-fold increase in cell lines (Makawita et al., 2011) and in study where the effect of pancreatic β-cells inducing immune-mediated diabetes was studied (Salem et al., 2014). Metabolism-related GGH has been found to be relevant and upregulated in gallbladder carcinomas (Washiro et al., 2008).

Most of the genes in the 9-gene classifier (ITGB1, EPHA2, IL1R2) are involved in the migration, immune pathways, adhesion and metastasis of PDAC or other cancers, that are specifically associated with the developmental events and signaling in the progression of cancer. To corroborate the involvement of these genes in PDAC progression and early stages of PDAC development, we evaluated the expression levels of these genes in the early lesions of PDAC precursors i.e., LIGD-IPMN, HGD-IPMN and InvCa-IPMN (**Figure 5**) [15]. Eight genes except GGH are upregulated in IPMA, IPMN, and IPMC as well as in PanINs, as compared to a normal pancreas, demonstrating their enhanced expression is linked with the progression of PDAC that occurs early during development of malignancy. The outcomes of our study clearly show that our 9-gene classifier reflect drivers of early defects during progression and development of PDAC. This argument is further strengthened by the survival analysis of the genes where five of the nine genes (CTSA, CTSD, EFNA4, IFI27 and IL1R2) are strongly related to discriminating better and poor survivors.

Since individuals with CP are at increased risk of developing PDAC and pathological discrimination is challenging between CP and PDAC which makes it important for a classifier to discriminate between these two disease stages. While other studies have performed meta-analysis of transcriptome data for PDAC to identify the genes that are overexpressed in PDAC (Iacobuzio-Donahue et al., 2003; López-Casas and López-Fernández, 2010; Munding et al., 2012), they are irrelevant in identifying the markers for prognosis of PDAC. Our 9-gene biomarker classifier accurately distinguished premalignant and malignant pancreatic lesions such as PanIN, IPMA, IPMN and IPMC from healthy pancreas. As all 9 genes of our classifier are upregulated in PanIN (as compare to normal pancreas) already, it indicates that these 9 genes are dysregulated in early lesions during the process of PDAC development and therefore could assist in an early detection of PDAC.

Further, to analyze the potential of the 9-gene biomarker in accurate classification of PDAC subjects versus healthy subjects

we compared our biomarker combination with previously known and established biomarkers. Our analysis also indicates that the 9-gene biomarker panel including multiple genes, rather than a single biomarker, is more powerful and had possibility to improve the specificity and selectivity for an accurate detection of PDAC. The idea behind generation of biomarker panel with the better identification in blood sample, in corroboration with the tissue studies, is fulfilled here. The previously established markers worked well in the tissue studies but could not show their similar potential in blood studies.

Further, the protein expression of selected biomarker genes was also examined to determine their association with PDAC at protein levels. The analysis depicted that multiple gene product/proteins corresponding to biomarkers genes depicted higher expression in pancreatic cancer tissues. Interestingly some marker (e.g., EFNA4, GGH) also depicted over-expression in other cancers indicating their association with tumor development and progression related hallmark processes. In recent years, multiple proteomics studies were performed to understand the proteome landscape of the PDAC but still lack in generating comprehensive picture due to technological limitations. Most of the proteomics technique can measure the expression of 2,000-3,000 proteins that is far from generating the global overview of proteome. High expression of Cathepsin family proteins specifically CTSD is noted in several proteomics studies which was also the case for Ephrin and Interferon gamma family markers (Chen et al., 2005; Cui et al., 2009; McKinney et al., 2011). Also, the expression of these genes is not found to be related to a particular cell-type in pancreatic cancer cell lineage. However, the fact that the overall study is based on bulk sequencing data cannot be overlooked and these cells may comprise of multiple cell-types which may or may not influence the overall methodology of marker selection. Overall, the protein-expression of the selected genes and their expression in multiple cell-types of pancreatic cancer is established. However, the aforementioned limitations have to be challenged before designing the diagnostic panel. The 9-gene markers identified here still needs validation in a bigger cohort for its potential in identifying accurately the early stages but this marker combination potentially has shown its discriminatory power across various blood and tissue datasets obtained from different sources and different platforms.

## 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

IK performed the bioinformatics analysis and wrote the manuscript. MB supervised the bioinformatics analysis and edited the manuscript. Both the authors read and approved the final manuscript.

## FUNDING

This study was supported through BIDMC CAO Innovation grant.

## ACKNOWLEDGMENTS

This manuscript has been released as a pre-print at medRxiv, Khatri and Bhasin (2020).

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.572284/full#supplementary-material

**FIGURE S1 |** Pathway enrichment analysis of the 74 PDAC-specific secretory genes.

**FIGURE S2 |** Upregulated Secretory genes in training datasets. **(A)** Heatmap of 27 upregulated secretory genes in PDAC for two of the three tissues and one of the two blood datasets. **(B)** PCA plots for each training datasets using 27 upregulated secretory genes.

**FIGURE S3 |** Performance of 9-gene PDAC classifier on training sets using leave one out cross-validation (LOOCV). **(A)** Diagnostic performance of the 9-gene PDAC classifier on the five training sets. Sensitivity (Sens) and Specificity (Spec) are indicated for each dataset. **(B)** AUC plot for 9-gene PDAC classifier on the three tissue training datasets. **(C)** AUC plot for 9-gene PDAC classifier on the two blood training datasets.

**FIGURE S4 |** The metrics for training datasets using the 9-biomarker panel genes. **(A)** Boxplot of the averaged expression of the genes across all the five training datasets. **(B)** PCA plots for each training datasets using the 9-biomarker panel genes.

**FIGURE S5 |** The assessment metrics for testing datasets using the 9-biomarker panel genes. **(A)** Heatmap of the 9 PDAC-upregulated marker genes. **(B)** PCA plots in six independent testing datasets.

**FIGURE S6 |** The assessment metrics for validation datasets using the 9-biomarker panel genes. Heatmaps **(A)** and PCA plots **(B)** based on biomarker panel genes in validation sets.

**FIGURE S7 |** The assessment metrics for PV1-3 dataset using the 9-biomarker panel genes. **(A)** PCA plots of three different prospective validation datasets. **(B)** Heatmaps of the 9-marker genes panel. **(C)** Boxplots of the expression of the genes.

**FIGURE S8 |** Survival curve of 9-gene-based PDAC classifier and combined genes.

**FIGURE S9 |** The assessment metrics for PV4 dataset using the 9-biomarker panel genes. **(A)** PCA plots for precursor lesions in three stages IPMA, IPMN and IPMC. **(B)** Heatmaps of the 9-marker genes panel. **(C)** Boxplots of the expression of the genes in precursor lesions.

**FIGURE S10 |** Comparative performance of 9-gene-based PDAC classifier with different previously established biomarkers. AUC plot for 9-gene-based PDAC classifier across the training and validation datasets. The measures of performances e.g., accuracy, sensitivity, specificity and AUC are mentioned in **Supplementary Table S3**.

**FIGURE S11 |** Expression of 9-gene markers in different pancreas cell-types in both healthy and tumor states. The expression of these genes is high in tumor state (CTSA, CTSD, EFNA4, GGH, HTATIP2, IFI27, and ITGB5) or they are not

expressed at all in healthy state (IL1R2 and PLBD1). This is also consistent with protein expression of the genes as measured by antibody staining experiments by HPA.

**FIGURE S12 |** Immunolabeling of protein expression of nine genes selected for the classifier in pancreatic cancer. Light blue is low staining; blue is moderate staining and brown is high.

**TABLE S1 |** Log2 fold change of the significantly DE genes identified from different training datasets.

**TABLE S2 |** Direction of differentially upregulated genes validated via boxplot analysis. Upregulated are shown with green background and ones with opposite direction are colored black.

**TABLE S3 |** Comparative performance of 9-gene PDAC Classifier with different previously established biomarkers in training, test and validation datasets. Sets with green background are datasets derived from blood. All mustard colored cells have AUC > 0.80 whereas light blue cells indicate low specificity or sensitivity despite of high AUC. For black shaded cells all the genes corresponding to the mentioned studies cannot be identified.

# REFERENCES


Arlt, A., Vorndamm, J., Müerköster, S., Yu, H., Schmidt, W. E., Fölsch, U. R., et al. (2002). Autocrine production of interleukin 1beta confers constitutive nuclear factor kappaB activity and chemoresistance in pancreatic carcinoma cell lines. *Cancer Res.* 62, 910–916.

Balasenthil, S., Huang, Y., Liu, S., Marsh, T., Chen, J., Stass, S. A., et al. (2017). A plasma biomarker panel to identify surgically resectable early-stage pancreatic cancer. *JNCI J. Natl. Cancer Inst.* 109:djw341. doi: 10.1093/jnci/djw341

Ballehaninna, U. K., and Chamberlain, R. S. (2012). The clinical utility of serum CA 19-9 in the diagnosis, prognosis and management of pancreatic adenocarcinoma: an evidence based appraisal. *J. Gastrointest. Oncol.* 3, 105–119. doi: 10.3978/j.issn.2078-6891.2011.021

Ballehaninna, U. K., and Chamberlain, R. S. (2013). Biomarkers for pancreatic cancer: promising new markers and options beyond CA 19-9. *Tumor Biol.* 34, 3279–3292. doi: 10.1007/s13277-013-1033-3

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B* 57, 289–300. doi: 10.2307/2346101

Bhasin, M. K., Ndebele, K., Bucur, O., Yee, E. U., Otu, H. H., Plati, J., et al. (2016). Meta-analysis of transcriptome data identifies a novel 5-gene pancreatic adenocarcinoma classifier. *Oncotarget* 7, 23263–23281. doi: 10.18632/oncotarget.8139

Brand, R. E., and Matamoros, A. (1998). Imaging techniques in the evaluation of adenocarcinoma of the pancreas. *Dig. Dis.* 16, 242–252. doi: 10.1159/000016872

Butler, A., Hoffman, P., Smibert, P., Papalexi, E., and Satija, R. (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* 36, 411–420. doi: 10.1038/nbt.4096

Chen, J., Bardes, E. E., Aronow, B. J., and Jegga, A. G. (2009). ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic Acids Res.* 37, W305–W311. doi: 10.1093/nar/gkp427

Chen, R., Yi, E. C., Donohoe, S., Pan, S., Eng, J., Cooke, K., et al. (2005). Pancreatic cancer proteome: the proteins that underlie invasion, metastasis, and immunologic escape. *Gastroenterology* 129, 1187–1197. doi: 10.1053/j.gastro.2005.08.001

Crnogorac-Jurcevic, T., Chelala, C., Barry, S., Harada, T., Bhakta, V., Lattimore, S., et al. (2013). Molecular analysis of precursor lesions in familial pancreatic cancer. *PLoS One* 8:e54830. doi: 10.1371/JOURNAL.PONE.0054830

Crnogorac-Jurcevic, T., Gangeswaran, R., Bhakta, V., Capurso, G., Lattimore, S., Akada, M., et al. (2005). Proteomic analysis of chronic pancreatitis and pancreatic adenocarcinoma. *Gastroenterology* 129, 1454–1463. doi: 10.1053/j.gastro.2005.08.012

Cui, Y., Tian, M., Zong, M., Teng, M., Chen, Y., Lu, J., et al. (2009). Proteomic analysis of pancreatic ductal adenocarcinoma compared with normal adjacent pancreatic tissue and pancreatic benign cystadenoma. *Pancreatology* 9, 89–98. doi: 10.1159/000178879

Dumartin, L., Whiteman, H. J., Weeks, M. E., Hariharan, D., Dmitrovic, B., Iacobuzio-Donahue, C. A., et al. (2011). AGR2 is a novel surface antigen that promotes the dissemination of pancreatic cancer cells through regulation of cathepsins B and D. *Cancer Res.* 71, 7091–7102. doi: 10.1158/0008-5472.can-11-1367

Dupraz, P., Cottet, S., Hamburger, F., Dolci, W., Felley-Bosco, E., and Thorens, B. (2000). Dominant negative MyD88 proteins inhibit interleukin-1beta /interferon-gamma -mediated induction of nuclear factor kappa B-dependent nitrite production and apoptosis in beta cells. *J. Biol. Chem.* 275, 37672–37678. doi: 10.1074/jbc.M005150200

Durinck, S., Moreau, Y., Kasprzyk, A., Davis, S., De Moor, B., Brazma, A., et al. (2005). BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics* 21, 3439–3440. doi: 10.1093/bioinformatics/bti525

Fesinmeyer, M. D., Austin, M. A., Li, C. I., De Roos, A. J., and Bowen, D. J. (2005). Differences in survival by histologic type of pancreatic cancer. *Cancer Epidemiol. Biomarkers Prev.* 14, 1766–1773. doi: 10.1158/1055-9965.EPI-05-0120

Frena, A. (2001). SPan-1 and exocrine pancreatic carcinoma. The clinical role of a new tumor marker. *Int. J. Biol. Markers* 16, 189–197. doi: 10.1177/172460080101600306

Grutzmann, R., Foerder, M., Alldinger, I., Staub, E., Brummendorf, T., Ropcke, S., et al. (2003). Gene expression profiles of microdissected pancreatic ductal adenocarcinoma. *Virchows Arch.* 443, 508–517. doi: 10.1007/s00428-003-0884-1

Harsha, H. C., Kandasamy, K., Ranganathan, P., Rani, S., Ramabadran, S., Gollapudi, S., et al. (2009). A compendium of potential biomarkers of pancreatic cancer. *PLoS Med.* 6:e1000046. doi: 10.1371/journal.pmed.1000046

Hyöty, M., Hyöty, H., Aaran, R. K., Airo, I., and Nordback, I. (1992). Tumour antigens CA 195 and CA 19-9 in pancreatic juice and serum for the diagnosis of pancreatic carcinoma. *Eur. J. Surg.* 158, 173–179.

Iacobuzio-Donahue, C. A., Maitra, A., Olsen, M., Lowe, A. W., van Heek, N. T., and Rosty, C. (2003). Exploration of global gene expression patterns in pancreatic adenocarcinoma using cDNA microarrays. *Am. J. Pathol.* 162, 1151–1162. doi: 10.1016/S0002-9440(10)63911-9

Iuga, C., Seicean, A., Iancu, C., Buiga, R., Sappa, P. K., Völker, U., et al. (2014). Proteomic identification of potential prognostic biomarkers in resectable pancreatic ductal adenocarcinoma. *Proteomics* 14, 945–955. doi: 10.1002/pmic.201300402

Kaplan, E. L., and Meier, P. (1958). Nonparametric estimation from incomplete observations. *J. Am. Stat. Assoc.* 53, 457–481. doi: 10.1080/01621459.1958.10501452

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

Khatri, I., and Bhasin, M. K. (2020). A transcriptomics-based meta-analysis combined with machine learning approach identifies a secretory biomarker panel for diagnosis of pancreatic adenocarcinoma. *medRxiv* [Preprint]. doi: 10.1101/2020.04.16.20061515

Kisiel, J. B., Raimondo, M., Taylor, W. R., Yab, T. C., Mahoney, D. W., Sun, Z., et al. (2015). New DNA methylation markers for pancreatic cancer: discovery, tissue validation, and pilot testing in pancreatic juice. *Clin. Cancer Res.* 21, 4473–4481. doi: 10.1158/1078-0432.CCR-14-2469

Koprowski, H., Herlyn, M., Steplewski, Z., and Sears, H. F. (1981). Specific antigen in serum of patients with colon carcinoma. *Science* 212, 53–55. doi: 10.1126/science.6163212

Koprowski, H., Steplewski, Z., Mitchell, K., Herlyn, M., Herlyn, D., and Fuhrer, P. (1979). Colorectal carcinoma antigens detected by hybridoma antibodies. *Somatic Cell Genet.* 5, 957–971. doi: 10.1007/bf01542654

Kosanam, H., Prassas, I., Chrystoja, C. C., Soleas, I., Chan, A., Dimitromanolakis, A., et al. (2013). Laminin, gamma 2 (LAMC2): A promising new putative pancreatic cancer biomarker identified by proteomic analysis of pancreatic adenocarcinoma tissues. *Mol. Cell. Proteomics* 12, 2820–2832. doi: 10.1074/mcp.M112.023507

Law, C. W., Chen, Y., Shi, W., and Smyth, G. K. (2014). voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol.* 15:R29. doi: 10.1186/gb-2014-15-2-r29

Law, C. W. M. (2013). *Precision Weights for Gene Expression Analysis*. Available online at: https://minerva-access.unimelb.edu.au/handle/11343/38150 (accessed October 11, 2017).

López-Casas, P. P., and López-Fernández, L. A. (2010). Gene-expression profiling in pancreatic cancer. *Expert Rev. Mol. Diagn.* 10, 591–601. doi: 10.1586/erm.10.43

Makawita, S., Smith, C., Batruch, I., Zhengţ, Y., Rü, F., Grü, R., et al. (2011). Integrated proteomic profiling of cell line conditioned media, and pancreatic juice for the identification of pancreatic cancer biomarkers. *Mol Cell Proteomics* 10:M111.008599.

McKinney, K. Q., Lee, Y. Y., Choi, H. S., Groseclose, G., Iannitti, D. A., Martinie, J. B., et al. (2011). Discovery of putative pancreatic cancer biomarkers using subcellular proteomics. *J. Proteomics* 74, 79–88. doi: 10.1016/j.jprot.2010.08.006

Mellby, L. D., Nyberg, A. P., Johansen, J. S., Wingren, C., Nordestgaard, B. G., Bojesen, S. E., et al. (2018). Serum biomarker signature-based liquid biopsy for diagnosis of early-stage pancreatic cancer. *J. Clin. Oncol.* 36, 2887–2894. doi: 10.1200/JCO.2017.77.6658

Munding, J. B., Adai, A. T., Maghnouj, A., Urbanik, A., Zöllner, H., Liffers, S. T., et al. (2012). Global microRNA expression profiling of microdissected tissues identifies miR-135b as a novel biomarker for pancreatic ductal adenocarcinoma. *Int. J. Cancer* 131, E86–E95. doi: 10.1002/ijc.26466

Ouyang, H., Gore, J., Deitz, S., and Korc, M. (2014). microRNA-10b enhances pancreatic cancer cell invasion by suppressing TIP30 expression and promoting EGF and TGF-β actions. *Oncogene* 33, 4664–4674. doi: 10.1038/onc.2013.405

Peng, J., Sun, B.-F., Chen, C.-Y., Zhou, J.-Y., Chen, Y.-S., Chen, H., et al. (2019). Single-cell RNA-seq highlights intra-tumoral heterogeneity and malignant progression in pancreatic ductal adenocarcinoma. *Cell Res.* 29, 725–738. doi: 10.1038/s41422-019-0195-y

Peran, I., Madhavan, S., Byers, S. W., and McCoy, M. D. (2018). Curation of the pancreatic ductal adenocarcinoma subset of the cancer genome atlas is essential for accurate conclusions about survival-related molecular mechanisms. *Clin. Cancer Res.* 24, 3813–3819. doi: 10.1158/1078-0432.CCR-18-0290

Ramasamy, A., Mondry, A., Holmes, C. C., and Altman, D. G. (2008). Key issues in conducting a meta-analysis of gene expression microarray datasets. *PLoS Med.* 5:e184. doi: 10.1371/journal.pmed.0050184

Ranganathan, P., Harsha, H. C., and Pandey, A. (2009). Molecular alterations in exocrine neoplasms of the pancreas. *Arch. Pathol. Lab. Med.* 133, 405–412. doi: 10.1043/1543-2165-133.3.405

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

Ruckdeschel, K., Mannel, O., and Schröttner, P. (2002). Divergence of apoptosis-inducing and preventing signals in bacteria-faced macrophages through myeloid differentiation factor 88 and IL-1 receptor-associated kinase members. *J. Immunol.* 168, 4601–4611. doi: 10.4049/jimmunol.168.9.4601

Rückert, F., Dawelbait, G., Winter, C., Hartmann, A., Denz, A., Ammerpohl, O., et al. (2010). Examination of apoptosis signaling in pancreatic cancer by computational signal transduction analysis. *PLoS One* 5:e12243. doi: 10.1371/journal.pone.0012243

Salem, H. H., Trojanowski, B., Fiedler, K., Maier, H. J., Schirmbeck, R., Wagner, M., et al. (2014). Long-term IKK2/NF- B signaling in pancreatic -cells induces immune-mediated diabetes. *Diabetes* 63, 960–975. doi: 10.2337/db13-1037

Schneider, J., and Schulze, G. (2003). Comparison of tumor M2-pyruvate kinase (tumor M2-PK), carcinoembryonic antigen (CEA), carbohydrate antigens CA 19-9 and CA 72-4 in the diagnosis of gastrointestinal cancer. *Anticancer Res.* 23, 5089–5093.

Shi, W.-D., Zhi, Q. M., Chen, Z., Lin, J.-H., Zhou, Z.-H., and Liu, L.-M. (2009). Identification of liver metastasis-related genes in a novel human pancreatic carcinoma cell model by microarray analysis. *Cancer Lett.* 283, 84–91. doi: 10.1016/J.CANLET.2009.03.030

Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W. M., et al. (2019). Comprehensive integration of single-cell data. *Cell* 177, 1888.e21–1902.e21. doi: 10.1016/j.cell.2019.05.031

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

Van den Broeck, A., Vankelecom, H., Van Eijsden, R., Govaere, O., and Topal, B. (2012). Molecular markers associated with outcome and metastasis in human pancreatic cancer. *J. Exp. Clin. Cancer Res.* 31:68. doi: 10.1186/1756-9966-31-68

Wang, J., Coombes, K. R., Highsmith, W. E., Keating, M. J., and Abruzzo, L. V. (2004). Differences in gene expression between B-cell chronic lymphocytic leukemia and normal B cells: a meta-analysis of three microarray studies. *Bioinformatics* 20, 3166–3178. doi: 10.1093/bioinformatics/bth381

Wang, W. S., Liu, X. H., Liu, L. X., Lou, W. H., Jin, D. Y., Yang, P. Y., et al. (2013). ITRAQ-based quantitative proteomics reveals myoferlin as a novel prognostic predictor in pancreatic adenocarcinoma. *J. Proteomics* 91, 453–465. doi: 10.1016/j.jprot.2013.06.032

Washiro, M., Ohtsuka, M., Kimura, F., Shimizu, H., Yoshidome, H., Sugimoto, T., et al. (2008). Upregulation of topoisomerase IIα expression in advanced gallbladder carcinoma: a potential chemotherapeutic target. *J. Cancer Res. Clin. Oncol.* 134, 793–801. doi: 10.1007/s00432-007-0348-0

Wilson, C. L., and Miller, C. J. (2005). Simpleaffy: a bioconductor package for affymetrix quality control and data analysis. *Bioinformatics* 21, 3683–3685. doi: 10.1093/bioinformatics/bti605

Xu, Z.-Y., Chen, J.-S., and Shu, Y.-Q. (2010). Gene expression profile towards the prediction of patient survival of gastric cancer. *Biomed. Pharmacother.* 64, 133–139. doi: 10.1016/J.BIOPHA.2009.06.021

Yoshida, Y., Kumar, A., Koyama, Y., Peng, H., Arman, A., Boch, J. A., et al. (2004). Interleukin 1 activates STAT3/nuclear factor-kappaB cross-talk via a unique TRAF6- and p65-dependent mechanism. *J. Biol. Chem.* 279, 1768–1776. doi: 10.1074/jbc.M311498200

**Conflict of Interest:** BIDMC will be filling patent on behalf of MB and IK on the use of biomarker panel for early PDAC diagnosis. MB is an equity holder at BiomaRx and Canomiks.

frontiers in Genetics

ORIGINAL RESEARCH
published: 12 November 2020
doi: 10.3389/fgene.2020.576442

# Integrative Computational Approach Revealed Crucial Genes Associated With Different Stages of Diabetic Retinopathy

*Nidhi Kumari[1,2,3], Aditi Karmakar[1,2,3], Saikat Chakrabarti[1,2,3]* and Senthil Kumar Ganesan[1,2,3]**

[1] Department of Structural Biology & Bioinformatics, CSIR-Indian Institute of Chemical Biology, Kolkata, India, [2] CSIR-IICB Translational Research Unit of Excellence (TRUE), Kolkata, India, [3] Academy of Scientific and Innovative Research (AcSIR), Ghaziabad, India

OPEN ACCESS

**Edited by:**
Amit Kumar Yadav, Translational Health Science and Technology Institute (THSTI), India

**Reviewed by:**
Sheng Yang, Nanjing Medical University, China
Duy Ngoc Do, Dalhousie University, Canada

***Correspondence:**
Saikat Chakrabarti
saikat@iicb.res.in
Senthil Kumar Ganesan
skumar@iicb.res.in

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

**Received:** 26 June 2020
**Accepted:** 07 October 2020
**Published:** 12 November 2020

**Citation:**
Kumari N, Karmakar A, Chakrabarti S and Ganesan SK (2020) Integrative Computational Approach Revealed Crucial Genes Associated With Different Stages of Diabetic Retinopathy. Front. Genet. 11:576442. doi: 10.3389/fgene.2020.576442

The increased incidence of diabetic retinopathy (DR) and the legacy effect associated with it has raised a great concern toward the need to find early diagnostic and treatment strategies. Identifying alterations in genes and microRNAs (miRNAs) is one of the most critical steps toward understanding the mechanisms by which a disease progresses, and this can be further used in finding potential diagnostic and prognostic biomarkers and treatment methods. We selected different datasets to identify altered genes and miRNAs. The integrative analysis was employed to find potential candidate genes (differentially expressed and aberrantly methylated genes that are also the target of altered miRNAs) and early genes (genes showing altered expression and methylation pattern during early stage of DR) for DR. We constructed a protein-protein interaction (PPI) network to find hub genes (potential candidate genes showing a greater number of interactions) and modules. Gene ontologies and pathways associated with the identified genes were analyzed to determine their role in DR progression. A total of 271 upregulated-hypomethylated genes, 84 downregulated-hypermethylated genes, 11 upregulated miRNA, and 30 downregulated miRNA specific to DR were identified. 40 potential candidate genes and 9 early genes were also identified. PPI network analysis revealed 7 hub genes (number of interactions >5) and 1 module (score = 5.67). Gene ontology and pathway analysis predicted enrichment of genes in oxidoreductase activity, binding to extracellular matrix, immune responses, leukocyte migration, cell adhesion, PI3K-Akt signaling pathway, ECM receptor interaction, etc., and thus their association with DR pathogenesis. In conclusion, we identified 7 hub genes and 9 early genes that could act as a potential prognostic, diagnostic, or therapeutic target for DR, and a few early genes could also play a role in metabolic memory phenomena.

**Keywords: diabetic retinopathy, integrative approach, candidate genes, hub genes, early genes, biomarker**

**Abbreviations:** DEGs, differentially expressed genes; DFU, diabetic foot ulcer; DMGs, differentially methylated genes; DN, diabetic nephropathy; DR, diabetic retinopathy; FVM, fibro-vascular membrane; NPDR, non-proliferative diabetic retinopathy; NV, neovascularization; PDR, proliferative diabetic retinopathy.

## INTRODUCTION

Diabetic retinopathy (DR), one of the major microvascular complications of diabetes, is affecting approximately 34.6% of diabetic individuals and has become the greatest threat to vision (Yau et al., 2012). It starts with few microaneurysms and dot hemorrhages during its initial stage, i.e., Non-Proliferative Diabetic Retinopathy (NPDR), and progresses to the sight-threatening stage called Proliferative Diabetic Retinopathy (PDR). Various structural abnormalities like thickening of basement membrane, pericyte loss, breakdown of blood-retinal barrier, etc., are also associated with DR. Studies have shown that neurodegeneration of ganglion cells is the most initial event of DR pathogenesis, which starts even before the formation of microaneurysm and dot hemorrhages (Barber and Baccouche, 2017). DR remains asymptomatic in its initial stages; however, symptoms like dark string floating in the visual field, blurred vision, etc., start appearing as the disease progresses and, if not treated, may end with loss of vision. DR is associated with alteration in various metabolic pathways like polyol pathway, hexosamine pathway, protein kinase C (PKC) pathway, accumulation of advanced glycosylation end products (AGEs), etc., which aggravates oxidative stress and inflammatory responses and thus the disease condition. Organelles like mitochondria and endoplasmic reticulum are highly affected in DR conditions. The alterations in the expression level, methylation pattern, and several other genetic and epigenetic modifications of various genes, especially those related to oxidative stress, inflammation, and angiogenesis, drive the pathogenesis and progression of DR by affecting multiple molecular pathways and functions (Wong et al., 2016).

Various epigenetic modifications such as DNA methylation, histone modifications, microRNA (miRNA), etc., occurring during the early stage of diabetes do not only regulate the expression of various genes but are also responsible for the metabolic memory phenomena (deleterious effect induced by prior glycemic exposure regardless of later glycemic control) associated with diabetes (Mishra and Kowluru, 2016; Kumari et al., 2020). This triggers the need for developing early diagnosis and treatment methods. Further, limitations of available treatments like its cost-effectiveness, variation in responses from patients to patients, unavailability in remote areas, etc., have raised the concern for better diagnosis and treatment strategies.

Almost all the complications are the result of and also lead to significant alterations in the expression pattern of various genes. There are various epigenetic, genetic, as well as other modifications responsible for such alterations. Identifying aberrantly expressed genes, miRNAs, altered methylation, and acetylation patterns are the first and the most critical step toward understanding the mechanisms by which the disease progresses, and this can be further used in finding the potential prognostic and treatment methods and also in identifying various biomarkers. Microarray profiling of genes is an emerging tool to screen significantly altered genes or miRNAs present in a specific disease condition. This tool can be exploited to identify candidate genes and diagnostic and prognostic biomarkers for a particular disease (Tarca et al., 2006; Moradifard et al., 2018).

The individual analysis of the array data is not very reliable and precise. This limitation can be overcome to some extent by overlapping usage of various relevant datasets (Curran and Hussong, 2009; S. Kim and Park, 2016). In the present study, integrative analysis of gene expression profiling microarray data, gene methylation profiling microarray data, and miRNA expression profiling microarray data were performed and various bioinformatics tools were utilized to find potential candidate genes and genes altered during early stage of DR, which may be used as a diagnostic or prognostic biomarkers specific for DR. Protein-protein interaction network construction, pathways, and functional analysis of identified genes were performed to investigate the molecular mechanisms associated with DR.

## MATERIALS AND METHODS

### Microarray Data and Processing

The data of gene expression profiling, gene methylation profiling, and miRNA expression profiling were obtained from Gene Expression Omnibus (GEO) datasets available at National Center for Biotechnology Information (NCBI)[1]. The first preference was given to datasets containing human samples for the specific disease followed by datasets containing greater number of samples and then the datasets from recent studies. The statistical significance, normalization, and quality of data present in datasets were ensured from the literature containing respective studies. We employed GEO2R tool[1] to download all the raw data ($p$-value adjusted to false discovery rate [FDR]) of a particular group of samples present in the selected dataset and identified the differentially expressed genes (DEGs), differentially methylated genes (DMGs), or differentially expressed miRNAs. In order to identify genes altered during early stage of DR, separate comparisons for PDR and NPDR were made from DR datasets containing PDR and NPDR samples.

The diabetic retinopathy (DR) gene expression profiling dataset GSE60436 (platform: GPL6884 Illumina HumanWG-6 v3.0 expression BeadChip) consisted of total 9 human samples (Japanese population) out of which 3 were taken from the normal retina and 6 from the fibrovascular membrane (FVM) of proliferative diabetic retinopathy (PDR) patients. The samples from PDR patients were grouped into active FVM (3 samples) and inactive FVM (3 samples) on the basis of presence or absence of neovascularization (NV) in the FVM, respectively (Ishikawa et al., 2015). To identify DEGs, we performed two sets of comparison: first, normal retina vs. inactive FVM (A), and second, normal retina vs. active FVM (B), with cut-off of $p$-value $< 0.05$ and absolute log fold change value ($|\log FC|$) $\geq 1.5$. However, we merged the data of both sets (A + B) as both contained the samples from PDR patients.

The diabetic retinopathy (DR) gene methylation profiling dataset GSE57362 (platform: GPL13534 Illumina Human

[1]https://www.ncbi.nlm.nih.gov/gds

Methylation450 BeadChip [HumanMethylation450_15017482]) consisted of total 265 human samples (Spanish population) out of which 8 were from normal neuroretina, 8 were from neuroretina of non-proliferative diabetic retinopathy (NPDR) patients, 9 were from FVM of DR patients, and the rest were from patients suffering from other ocular diseases (Berdasco et al., 2017). Here also two sets of comparison were performed: first, normal neuroretina vs. neuroretina of NPDR with |log FC| > = 0.2 (G), and second, normal neuroretina vs. FVM of PDR with |log FC| > = 0.5 (H), to identify DMGs with $p$-value < 0.05. Here, we have set less threshold for |log FC| with the assumption that fold change depends on many factors like type of study performed, stage of disease at which sample was collected, methods used to perform the study, etc. Hence we assumed that the fold change in methylation profiling study might be far lower than that in the expression profiling study (Maag et al., 2017; Raman et al., 2018; Abdulrahim et al., 2019; He et al., 2019; Yang et al., 2019).

The diabetic retinopathy (DR) miRNA expression profiling dataset GSE140959 (platform: GPL16384 [miRNA-3] Affymetrix Multispecies miRNA-3 Array) consisted of total 73 human samples (from United States) of macular hole (MH), PDR, and NPDR patients from aqueous humor (10 MH, 4 NPDR, 10 PDR), vitreous humor (10 MH, 4 NPDR, 10 PDR), and plasma (10 MH, 4 NPDR, 11 PDR) (Smit-McBride et al., 2020). For this dataset a total of 4 comparisons were made with $p$-value < 0.05 and |log FC| > = 1.5: first, aqueous and vitreous humor of normal vs. NPDR (C'); second, aqueous and vitreous humor of normal vs. PDR (D'); third, plasma of normal vs. NPDR (E'); and fourth, plasma of normal vs. PDR (F'). Plasma samples were compared separately with the thought of identifying any circulatory biomarker.

The diabetic nephropathy (DN) gene expression profiling dataset GSE1009 (platform: GPL8300 [HG_U95Av2] Affymetrix Human Genome U95 Version 2 Array) consisted of total 6 human samples (from Netherlands) out of which 3 were from the glomeruli of normal kidney and 3 from the glomeruli obtained from the diabetic nephropathy kidney (Baelde et al., 2004), and the DEGs were identified by performing comparison between glomeruli of normal kidney vs. glomeruli of DN kidney (1) with cut-off of $p$-value < 0.05 and |log FC| ≥ 1.5.

The diabetic nephropathy (DN) miRNA expression profiling dataset GSE51674 (platform: GPL10656 Agilent-029297 Human miRNA Microarray v14 Rev.2 [miRNA ID version]) consisted of total 16 human samples (from Italy) out of which 4 were from kidney of healthy control, 6 were from kidney of DN patients, and 6 from kidney of diabetic patients with membranous nephropathy (Conserva et al., 2019), and comparison was made between kidney tissue sample of normal vs. DN individuals (3) with $p$-value < 0.05 and |log FC| ≥ 1.5.

The diabetic foot ulcer (DFU) gene expression profiling dataset GSE80178 (platform: GPL16686 [HuGene-2_0-st] Affymetrix Human Gene 2.0 ST Array [transcript (gene) version]) consisted of total 12 human samples (from United States) out of which 6 were of diabetic foot ulcer, 3 of diabetic foot skin, and 3 of non-diabetic foot skin (Ramirez et al., 2018), and DEGs were identified from comparison between non-diabetic foot skin vs. diabetic foot ulcer (2) with cut-off value of $p$-value < 0.05 and |log FC| ≥ 1.5.

The diabetic foot ulcer (DFU) miRNA expression profiling dataset GSE84971 (platform: GPL17537 nCounter Human miRNA Expression Assay, V2) consisted of total 6 human foot fibroblast samples (from United States) out of which 3 were from diabetic foot ulcer and 3 from non-diabetic foot (Liang et al., 2016), and the comparison between foot fibroblast samples of non-diabetic foot vs. diabetic foot ulcer (4) was made with $p$-value < 0.05 and |log FC| ≥ 1.5.

Datasets and the sets of comparison are summarized in **Table 1**, and the overall work-flow of the study is summarized in **Figure 1**.

## Determination of Specific and Overlapping Genes and miRNAs

We got two sets of data from each comparison [upregulated (log FC ≥ 1.5) and downregulated (log FC ≤ −1.5) from expression data and hypermethylated (log FC ≥ 0.2 for NPDR and log FC ≥ 0.5 for PDR) and hypomethylated (log FC ≤ −0.2 for NPDR and log FC ≤ −0.5 for PDR) from methylation data]. The specific and overlapping genes or miRNAs were determined using online software Draw Venn Diagram[1].

### DR Specific Genes and miRNAs

We performed stepwise comparisons. Initially, to find DR specific aberrantly expressed genes (specific AB) and DR specific aberrantly expressed miRNAs [specific (C'+D'+E'+F')], we compared differentially expressed genes or miRNAs data of DR with that of DN and DFU (**Figure 1**) and excluded all those genes and miRNAs that were not exclusively present in DR from further analysis.

### Potential DR Candidate Genes

We assumed potential DR candidate genes as the genes that showed altered expression as well as methylation pattern and were also the target of altered miRNAs. To identify potential candidate genes for DR, we compared DR specific differentially expressed genes (specific AB), differentially methylated genes (G+H), and targets of the DR specific altered miRNA (C+D+E+F) (**Figure 1**). The genes that were common among all the three groups were considered as potential candidate genes for DR. To find the targets of altered miRNA, we used miRTarBase[2] and chose the targets on the basis of strong experimental evidence such as Reporter assay, Western blot, and qPCR.

### Genes Involved in Early Stage of DR

The criteria for choosing early genes, i.e., the genes involved in an early stage of DR, was to find genes that show differential expression and aberrant methylation pattern in the early stage (NPDR) of DR. So, we compared DR specific differentially expressed genes (specific AB), differentially methylated genes in NPDR (G), and differentially methylated genes in PDR (H). Genes that were present in (specific AB) and (G) were considered as the genes altered during early stage of the disease and can be targeted for early diagnosis and treatment (**Figure 1**).

[2]https://www.ncbi.nlm.nih.gov/geo/geo2r/

**TABLE 1 |** Datasets and the set of comparisons.

| Dataset | Comparisons | No. of sample control/diseased | Platform | References |
|---|---|---|---|---|
| GSE60436 (DR_mRNA) | Normal retina vs. inactive FVM of PDR (A) | 3/3 | GPL6884 | Ishikawa et al., 2015 |
| | Normal retina vs. active FVM of PDR (B) | 3/3 | | |
| GSE57362 (DR_methylation) | Normal neuroretina vs. neuroretina of NPDR (G) | 8/8 | GPL13534 | Berdasco et al., 2017 |
| | Normal neuroretina vs. FVM of PDR (H) | 8/9 | | |
| GSE140959 (DR_miRNA) | Aqueous and vitreous humor of normal vs. NPDR (C') | 10/4 and 10/4 | GPL16384 | Smit-McBride et al., 2020 |
| | Aqueous and vitreous humor of normal vs. PDR (D') | 10/10 and 10/10 | | |
| | Plasma of normal vs. NPDR (E') | 10/4 | | |
| | Plasma of normal vs. PDR (F') | 10/11 | | |
| GSE1009 (DN_mRNA) | Glomeruli of normal kidney vs. DN kidney (1) | 3/3 | GSL8300 | Baelde et al., 2004 |
| GSE51674 (DN_miRNA) | Kidney tissue sample of normal vs. DN (3) | 4/6 | GPL10656 | Conserva et al., 2019 |
| GSE80178 (DFU_mRNA) | Non-diabetic foot skin vs. DFU (2) | 3/6 | GPL16686 | Ramirez et al., 2018 |
| GSE84971 (DFU_miRNA) | Foot fibroblast sample of non-diabetic foot vs. DFU (4) | 3/3 | GPL17537 | Liang et al., 2016 |

## PPI Network Construction, Hub Gene, and Module Identification

We considered hub genes as those potential candidate genes of DR that possess a large number of interactions. Search Tool for the Retrieval of Interacting Genes (STRING) database is one of the most familiar tools to determine the known and predicted interactions among a set of proteins. STRING version 11.0[3] was used to construct the interaction network between potential candidate genes with sources of interactions including experiments, databases, text mining, co-occurrence, co-expression, and protein homology. A high confidence cut off $\geq 0.7$ of minimum interaction score was used to extract the interactions. With the help of Molecular Complex Detection (MCODE) (Bader and Hogue, 2003) and CytoHubba (Chin et al., 2014) applications of Cytoscape (Shannon et al., 2003) we determined highly interconnected clusters or module and hub genes, respectively, present in our PPI networks (**Figure 1**).

## Gene Ontology and Pathway Analysis

Though gene ontology (GO) provides various biological processes, molecular functions, and sub-cellular localizations of genes, it doesn't contain any information about the pathways that are associated with the genes. Various subsets of GO and pathways are interdependent and interconnected with each other, so to understand the mechanisms by which a gene works, it is necessary to determine various gene ontologies along with their associated pathways. In this study, gene ontology and pathway analysis were performed for potential DR candidate genes, genes altered during the early stage of DR, i.e., NPDR stage, hub genes, and genes present in the interconnected module. Gene Ontology Resource (GOR)[3] and Database for Annotation Visualization and Integrated Discovery (DAVID)[4] are among the most well-known tools to perform gene enrichment analysis. The enrichment analysis of Gene Ontology (GO) (GO biological process complete and GO molecular function complete) was performed using GOR database while that of KEGG pathways was performed using DAVID database with cutoff value of FDR $p < 0.05$. Further, to go into the details of each individual genes present in the hub genes and early genes of DR, we determined the GO and KEGG pathways for each of those genes separately using QuickGo[5] and KEGG[6] databases (**Figure 1**).

# RESULTS

## Identification of Altered Genes and miRNAs in DR

The GEO2R analysis of different gene expression profiling datasets identified total 743 upregulated and 971 downregulated genes in DR, 855 upregulated and 408 downregulated genes in DN, 353 upregulated and 864 downregulated genes in DFU, respectively. In the case of gene methylation profiling of DR dataset total 81 hypermethylated genes in NPDR, 83 hypomethylated genes in NPDR, 584 hypermethylated genes in PDR, and 3699 hypomethylated genes in PDR were identified. The miRNA expression profiling of different datasets identified total 11 upregulated and 30 downregulated miRNA in DR, 126 upregulated and 35 downregulated miRNA in DN, 27

[3]https://string-db.org/

[4]https://david.ncifcrf.gov/

[5]https://www.ebi.ac.uk/QuickGO/

[6]https://www.genome.jp/kegg/pathway.html

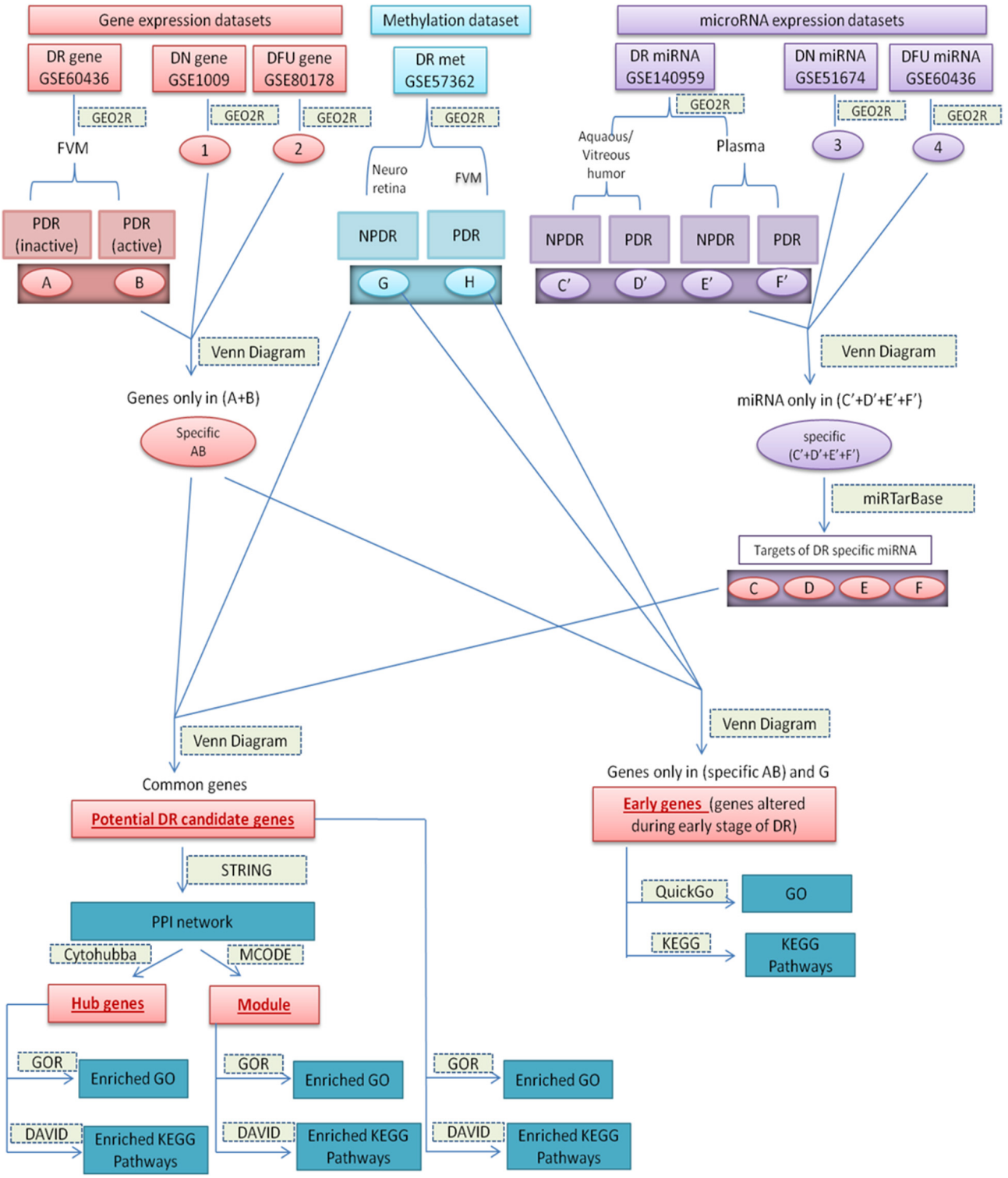



**FIGURE 1 |** Diagrammatic representation of overall work-flow. **(A)**: (DEGs in inactive FVM of PDR); **(B)**: (DEGs in active FVM of PDR); **(G)**: (DMGs in neuroretina of NPDR); **(H)**: (DMGs in FVM of PDR); **(1)**: (DEGs in glomeruli of DN kidney); **(2)**: (DEGs in DFU); **(3)**: (differentially expressed miRNAs in kidney tissue of DN); **(4)**: (differentially expressed miRNAs in foot fibroblast sample of DFU); **(C')**: (differentially expressed miRNAs in Aqueous and vitreous humor of NPDR); **(D')**: (differentially expressed miRNAs in Aqueous and vitreous humor of PDR); **(E')**: (differentially expressed miRNAs in plasma of NPDR); **(F')**: (differentially expressed miRNAs in plasma of PDR); **(Specific AB)**: (DEGs specific for DR); **(Specific C'–F')**: (differentially expressed miRNAs specific for DR); **(C–F)**: (targets of DR specific differentially expressed miRNA).

upregulated and 1 downregulated miRNA in DFU, respectively (**Supplementary Table 1**).

## Identification of Potential DR Candidate Genes

### DR Specific Genes and miRNAs

The Venn diagram revealed various specific and overlapping genes. Comparing gene and miRNA expression datasets of DR, DN, and DFU revealed total 681 upregulated genes [specific (AB).u], 884 downregulated genes [specific (AB).d], 11 upregulated miRNA, and 30 downregulated miRNA specific for DR (**Figures 2**, **3**). **Table 2** enlists all miRNA specific for DR. Further, among 681 up-regulated genes, 271 genes were also found to be hypo-methylated, and 78 were the targets of down-regulated miRNA, and among 884 down-regulated genes 84 genes were also hyper-methylated and 8 were the targets of up-regulated miRNA (**Figure 4** and **Supplementary Table 1**).

### Potential DR Candidate Gene

A total of 40 potential DR candidate genes (genes showing altered expression as well as methylation pattern and were also the target of altered miRNAs) were identified. All of them were upregulated, hypomethylated, and targets of downregulated miRNA (**Figure 4** and **Supplementary Table 1**).

## Genes Involved in Early Stage of DR

Various pathological changes in retina of diabetic individual start even before the appearance of DR associated symptoms (Vujosevic et al., 2019). Thus DR remains asymptomatic during its initial stages, and by the time symptoms appear, the individual already suffers with some vision loss. The available treatment can preserve the remaining vision but can't compensate for the already lost vision (Ellis et al., 2013). Further, metabolic memory phenomenon is believed to occur due to various epigenetic modifications occurring during the early stage of the disease (Intine and Sarras, 2012; Maghbooli et al., 2015). Hence, determining genes that play a critical role during early stage of the disease could help in preventing the disease progression during early stage of DR and could also help in finding a way to abolish metabolic memory phenomenon. We found a total of 9 genes showing differential expression and methylation pattern in the early stage of DR, i.e., NPDR out of which 5 (*NR1H4*, *ROCK2*, *HTATIP2*, *UHRF1*, and *NTM*) were upregulated-hypomethylated and 4 (*MAPT*, *FAM69C*, *FHOD3*, and *IGSF21*) were downregulated-hypermethylated genes (**Figure 5**, **Table 3**, and **Supplementary Table 1**). The identified early genes were found to be involved in one or more crucial events associated with DR progression like angiogenesis, oxidative stress, inflammation, etc. Further, some of the early genes like *ROCK2* (Koch et al., 2014; Lu et al., 2020), *UHRF1* (Ramesh et al., 2016), and *MAPT* (C. C. Zhang et al., 2016) are shown to participate in neurodegeneration, which is one of the earliest events in DR development. Also, in the present study, one of the identified early genes, i.e., *NR1H4,* was also found to be the target of one of the downregulated miRNAs (has-mir-192) identified in plasma sample of NPDR cases.

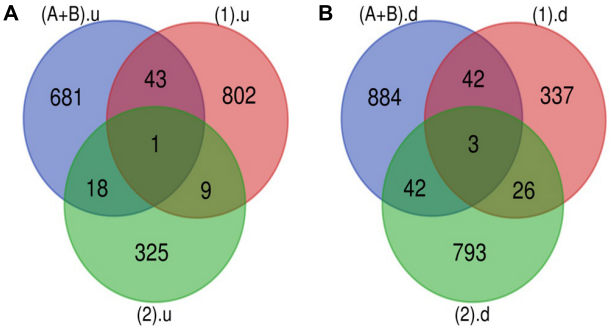



**FIGURE 2 |** Genes specific for DR. **(A)** Up-regulated genes: total 743 (681 + 18 + 1 + 43) up-regulated genes are present in DR out of which 681 are uniquely present in DR cases. **(B)** Down-regulated genes: total 971 (884 + 42 + 3 + 42) down-regulated genes are present in DR out of which 884 are uniquely present in DR cases (A+B).u: (up-regulated genes in inactive and active FVM of DR); (1).u: (up-regulated genes in DN); (2).u: (up-regulated genes in DFU); (A+B).d: (down-regulated genes in inactive and active FVM of DR); (1).d: (down-regulated genes in DN); (2).d: (down-regulated genes in DFU).

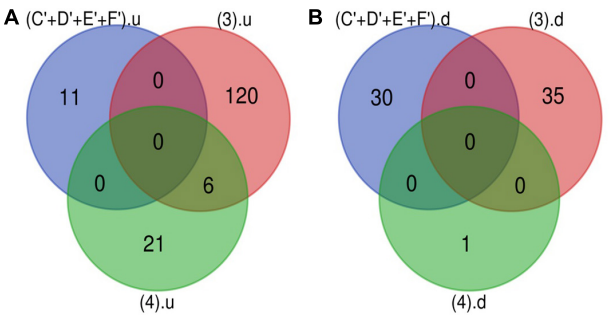



**FIGURE 3 |** microRNA specific for DR. **(A)** Upregulated miRNA: 11 up-regulated miRNAs are uniquely present in DR cases. **(B)** Downregulated miRNA: 30 down-regulated miRNAs are uniquely present in DR cases (C'+D'+E'+F').u: (up-regulated miRNA in DR); (3).u: (up-regulated miRNA in DN); (4).u: (up-regulated miRNA in DFU); (C'+D'+E'+F').d: (down-regulated miRNA in DR); (3).d: (down-regulated miRNA in DN); (4).d: (down-regulated miRNA in DFU).

## PPI Network Construction, Hub Genes, and Module Identification

The PPI network of potential candidate genes showed a total of 26 interacting nodes with minimum interaction score of 0.7 (high confidence) (**Figure 6A**). MCODE detected 1 module having MCODE score 5.67 and number of node 13 (**Figure 6B**) while CytoHubba revealed 7 hub genes (*FN1*, *IL-6*, *COL1A2*, *COL4A1*, *COL4A2*, *SPARC*, and *MMP9*) (**Table 3**) with number of interactions >5 in the PPI network. DR specific miRNAs associated with hub genes are listed in **Table 4**. Most of the identified hub genes belong to the collagen group of extracellular matrix. Though evidence suggests involvement of extracellular matrix component in DR pathogenesis, not much study has been completed on collagen in association with DR. However, genes like *COL4A1* (Alavi et al., 2016), *COL4A2* (Alavi et al., 2016), and

**TABLE 2 |** Differentially expressed DR specific miRNA.

| | |
|---|---|
| Up-regulated miRNA | hsa-mir-320a, ssc-mir-24-2*, hsa-mir-320d-2, hsa-mir-455, hsa-mir-320d-1, tni-mir-23a-2*, tni-mir-23a-1*, ssc-mir-24-1*, hsa-let-7b, lca-mir-23a*, tni-mir-23a-3* |
| Down-regulated miRNA | hsa-mir-16-2, hsa-mir-486, hsa-mir-20b, hsa-mir-15b, hsa-let-7i, hsa-mir-16-1, hsa-mir-30e, hsa-mir-502, hsa-mir-17, hsa-mir-20a, hsa-mir-532, hsa-mir-18a, hsa-mir-222, hsa-mir-363, hsa-mir-194-2, hsa-mir-660, hsa-mir-194-1, hsa-mir-29a, hsa-let-7g, hsa-mir-130a, hsa-mir-27a, hsa-mir-192, hsa-mir-106b, hsa-mir-150, hsa-mir-106a, hsa-mir-25, hsa-mir-451, hsa-mir-15a, hsa-mir-30d, hsa-mir-126 |

**microRNA from species other than human. Human homologous of these miRNA were considered to find their targets.*

*FN1* (Moradipoor et al., 2016) are found to have association with the DR pathogenesis. Studies have also found *SPARC* (Fu et al., 2019), *IL-6* (Rojas et al., 2011), and *MMP9* (Kowluru et al., 2012) playing roles in DR development by affecting one or more factors responsible for DR like angiogenesis, inflammation, etc.

## Gene Ontology and KEGG Pathway Analysis

The gene ontology and KEGG pathway analysis revealed many biological processes, molecular functions, and pathways linked with potential candidate genes, hub genes, genes present in module, and some of the early genes of DR that can play an essential role in the pathogenesis of DR by regulating each other, enhancing pathological activities, and forming other cross communications.

The enrichment analysis revealed that the potential candidate genes, hub genes, and genes present in modules were enriched in molecular functions like binding to protein, organic cyclic compound, ions and extracellular matrix; hydrolase, oxidoreductase, transferase, and catalytic activity; transcription regulation; signaling receptor; enzyme and receptor regulation, etc. Further, the enriched biological processes were various cellular processes like signal transduction, movement of cells, cellular metabolic processes, cellular response to stimulus, regulation of various biological processes and molecular functions, immune response, leukocyte migration, oxidation-reduction process, metabolic processes, cell adhesions, etc. The enriched KEGG pathways were PI3K-Akt signaling pathway, ECM-receptor interaction, Focal adhesion, TNF signaling pathway, Toll-like receptor signaling pathway, Protein digestion and absorption, NOD-like receptor signaling pathway, Chemokine signaling pathway, etc. Some of the enriched GO and KEGG pathways of hub genes are shown in **Table 5** while the list of probable DR associated enriched GO and KEGG pathways of potential candidate genes, hub genes, and genes present in module are provided in **Supplementary Table 2**.

The gene ontology and pathway analysis of individual hub genes and early genes showed that many of those genes were involved in biological processes, molecular functions, and pathways that are or can be associated with DR pathogenesis. For example, biological processes like angiogenesis, inflammatory response, neurogenesis, blood vessel development, extracellular matrix organization, etc.; molecular functions like protein binding, receptor binding, collagen binding, growth factor activity, extracellular matrix structural constituent, etc.; KEGG pathways like ECM receptor interaction, AGE-RAGE signaling pathways, PI3K-Akt signaling pathway, focal adhesion, etc. were associated with one or more genes. **Supplementary Table 3** enlists the gene ontologies and KEGG pathways of individual hub genes and early genes based on their probable association with DR pathogenesis.

# DISCUSSION

The complications associated with human health results from alterations in gene expression pattern, either by genetic,

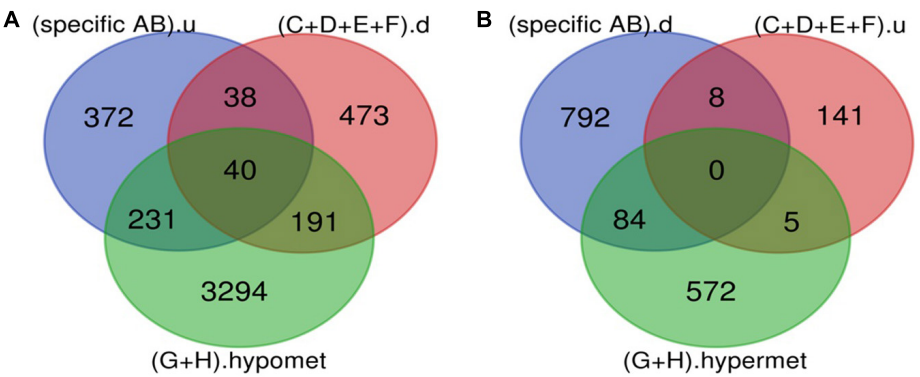



**FIGURE 4 |** Potential DR candidate genes. **(A)** Upregulated genes_Down-regulated miRNA targets_Hypomethylated genes: 40 up-regulated and hypomethylated genes were also the target of DR specific down-regulated miRNA. **(B)** Down-regulated genes_Up regulated miRNA targets_Hypermethylated genes: 0 down-regulated and hypermethylated were the target of DR specific up-regulated miRNA (specific AB).u: (up-regulated genes specific for DR); (C+D+E+F).d: (targets of DR specific down-regulated miRNA); (G+H).hypomet: (hypomethylated genes in NPDR and PDR); (specific AB).d: (down-regulated genes specific for DR); (C+D+E+F).u: (targets of DR specific up-regulated miRNA); (G+H).hypermet: (hypermethylated genes in NPDR and PDR).

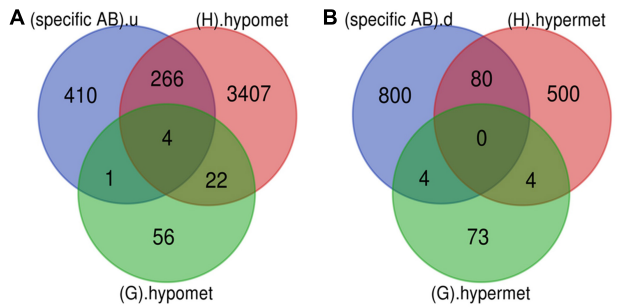



**FIGURE 5 |** Early genes. **(A)** Up-regulated and hypomethylated genes: total 83 genes were hypomethylated during NPDR while 3699 during PDR out of which 271 genes were also up-regulated in DR cases. 5 up-regulated and hypomethylated genes were present in NPDR stage of DR.
**(B)** Down-regulated and hypermethylated genes: total 81 genes were hypermethylated during NPDR while 584 during PDR out of which 84 genes were also downregulated in DR cases. 4 down-regulated and hypermethylated genes were present in NPDR stage of DR (specific AB).u: (up-regulated genes specific for DR); (G).hypomet: (hypomethylated genes in NPDR); (H).hypomet: (hypomethylated genes in PDR); (specific AB).d: (down-regulated genes specific for DR); (G).hypermet: (hypermethylated genes in NPDR); (H).hypermet: (hypermethylated genes in PDR).

**TABLE 3 |** List of probable DR associated genes.

| Categories | | Genes |
|---|---|---|
| Hub genes | | *FN1* (Fibronectin 1), *IL6* (Interleukin 6), *COL1A2* (Collagen Type I Alpha 2 Chain), *COL4A1* (Collagen Type IV Alpha 1 Chain), *COL4A2* (Collagen Type IV Alpha 2 Chain), *SPARC* (Secreted Protein Acidic And Cysteine Rich), *MMP9* (Matrix Metallopeptidase 9) |
| Early genes | Upregulated and hypomethylated | *NR1H4* (Nuclear Receptor Subfamily 1 Group H Member 4), *ROCK2* (Rho Associated Coiled-Coil Containing Protein Kinase 2), *HTATIP2* (HIV-1 Tat Interactive Protein 2), *UHRF1* (Ubiquitin Like With PHD And Ring Finger Domains 1), *NTM* (Neurotrimin) |
| | Downregulated and hypermethylated | *MAPT* (Microtubule Associated Protein Tau), *FAM69C* (Family With Sequence Similarity 69 Member C), *FHOD3* (Formin Homology 2 Domain Containing 3), *IGSF21* (Immunoglobin Superfamily Member 21) |

*Hub genes: potential candidate genes of DR which possess large number of interaction. Early genes: genes showing altered expression and methylation pattern during an early stage of DR.*

epigenetic modifications or other mechanisms. Today, the increased rate of diabetic incidences has also increased the rate of its associated complications, and diabetic retinopathy that affects approximately 34.6% of diabetic individuals (Yau et al., 2012) accounts for about 4.8% cases of blindness worldwide (Drake, 2007). Also available DR treatments suffer from one or more limitations such as economic burden, variability in drug response among patients, accessibility of the healthcare in rural areas, etc. Further, metabolic memory phenomena associated with diabetes has increased a great concern for early diagnosis and treatment strategies. Therefore, determining DR specific potential genes, genes altered during early stage of DR, their functions, molecular pathways, and interacting partners may lead to the finding of early diagnostic and better treatment methods. DNA methylation and miRNAs are among the various epigenetic modifications that are responsible for alterations in various genes expression during DR pathogenesis (Maghbooli et al., 2015; X. Zhang et al., 2017). Thus, they may play a crucial role in regulation of various biological processes, functions, and pathways associated with DR. Hence integration of gene expression profiling, gene methylation profiling, and miRNA expression profiling data could help in identification of more accurate and specific genes that may play an indispensable role in DR progression and pathogenesis.

Abnormal inflammation, oxidative stress, and neovascularization are the prime events responsible for vision loss in DR. The inflammatory responses like adhesion of leukocytes with endothelial cells and their migration toward the inflamed area aggravates the pathogenesis. Further neurodegeneration is another event observed during early stages of DR. Hence, the product of any genes whose pathways, functions, or processes affect these events either directly or indirectly can be involved in the disease progression.

Regarding the individual genes, we find that most of the hub genes belong to the collagen group of extracellular matrix. Studies have shown various extracellular matrix components to be involved in the development of DR, but only a few studies have been done on collagen in context to DR. There is not much study done on *COL1A2* (Collagen Type I Alpha 2 Chain) about DR. However, Type IV collagen, the major protein of basement membrane matrix, shows increase in its expression level in vitreous and probably also in serum with the duration of diabetes and is exalted in DR condition (Kotajima et al., 2001). Mutation in *COL4A1* (Collagen Type IV Alpha 1 Chain) and *COL4A2* (Collagen Type IV Alpha 2 Chain) has been found to elevate the risk of DR development by causing various abnormalities like vascular lesions, raising the expression of *Vegfa, Pdgfb,* and *Pgf* leading to neovascularization (Alavi et al., 2016). *COL4A1* is also associated with obesity, one of the risk factors of diabetes. Moreover, *FN1* (Fibronectin 1), a gene encoding fibronectin, is found to be up-regulated in T2DM and might be involved in angiogenesis, inflammatory response, and cell adhesion. Its level is increased in various tissues including retina, thus changing extracellular matrix (ECM) in endothelium and promoting damage to vessels wall. Also, endothelin- (ET-) dependent pathway is involved in the up-regulation of FN-1 during diabetes that involves activation of NF-kβ and AP1 transcription factors (Moradipoor et al., 2016). Further, IL-6 (Interleukin-6), which is a potent proinflammatory cytokine, plays an essential role in DR pathogenesis. Knockout of IL-6 resulted in reduced leukocytes adhesion in retinal blood vessels and TNF-alpha level in microglial cells of retina (Rojas et al., 2011). However, one study showed that IL-6 protects muller cells from glucose toxicity, thus playing a protective role in DR (Coughlin et al., 2019). *MMP9* (Matrix Metallopeptidase 9) encodes protein that is involved in the breakdown of extracellular matrix. It is involved

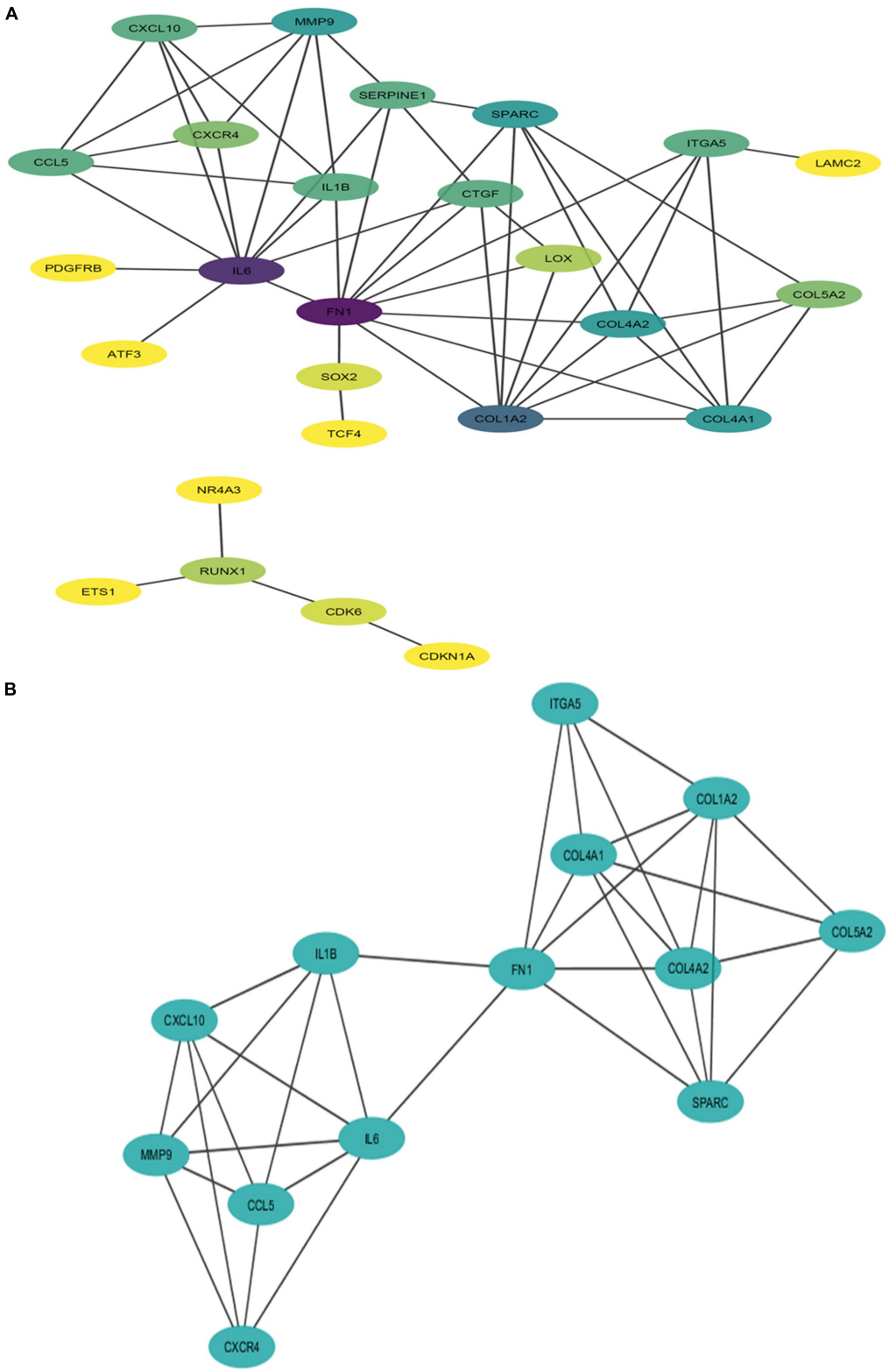



**FIGURE 6 |** Interaction network and module of potential DR candidate genes. **(A)** Protein-protein interaction (PPI) network of potential DR candidate genes: the intensity of node color denotes the degree of interactions it has with other nodes (dark purple color denotes the highest number of interactions followed by blue, light blue, green, light green, etc., and yellow denotes the lowest number of interaction). **(B)** Module (MCODE score 5.67 and number of nodes 13) obtained from PPI network of potential DR candidate genes.

in DR development and progression by accelerating apoptosis of retinal capillary cells in the early phase of DR and angiogenesis in the later phase (Kowluru et al., 2012). The level of MMP-9 differs with the stages of DR and was found to contribute more than MMP-1 in DR pathogenesis (Kwon et al., 2016). Various histone modifications, DNA methylations, and their role in metabolic memory formation (Mishra and Kowluru, 2016; Kumari et al., 2020) are reported for *MMP9* during hyperglycemic conditions. *SPARC* (Secreted Protein Acidic And Cysteine Rich), a gene that encodes cysteine-rich acidic matrix-associated protein, is also involved in the development of DR. Retinal basement membrane of Type2DM patients showing thickening and permeability changes is found to secrete the protein encoded by *SPARC* (Watanabe et al., 2009). Further, *SPARC* was also found to mediate cellular adhesion, cell migration, and angiogenesis.

Moving to genes altered in early stage of DR, except for one study on *IGSF21* (Immunoglobin Superfamily Member 21) (Lin et al., 2016), none of the genes have been studied in context to DR. However, *ROCK2* (Rho Associated Coiled-Coil Containing Protein Kinase 2) (Koch et al., 2014; Lu et al., 2020), *UHRF1* (Ubiquitin Like With PHD And Ring Finger Domains 1) (Ramesh et al., 2016), and *MAPT* (Microtubule Associated Protein Tau) (C. C. Zhang et al., 2016) are shown to be associated with neurodegeneration. As neurodegeneration has been observed as one of the earliest events in the onset of DR, these genes might be responsible for retinal pathological changes in early DR and might also play a role in metabolic memory formation. Additionally, there are so many studies done on *ROCK1* but not on *ROCK2*. However, ROCK has an essential role in the pathogenesis of DR. It affects the expression and function of adhesion molecules and its inhibitor significantly reduced this adhesion process by reducing the activation of ROCK. ROCK pathway also plays a critical role in angiogenesis (Arita et al., 2010). Moreover, abnormal ROCK pathways are responsible for various neurological disorders. In one study ROCK inhibitor was shown to increase the regeneration of retinal ganglion cell (Lingor et al., 2007). Another study showed increase in ROCKII protein level in NMDA-induced retinal neurotoxicity, and its inhibitor acted as neuroprotective agent by abolishing the increase in ROCKII level (Kitaoka et al., 2004). Apart from this, *UHRF1,* which encode a protein that regulates DNA and histone methylation, and *NR1H4* (Nuclear Receptor Subfamily 1 Group H Member 4), which encode ligand-activated transcription factor, are also found to be linked with inflammation (Fiorucci et al., 2010; Wang et al., 2018), oxidative stress (Gai et al., 2017; J. K. Kim J. K. et al., 2020), and angiogenesis (Guo and Mo, 2020). Interestingly, in the present study, *NR1H4* was also found to be one of the targets of down-regulated miRNA identified in a plasma sample of NPDR cases and thus can act as a preferred candidate in studies concerned with identification of circulatory prognostic biomarker for DR.

This study revealed many hub genes and few early genes that have the potential to act as a target in future DR research, but this study suffers from its own limitations. During the

**TABLE 4 |** Hub genes and the associated altered miRNA.

| Hub genes* | miRNA* |
|---|---|
| *FN1* | hsa-let-7g |
| *IL6* | hsa-mir-451, has-mir-106a |
| *COL1A2* | hsa-let-7g, has-mir-25, has-mir-29a |
| *COL4A1* | hsa-mir-29a |
| *COL4A2* | hsa-mir-29a |
| *SPARC* | hsa-mir-29a |
| *MMP9* | hsa-mir-451, has-mir-15b |

**All hub genes were up-regulated and their associated miRNAs were down-regulated.*

**TABLE 5 |** List of some of the enriched GO and KEGG pathways of hub genes.

| Terms | GO/KEGG pathways | FDR value |
|---|---|---|
| Biological Processes (BP) | Extracellular structure organization (GO:0043062) | 1.59E-06 |
| | Endodermal cell differentiation (GO:0035987) | 1.21E-03 |
| | Cellular response to organic substance (GO:0071310) | 1.27E-03 |
| | Cellular response to chemical stimulus (GO:0070887) | 2.88E-03 |
| | Response to organic substance (GO:0010033) | 3.11E-03 |
| | Circulatory system development (GO:0072359) | 4.70E-03 |
| | Platelet activation (GO:0030168) | 9.50E-03 |
| | Collagen-activated tyrosine kinase receptor signaling Pathway (GO:0038063) | 9.96E-03 |
| | Blood vessel development (GO:0001568) | 1.02E-02 |
| | Positive regulation of cell migration (GO:0030335) | 1.08E-02 |
| Molecular Functions (MF) | Platelet-derived growth factor binding (GO:0048407) | 7.15E-03 |
| | Extracellular matrix structural constituent conferring tensile strength (GO:0030020) | 7.24E-04 |
| | Collagen binding (GO:0005518) | 1.69E-03 |
| | Extracellular matrix structural constituent (GO:0005201) | 3.76E-06 |
| | Structural molecule activity (GO:0005198) | 1.48E-03 |
| KEGG pathways | hsa04512:ECM-receptor interaction | 0.018044733 |
| | hsa04151:PI3K-Akt signaling pathway | 0.028095395 |
| | hsa05200:Pathways in cancer | 0.047128876 |

selection of probable candidate genes in the early stage of DR (early genes), only gene methylation dataset was taken into consideration due to unavailability of NPDR samples in the gene expression (mRNA expression) dataset, and this is one of the major limitations of this study. The miRNA data was also not considered here because targets of miRNAs are diverse, which may include many unrelated genes and therefore may decrease the specificity of the identified genes to the particular context.

Further, as different datasets differ in the source of population for sample collection, and also in some cases single dataset contains sample from different parts of body, integration of these may lead to heterogeneity and affect the results. However, by maintaining the statistical significance of data, the integration of multiple datasets can be beneficial in capturing multiple molecular alterations, thus improving the prediction accuracy while the presence of diverse source of population and different regions of body can also lead to identification of potential global candidate genes for the disease.

Also, not sufficient studies are available for the identified genes with respect to DR. Furthermore, validation of the obtained results is necessary for the genes to be considered as a representative gene for DR.

## CONCLUSION

Diabetic retinopathy is a consequence of multiple altered metabolic processes, biological functions, and pathways that are linked among themselves in one or the other ways, and these alterations are in turn associated with one or more altered expression of genes.

The study identified 7 hub genes (*FN1*, *IL-6*, *COL1A2*, *COL4A1*, *COL4A2*, *SPARC*, and *MMP9*) that could play a potential role in the aggravation of DR pathogenesis. Further, some of the early genes like *NR1H4* and those participating in neurodegeneration (*ROCK2*, *UHRF1*, and *MAPT*) could be responsible for early pathological changes in DR and formation of metabolic memory and can be used as a potential prognostic biomarker and early therapeutic targets for DR.

## DATA AVAILABILITY STATEMENT

The data of gene expression profiling, gene methylation profiling, and miRNA expression profiling were obtained from Gene Expression Omnibus (GEO) datasets available at the National Center for Biotechnology Information (NCBI) (https://www.ncbi.nlm.nih.gov/gds): GSE60436, GSE57362, GSE140959, GSE1009, GSE51674, GSE80178, and GSE84971.

## AUTHOR CONTRIBUTIONS

NK, SC, and SG gave equal contribution in framing research topics, collecting study materials, performing the entire research work, analyzing the data, preparing the manuscript, proofreading, and so forth. AK wrote the introduction section of the manuscript and was involved in the data analysis. All authors read and approved the final manuscript.

## FUNDING

This study was supported by institutional grant No. P07-MLP-120 from CSIR-Indian Institute of Chemical Biology, Kolkata, India.

## ACKNOWLEDGMENTS

NK thanks CSIR-Award No: 1121732018 and AK thanks UGC-Award No: 16-[Dec-2017/2018] India for their fellowships. NK and AK thank Ishita Mukherjee (CSIR fellow) and Krishna Kumar (DBT fellow) for helping with various bioinformatics tools.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.576442/full#supplementary-material

## REFERENCES

Abdulrahim, J. W., Kwee, L. C., Grass, E., Siegler, I. C., Williams, R., Karra, R., et al. (2019). Epigenome-wide association study for all-cause mortality in a cardiovascular cohort identifies differential methylation in castor zinc finger 1 (CASZ 1). *J. Am. Heart Assoc.* 8:e013228.

Alavi, M. V., Mao, M., Pawlikowski, B. T., Kvezereli, M., Duncan, J. L., Libby, R. T., et al. (2016). Col4a1 mutations cause progressive retinal neovascular defects and retinopathy. *Sci. Rep.* 6:18602. doi: 10.1038/srep18602

Arita, R., Hata, Y., and Ishibashi, T. (2010). ROCK as a therapeutic target of diabetic retinopathy. *J. Ophthalmol.* 2010:175163. doi: 10.1155/2010/175163

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

Baelde, H. J., Eikmans, M., Doran, P. P., Lappin, D. W., de Heer, E., and Bruijn, J. A. (2004). Gene expression profiling in glomeruli from human kidneys with diabetic nephropathy. *Am. J. Kidney Dis.* 43, 636–650. doi: 10.1053/j.ajkd.2003.12.028

Barber, A. J., and Baccouche, B. (2017). Neurodegeneration in diabetic retinopathy: potential for novel therapies. *Vision Res.* 139, 82–92. doi: 10.1016/j.visres.2017.06.014

Berdasco, M., Gomez, A., Rubio, M. J., Catala-Mora, J., Zanon-Moreno, V., Lopez, M., et al. (2017). DNA methylomes reveal biological networks involved in human eye development, functions and associated disorders. *Sci. Rep.* 7:11762. doi: 10.1038/s41598-017-12084-1

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

Conserva, F., Barozzino, M., Pesce, F., Divella, C., Oranger, A., Papale, M., et al. (2019). Urinary miRNA-27b-3p and miRNA-1228-3p correlate with the progression of kidney fibrosis in diabetic nephropathy. *Sci. Rep.* 9:11357. doi: 10.1038/s41598-019-47778-1

Coughlin, B. A., Trombley, B. T., and Mohr, S. (2019). Interleukin-6 (IL-6) mediates protection against glucose toxicity in human Muller cells via activation of VEGF-A signaling. *Biochem. Biophys. Res. Commun.* 517, 227–232. doi: 10.1016/j.bbrc.2019.07.044

Curran, P. J., and Hussong, A. M. (2009). Integrative data analysis: the simultaneous analysis of multiple data sets. *Psychol. Methods* 14, 81–100. doi: 10.1037/a0015914

Drake, L. (2007). Prevention of blindness from Diabetes mellitus–report of a WHO consultation in Geneva, Switzerland, 9-11 November 2005. *Nursing Standard* 21, 30–31.

Ellis, D., Burgess, P. I., and Kayange, P. (2013). Management of diabetic retinopathy. *Malawi. Med. J.* 25, 116–120.

Fiorucci, S., Cipriani, S., Mencarelli, A., Renga, B., Distrutti, E., and Baldelli, F. (2010). Counter-regulatory role of bile acid activated receptors in immunity and inflammation. *Curr. Mol. Med.* 10, 579–595. doi: 10.2174/1566524011009060579

Fu, Y., Tang, M., Xiang, X., Liu, K., and Xu, X. (2019). Glucose affects cell viability, migration, angiogenesis and cellular adhesion of human retinal capillary endothelial cells via SPARC. *Exp. Ther. Med.* 17, 273–283. doi: 10.3892/etm.2018.6970

Gai, Z., Chu, L., Xu, Z., Song, X., Sun, D., and Kullak-Ublick, G. A. (2017). Farnesoid X receptor activation protects the kidney from ischemia-reperfusion damage. *Sci. Rep.* 7:9815. doi: 10.1038/s41598-017-10168-6

Guo, Z., and Mo, Z. (2020). [Role of UHRF1 in methylation regulation and angiogenesis]. *Zhonghua Yi Xue Yi Chuan Xue Za Zhi* 37, 200–204. doi: 10.3760/cma.j.issn.1003-9406.2020.02.025

He, J., Zheng, W., Lu, M., Yang, X., Xue, Y., and Yao, W. (2019). A controlled heat stress during late gestation affects thermoregulation, productive performance, and metabolite profiles of primiparous sow. *J. Therm. Biol.* 81, 33–40. doi: 10.1016/j.jtherbio.2019.01.011

Intine, R. V., and Sarras, M. P. Jr. (2012). Metabolic memory and chronic diabetes complications: potential role for epigenetic mechanisms. *Curr. Diab. Rep.* 12, 551–559. doi: 10.1007/s11892-012-0302-7

Ishikawa, K., Yoshida, S., Kobayashi, Y., Zhou, Y., Nakama, T., Nakao, S., et al. (2015). Microarray analysis of gene expression in fibrovascular membranes excised from patients with proliferative diabetic retinopathy. *Invest. Ophthalmol. Vis. Sci.* 56, 932–946. doi: 10.1167/iovs.14-15589

Kim, J. K., Kan, G., Mao, Y., Wu, Z., Tan, X., He, H., et al. (2020). UHRF1 downmodulation enhances antitumor effects of histone deacetylase inhibitors in retinoblastoma by augmenting oxidative stress-mediated apoptosis. *Mol. Oncol.* 14, 329–346. doi: 10.1002/1878-0261.12607

Kim, S., and Park, T. (2016). An overview of integrative analysis in cancer studies. *Integrat. Cancer Sci. Therap.* 3, 484–485. doi: 10.15761/icst.1000193

Kitaoka, Y., Kitaoka, Y., Kumai, T., Lam, T. T., Kuribayashi, K., Isenoumi, K., et al. (2004). Involvement of RhoA and possible neuroprotective effect of fasudil, a Rho kinase inhibitor, in NMDA-induced neurotoxicity in the rat retina. *Brain Res.* 1018, 111–118. doi: 10.1016/j.brainres.2004.05.070

Koch, J. C., Tonges, L., Barski, E., Michel, U., Bahr, M., and Lingor, P. (2014). ROCK2 is a major regulator of axonal degeneration, neuronal death and axonal regeneration in the CNS. *Cell Death Dis.* 5:e1225. doi: 10.1038/cddis.2014.191

Kotajima, N., Kanda, T., Yuuki, N., Kimura, T., Kishi, S., Fukumura, Y., et al. (2001). Type IV collagen serum and vitreous fluid levels in patients with diabetic retinopathy. *J. Int. Med. Res.* 29, 292–296.

Kowluru, R. A., Zhong, Q., and Santos, J. M. (2012). Matrix metalloproteinases in diabetic retinopathy: potential role of MMP-9. *Expert Opin. Investig. Drugs* 21, 797–805. doi: 10.1517/13543784.2012.681043

Kumari, N., Karmakar, A., and Ganesan, S. K. (2020). Targeting epigenetic modifications as a potential therapeutic option for diabetic retinopathy. *J. Cell Physiol.* 235, 1933–1947. doi: 10.1002/jcp.29180

Kwon, J. W., Choi, J. A., and Jee, D. (2016). Matrix metalloproteinase-1 and matrix metalloproteinase-9 in the aqueous Humor of diabetic macular edema patients. *PLoS One* 11:e0159720. doi: 10.1371/journal.pone.0159720

Liang, L., Stone, R. C., Stojadinovic, O., Ramirez, H., Pastar, I., Maione, A. G., et al. (2016). Integrative analysis of miRNA and mRNA paired expression profiling of primary fibroblast derived from diabetic foot ulcers reveals multiple impaired cellular functions. *Wound Repair. Regen* 24, 943–953. doi: 10.1111/wrr.12470

Lin, X., Wang, J., Yun, L., Jiang, S., Li, L., Chen, X., et al. (2016). Association between LEKR1-CCNL1 and IGSF21-KLHDC7A gene polymorphisms and diabetic retinopathy of type 2 diabetes mellitus in the Chinese Han population. *J. Gene Med.* 18, 282–287. doi: 10.1002/jgm.2926

Lingor, P., Teusch, N., Schwarz, K., Mueller, R., Mack, H., Bahr, M., et al. (2007). Inhibition of Rho kinase (ROCK) increases neurite outgrowth on chondroitin sulphate proteoglycan in vitro and axonal regeneration in the adult optic nerve in vivo. *J. Neurochem.* 103, 181–189. doi: 10.1111/j.1471-4159.2007.04756.x

Lu, W., Wen, J., and Chen, Z. (2020). Distinct roles of ROCK1 and ROCK2 on the cerebral ischemia injury and subsequently neurodegenerative changes. *Pharmacology* 105, 3–8. doi: 10.1159/000502914

Maag, J. L., Kaczorowski, D. C., Panja, D., Peters, T. J., Bramham, C. R., Wibrand, K., et al. (2017). Widespread promoter methylation of synaptic plasticity genes in long-term potentiation in the adult brain in vivo. *BMC Genomics* 18:250. doi: 10.1186/s12864-017-3621-x

Maghbooli, Z., Hossein-nezhad, A., Larijani, B., Amini, M., and Keshtkar, A. (2015). Global DNA methylation as a possible biomarker for diabetic retinopathy. *Diabetes Metab. Res. Rev.* 31, 183–189. doi: 10.1002/dmrr.2584

Mishra, M., and Kowluru, R. A. (2016). The role of DNA methylation in the metabolic memory phenomenon associated with the continued progression of diabetic retinopathy. *Invest. Ophthalmol. Vis. Sci.* 57, 5748–5757. doi: 10.1167/iovs.16-19759

Moradifard, S., Hoseinbeyki, M., Ganji, S. M., and Minuchehr, Z. (2018). Analysis of microRNA and gene expression profiles in Alzheimer's disease: a meta-analysis approach. *Sci. Rep.* 8:4767. doi: 10.1038/s41598-018-20959-0

Moradipoor, S., Ismail, P., Etemad, A., Wan Sulaiman, W. A., and Ahmadloo, S. (2016). Expression profiling of genes related to endothelial cells biology in patients with Type 2 Diabetes and patients with prediabetes. *Biomed. Res. Int.* 2016:1845638. doi: 10.1155/2016/1845638

Nagaraju, G. P., and Sharma, D. (2011). Anti-cancer role of SPARC, an inhibitor of adipogenesis. *Cancer Treat. Rev.* 37, 559–566. doi: 10.1016/j.ctrv.2010.12.001

Raman, A. T., Pohodich, A. E., Wan, Y. W., Yalamanchili, H. K., Lowry, W. E., Zoghbi, H. Y., et al. (2018). Apparent bias toward long gene misregulation in MeCP2 syndromes disappears after controlling for baseline variations. *Nat. Commun.* 9:3225. doi: 10.1038/s41467-018-05627-1

Ramesh, V., Bayam, E., Cernilogar, F. M., Bonapace, I. M., Schulze, M., Riemenschneider, M. J., et al. (2016). Loss of Uhrf1 in neural stem cells leads to activation of retroviral elements and delayed neurodegeneration. *Genes Dev.* 30, 2199–2212. doi: 10.1101/gad.284992.116

Ramirez, H. A., Pastar, I., Jozic, I., Stojadinovic, O., Stone, R. C., Ojeh, N., et al. (2018). *Staphylococcus aureus* triggers induction of miR-15B-5P to diminish DNA repair and deregulate inflammatory response in diabetic foot ulcers. *J. Invest. Dermatol.* 138, 1187–1196. doi: 10.1016/j.jid.2017.11.038

Rojas, M. A., Zhang, W., Xu, Z., Nguyen, D. T., Caldwell, R. W., and Caldwell, R. B. (2011). Interleukin 6 has a critical role in diabetes-induced retinal vascular inflammation and permeability. *Invest. Ophthalmol. Vis. Sci.* 52, 1003–1003.

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

Smit-McBride, Z., Nguyen, A. T., Yu, A. K., Modjtahedi, S. P., Hunter, A. A., Rashid, S., et al. (2020). Unique molecular signatures of microRNAs in ocular fluids and plasma in diabetic retinopathy. *PLoS One* 15:e0235541. doi: 10.1371/journal.pone.0235541

Tarca, A. L., Romero, R., and Draghici, S. (2006). Analysis of microarray experiments of gene expression profiling. *Am. J. Obstet. Gynecol.* 195, 373–388. doi: 10.1016/j.ajog.2006.07.001

Vujosevic, S., Muraca, A., Alkabes, M., Villani, E., Cavarzeran, F., Rossetti, L., et al. (2019). Early microvascular and neural changes in patients with Type 1 and Type 2 diabetes mellitus without clinical signs of diabetic retinopathy. *Retina* 39, 435–445. doi: 10.1097/IAE.0000000000001990

Wang, B. C., Lin, G. H., Wang, B., Yan, M., He, B., Zhang, W., et al. (2018). UHRF1 suppression promotes cell differentiation and reduces inflammatory reaction in anaplastic thyroid cancer. *Oncotarget* 9, 31945–31957. doi: 10.18632/oncotarget.10674

Watanabe, K., Okamoto, F., Yokoo, T., Iida, K. T., Suzuki, H., Shimano, H., et al. (2009). SPARC is a major secretory gene expressed and involved in the

development of proliferative diabetic retinopathy. *J. Atheroscler. Thromb.* 16, 69–76. doi: 10.5551/jat.e711

Wong, T. Y., Cheung, C. M., Larsen, M., Sharma, S., and Simo, R. (2016). Diabetic retinopathy. *Nat. Rev. Dis. Primers* 2:16012. doi: 10.1038/nrdp.2016.12

Yang, W., Rosenstiel, P., and Schulenburg, H. (2019). aFold – using polynomial uncertainty modelling for differential gene expression estimation from RNA sequencing data. *BMC Genomics* 20:364. doi: 10.1186/s12864-019-5686-1

Yau, J. W., Rogers, S. L., Kawasaki, R., Lamoureux, E. L., Kowalski, J. W., Bek, T., et al. (2012). Global prevalence and major risk factors of diabetic retinopathy. *Diabetes Care* 35, 556–564.

Zhang, C. C., Xing, A., Tan, M. S., Tan, L., and Yu, J. T. (2016). The role of MAPT in neurodegenerative diseases: genetics mechanisms and therapy. *Mol. Neurobiol.* 53, 4893–4904. doi: 10.1007/s12035-015-9415-8

Zhang, X., Zhao, L., Hambly, B., Bao, S., and Wang, K. (2017). Diabetic retinopathy: reversibility of epigenetic modifications and new therapeutic targets. *Cell Biosci.* 7:42. doi: 10.1186/s13578-017-0167-1

**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 Neuroscience

BRIEF RESEARCH REPORT
published: 25 November 2020
doi: 10.3389/fnins.2020.591196

# Anticonvulsants and Chromatin-Genes Expression: A Systems Biology Investigation

*Thayne Woycinck Kowalski[1,2,3,4,5,6,7]*, Julia do Amaral Gomes[1,2,3,4,5], Mariléa Furtado Feira[1,2,3,4], Ágata de Vargas Dupont[2,4], Mariana Recamonde-Mendoza[7,8] and Fernanda Sales Luiz Vianna[1,2,3,4,5]**

[1] Postgraduation Program in Genetics and Molecular Biology, Universidade Federal do Rio Grande do Sul (UFRGS), Porto Alegre, Brazil, [2] Laboratory of Immunobiology and Immunogenetics, Universidade Federal do Rio Grande do Sul (UFRGS), Porto Alegre, Brazil, [3] National Institute of Population Medical Genetics (INAGEMP), Porto Alegre, Brazil, [4] Genomic Medicine Laboratory, Hospital de Clínicas de Porto Alegre (HCPA), Porto Alegre, Brazil, [5] National System of Information on Teratogenic Agents (SIAT), Medical Genetics Service, Hospital de Clínicas de Porto Alegre (HCPA), Porto Alegre, Brazil, [6] Centro Universitário CESUCA, Cachoeirinha, Brazil, [7] Bioinformatics Core, Hospital de Clínicas de Porto Alegre (HCPA), Porto Alegre, Brazil, [8] Institute of Informatics, Universidade Federal do Rio Grande do Sul (UFRGS), Porto Alegre, Brazil

OPEN ACCESS

**Edited by:**
*Sanjay Kumar Banerjee, National Institute of Pharmaceutical Education and Research (Guwahati), India*

**Reviewed by:**
*Hugo Tovar, Instituto Nacional de Medicina Genómica (INMEGEN), Mexico*
*Lorena Aguilar Arnal, National Autonomous University of Mexico, Mexico*

***Correspondence:**
*Thayne Woycinck Kowalski*
*thaynewk@gmail.com*
*Fernanda Sales Luiz Vianna*
*fslvianna@gmail.com*

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

**Received:** *04 August 2020*
**Accepted:** *27 October 2020*
**Published:** *25 November 2020*

**Citation:**
*Kowalski TW, Gomes JA, Feira MF, Dupont ÁV, Recamonde-Mendoza M and Vianna FSL (2020) Anticonvulsants and Chromatin-Genes Expression: A Systems Biology Investigation. Front. Neurosci. 14:591196. doi: 10.3389/fnins.2020.591196*

Embryofetal development is a critical process that needs a strict epigenetic control, however, perturbations in this balance might lead to the occurrence of congenital anomalies. It is known that anticonvulsants potentially affect epigenetics-related genes, however, it is not comprehended whether this unbalance could explain the anticonvulsants-induced fetal syndromes. In the present study, we aimed to evaluate the expression of epigenetics-related genes in valproic acid, carbamazepine, or phenytoin exposure. We selected these three anticonvulsants exposure assays, which used murine or human embryonic stem-cells and were publicly available in genomic databases. We performed a differential gene expression (DGE) and weighted gene co-expression network analysis (WGCNA), focusing on epigenetics-related genes. Few epigenetics genes were differentially expressed in the anticonvulsants' exposure, however, the WGCNA strategy demonstrated a high enrichment of chromatin remodeling genes for the three drugs. We also identified an association of 46 genes related to Fetal Valproate Syndrome, containing *SMARCA2* and *SMARCA4,* and nine genes to Fetal Hydantoin Syndrome, including *PAX6, NEUROD1,* and *TSHZ1*. The evaluation of stem-cells under drug exposure can bring many insights to understand the drug-induced damage to the embryofetal development. The candidate genes here presented are potential biomarkers that could help in future strategies for the prevention of congenital anomalies.

**Keywords: WGCNA, epigenetics, antiepileptics, teratogen, valproic acid, phenytoin, fetal hydantoin syndrome, fetal valproate syndrome**

## INTRODUCTION

Embryogenesis is a stepwise controlled process, which requires specific gene expression orchestrated by signaling networks (Rape, 2017). During the embryo development, epigenetics modifications are essential for the correct expression of these highly orchestrated genes, hence enabling the transition from pluripotent stem-cells until its final differentiation state (Rape, 2017; Jambhekar et al., 2019).

The lack of proper chromatin modifications can be lethal during embryogenesis or lead to the occurrence of congenital anomalies (Jambhekar et al., 2019). Embryogenesis failures can be caused by genetic factors or external stimuli, named teratogens (De Santis et al., 2004; Worley et al., 2018). According to epidemiologic studies, it is believed that teratogens cause 10–15% of all the congenital anomalies (Gilbert-Barness, 2010); however, there are many barriers in regard to the proper teratogen identification and understanding of its molecular mechanisms.

Few studies have assessed the potential of a teratogenic drug disrupting epigenetics mechanisms, being these studies restricted especially to alcohol and valproic acid use during pregnancy, and their induced histone hyperacetylation (Tung and Winn, 2010; Gupta et al., 2016; Mazzu-Nascimento et al., 2017). On the other hand, assays in embryonic stem-cells are constantly used in the developmental toxicity field, providing a better comprehension of the drug-induced perturbation in development (Worley et al., 2018; Leigh et al., 2020). These perturbations could be assessed by evaluating how these proteins interact with each other in a biological network, which is systems biology field of research.

From a systems biology perspective, these gene expression perturbations could be identified by network and co-expression analyses, helping to hypothesize which epigenetics mechanisms are teratogen-affected.

Hence, the aim of the present study is to evaluate the effect of anticonvulsant drugs, known for their teratogenic effects, in the expression of epigenetics machinery genes. For its accomplishment, we performed a secondary expression analysis in murine or human embryonic stem-cells (mESC and hESC) exposed to these drugs, and evaluated the results through systems biology strategies, especially the weighted gene correlation network analysis (WGCNA). WGCNA is a consolidated screening method to identify biomarker candidates or therapeutic targets; it associates gene expression and external traits to identify modules of highly correlated genes (Langfelder and Horvath, 2008). Finally, we hypothesized the main genes and epigenetics mechanisms that might be perturbed in these teratogens' exposure.

# METHODS

## Teratogens Selection

Careful literature research was performed to select only drugs with proven teratogenic effects in the human embryo or fetus, and with established animal models. These molecules were named major teratogens and assessed in the DrugBank database to obtain its pharmaceutical class and variant names. Anticonvulsants were the chosen class of study by convenience, according to the availability of genomic expression assays.

## Bioinformatics Analysis

Gene expression studies were obtained through research mechanisms in the ArrayExpress and Gene Expression Omnibus (GEO) databases, using the name of the drugs selected in the search mechanism. Filters were applied to select only exposure studies in murine or human embryonic stem-cells (mESC or hESC). Despite only microarray studies being selected, RNA-seq assays were also considered.

Differential gene expression analysis was performed in the R v.3.6.2, applying robust multiaverage (RMA) normalization, and using the *affy* and *limma* packages. The following comparisons were executed: valproic acid, carbamazepine, phenytoin, methotrexate, and warfarin exposure assays were set against unexposed stem-cells; mESC and hESC selected assays were evaluated separately. All the genes with logFC $> 1.5$ and adjusted *P*-value for false discovery rate (FDR) $< 0.05$ were considered upregulated; logFC $< -1.5$ and the same adjusted *P*-value for FDR were set as parameters for the downregulated genes.

Gene ontologies and Reactome enrichment analysis were also performed in the R v.3.6.2, using the *clusterprofileR* package, considering only significantly enriched ontologies or pathways (FDR $< 0.05$). Orthologs assessment was performed using the *BiomaRt* package. Only orthologs of high confidence were included, according to the Ensembl Orthology Quality Contro[1].

Human Phenotype Ontology (HPO) database was assessed in the link[2]. Venn diagrams were performed in the Bioinformatics and Evolutionary Genomics webtool, from the Ghent University[3].

## Systems Biology Analysis

Weighted gene correlation network analysis (WGCNA) was performed in the R v.3.6.2 with the homonym package; as in DGE analysis, mESC, and hESC datasets were evaluated separately. Data heterogeneity included differences in dose and time of exposure for the mESC studies, hence a consensus analysis was used, as recommended in the WGCNA package tutorials. The 20% probes with larger expression variance were included in the analysis. A thresholding power was set in 12, according to the topology of the data. Default minimum and maximum module sizes were used, comprising of at least 30 and maximum of 3,000 genes per module. Gene significance was set in 0.1. This measure helps to obtain the biologically relevant genes. We selected 0.1 as a threshold value because this is an exploratory study, hence we wanted to collect all the biologically relevant genes. Further phenotype-associated genes were still to be filtered, what would also help to reduce any noise (non-relevant genes).

More information about the WGCNA parameters can be encountered in the tutorials provided by the developers[4].

Protein-protein interaction networks were generated with the STRING v.11 database webtool, comprising only query proteins, and with a minimum required interaction score of 0.4 (default). The confidence score is a probability that evaluates whether the proteins are included in the same metabolic pathway (von Mering et al., 2005). The medium score we selected might include false positives, therefore, we filtered for experimental data only, to exclude computational predicted interactions. Further network

[1]https://www.ensembl.org/info/genome/compara/Ortholog_qc_manual.html
[2]https://hpo.jax.org/app/
[3]http://bioinformatics.psb.ugent.be/webtools/Venn/
[4]https://horvath.genetics.ucla.edu/html/CoexpressionNetwork/Rpackages/WGCNA

statistics were performed in the Cytoscape v.3.7.2. Considering the size of the network, we performed global centrality analysis, evaluating betweenness centrality as the size and closeness centrality as the color of the nodes. The DyNet v. 1.0.0 Cytoscape application was used for network comparison, using a *prefuse force directed layout* in the network combinations for the WGCNA results and HPO data for the teratogenic syndromes.

# RESULTS

## Teratogens, Expression Datasets Selection, and Epigenetics Genes

We searched for gene expression studies in stem-cells exposed to 28 different teratogens. After a careful evaluation, the anticonvulsants valproic acid, carbamazepine, and phenytoin were chosen for expression and systems biology analysis; the three drugs are folic acid antagonists (Matok et al., 2009). For comparison purposes, methotrexate and warfarin were selected. Methotrexate is an antineoplastic agent and a folic acid antagonist, whilst warfarin is an anticoagulant (De Santis et al., 2004). **Supplementary Table 1** comprises the phenotypical spectrum of the embryopathies induced by the drug selected.

Four studies were selected for gene expression and systems biology analysis: E-MTAB-300, E-TABM-1205, and E-TABM-1216 (van Dartel et al., 2011; Theunissen et al., 2012, 2013), from ArrayExpress database (European Bioinformatics Institute, EBI), and GSE64123 (Schulpen et al., 2015) from the Gene Expression Omnibus (GEO) database (National Center of Biotechnology and Information, NCBI). The studies comprised assays evaluating valproic acid ($n$ = 32), carbamazepine ($n$ = 24), and phenytoin ($n$ = 16) exposure in mESC, being all performed in the same platform, of the same laboratory. For comparison purposes, methotrexate ($n$ = 8), warfarin ($n$ = 8), and non-exposed cells ($n$ = 13) were also used in the analysis. Separately, one assay of valproic acid ($n$ = 28) or carbamazepine ($n$ = 26) in hESC was evaluated and compared to unexposed cells ($n$ = 27). The studies selected were all microarray assays from the Affymetrix platforms (Thermo Fisher Scientific, United States). Full characteristics of the assays are available in the ArrayExpress and GEO databases.

In this study, we aimed to focus only in the expression effects on the epigenetics machinery genes. Hence, we performed a Gene Ontology (GO) research, to select all the genes that might be relevant in this scenario. We encountered 593 ontologies related to epigenetics mechanisms (**Supplementary Table 2**) that were used to filter the epigenetics genes after the DGE and WGCNA analyses were completed. This selection provided 2,091 *Homo sapiens* genes and 1,918 *Mus musculus* genes (**Supplementary Table 3**).

The diagram available in **Figure 1** demonstrates the gene filters applied in the following bioinformatics and systems biology analysis.

## Differential Gene Expression Analysis

Despite the epigenetics machinery genes being restrictedly regulated, we evaluated whether the selected teratogens could influence in their gene expression. We evaluated each dataset separated by concentration, time of exposure, and teratogen. Similar results to the ones already published by the group that performed the primary analysis in these datasets were encountered (Theunissen et al., 2012, 2013; Schulpen et al., 2015). Hence, we do not present it. Then, we joined the samples of cells exposed to different concentrations in different time-points for a same drug. This union was especially with the intention of evaluating which epigenetics-related genes are deregulated, independently of the concentration and time of exposure. We compared the differentially expressed genes to the epigenetics-related ones selected with the GO analysis.

Using the epigenetics machinery genes filter for the mESC assays, only the genes *Tshz1* and *Pax6* were upregulated in phenytoin or valproic acid exposure; however, both were downregulated in carbamazepine, methotrexate, or warfarin exposure. An opposite effect was seen for *Eomes* gene, which was downregulated when in exposure of valproic acid or phenytoin, and upregulated after carbamazepine, warfarin, or methotrexate treatment. *Lef1* and *Meis1* also had discordant results between the teratogens. **Supplementary Table 4** comprises a complete list of the logFC values, GO, and these genes' main functions, which were identified as mainly related to chromatin binding (GO:0003682).

When evaluating the hESC study, only one gene related to epigenetics mechanisms was downregulated in valproic acid exposure, and eight were upregulated. In carbamazepine treatment, only two downregulated genes were identified. None were in common between both drugs. **Supplementary Table 5** comprises the main characteristics for the genes differentially expressed in the hESC exposure assay.

In summary, few epigenetics-related genes were differentially expressed in the anticonvulsants' exposure. However, it was not possible to confirm whether these genes correlated expression was also unaffected. Hence, to perform a co-expression evaluation, we proceeded with the WGCNA analysis.

## Weighted Gene Correlation Network Analysis (WGCNA)

WGCNA analysis was applied to better comprehend which genes of the epigenetics machinery are mostly affected by the chosen drugs.

According to its developers, WGCNA can only be applied in sets with a high number of samples (preferentially above 15), hence methotrexate and warfarin were excluded from this analysis. All the samples used were of the fourth day after exposure to different drug concentrations (**Supplementary Table 6**), and they were all retrieved from E-TABM-1205 and E-TABM-1216. As in the DGE analysis, we wanted to verify the variable co-expression when using different concentrations of these drugs. According to the suggestions given by the WGCNA package developers, a variance filter was applied to select only the genes with higher deviation from the genes' mean expression.

**Supplementary Figure 1** graphically represents the filters applied in the mESC datasets, until we obtained a final list of genes related to epigenetic mechanisms. First, a variance filter was applied in mESC assays and provided 7,588 probes for WGCNA

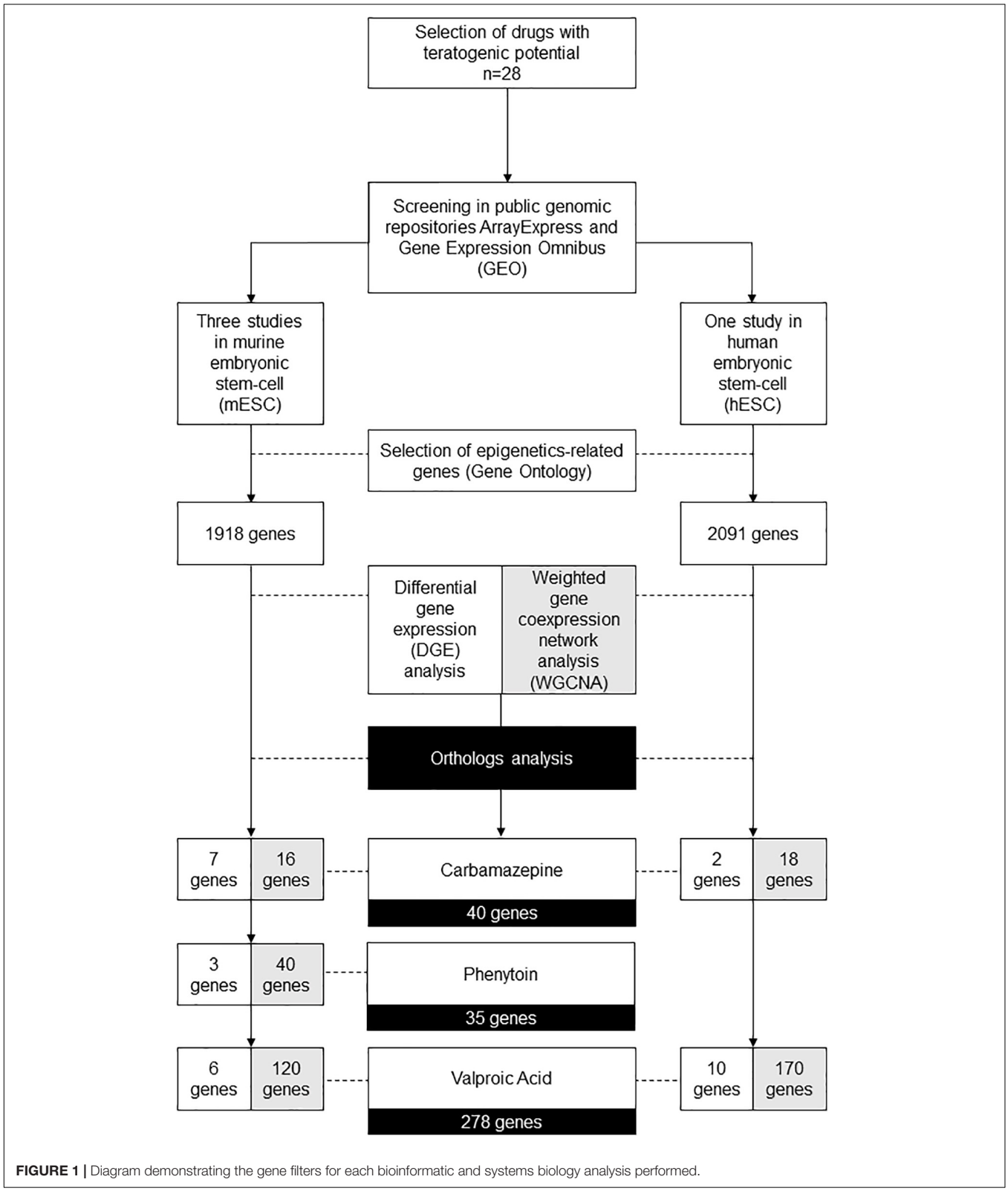



**FIGURE 1 |** Diagram demonstrating the gene filters for each bioinformatic and systems biology analysis performed.

analysis. The following modules were encountered for each anticonvulsant: eight for carbamazepine, 13 for phenytoin, and nine for valproic acid. Second, we evaluated the highly significant modules, considering only the modules with a gene ratio of at least 0.1; this cutoff implies all the modules selected had a high ratio of clustered genes that might be associated to phenotypical

traits. Hence, of the 7,588 probes with greatest variance, it was possible to identify one highly significant module for valproic acid, containing 1,124 genes, and two modules for carbamazepine and phenytoin each. Carbamazepine modules contained 113 and 50 genes, whilst phenytoin modules had both 227 genes included. Cluster dendrogram is available in **Supplementary Figure 2**.

Finally, we evaluated the genes enriched in these modules, filtering only for the ones that were related with epigenetics mechanisms, according to the list we previously obtained, which is available in **Supplementary Table 3**. In regard to this GO analysis, 120 probes of the valproic acid significant module were related to epigenetics mechanisms, 40 from the two significant phenytoin modules and 16 from the carbamazepine ones had epigenetics role. *Prdm14* was the hub gene for one of the significant modules presented in phenytoin exposure; the other modules did not have epigenetics gene as their main or most connected hub.

The same process was applied in the hESC exposure assay, and is graphically represented in **Supplementary Figure 3**. First, WGCNA analysis was performed in the 10,943 probes with greater variance. The dendrogram with the hierarchical clustering method applied by WGCNA is available in **Supplementary Figure 4**. Second, there were 217 and 3,776 genes presented in significant modules for carbamazepine and valproic acid, respectively. These genes were presented in three significant modules of the 13 identified, when evaluating the cells with carbamazepine exposure, and four significant modules in eight were identified, when in valproic acid treatment. Finally, we filtered for the genes related to epigenetics mechanisms, and obtained a list of eighteen probes of the carbamazepine assay, and 170 probes of the valproic acid treatment. One of the significant modules for valproic acid has an epigenetics machinery gene as its main hub: *ZMYND11*.

In summary, at the end of WGCNA, it is possible to identify modules of highly clustered genes, that might have an association with phenotypical traits, or even experimental conditions. For all the modules identified, we selected only the epigenetics related genes. From this filtered list, we performed an ortholog analysis to obtain a final list of genes for the three drugs evaluated, also comprising the genes identified in differential expression analysis. Hence, at the end of these analyses, 278 genes for valproic acid, 40 genes for carbamazepine, and 35 genes for phenytoin (**Figure 1**). The complete list of the genes is available in **Supplementary Table 7**.

After achieving a final list of genes, we aimed to evaluate its association to the clinical traits by performing network statistics analysis and evaluating gene-phenotype association.

## Network Statistics

Besides the WGCNA analysis, other network statistics were applied to evaluate the main characteristics of the genes identified in the previous step. Our aim was to verify whether the epigenetics mechanisms deregulated by the different teratogens were similar for each drug.

When evaluating carbamazepine and phenytoin candidate genes, it was not possible to assemble a protein-protein interaction network, probably because of the small number of genes selected. Hence, a valproic acid network was generated, and compared to a network containing all the genes selected for the three drugs (**Figure 2A**).

Network statistics analysis was also performed for valproic acid, evaluating betweenness and closeness centrality to identify the genes with bigger information flow (**Figure 2B**). *CREBBP* was the gene with the bigger betweenness centrality, although we highlight the chromatin remodeling genes (*SMARCA4, SMARCA2, SMARCD1, and SMARCD3*) in the center of the network.

To verify what are the main epigenetics mechanisms associated to the selected genes, we performed another GO analysis, and evaluated the main pathways they were included, according to the Reactome database. Significantly enriched GO and Reactome pathways can be assessed in **Figures 3A–F**. Besides epigenetics pathway ontologies, many embryo development ontologies were also enriched.

Therefore, after network statistics evaluation, chromatin remodeling was the main epigenetic mechanism suggested to be affected in the anticonvulsants' exposure. The following analysis was intended to understand if these genes might have a role in the phenotypical spectrum of these teratogens-induced embryopathies.

## Gene-Phenotype Associations

Systems biology analyses provided several epigenetic genes potentially deregulated by the anticonvulsant drugs here evaluated. To better comprehend how these genes could also influence in the teratogenic potential of these drugs, we evaluated the phenotypical spectrum of the embryopathies caused by these teratogens, as comprised in **Supplementary Table 1**.

To assess the gene-phenotype association, Human Phenotype Ontology (HPO) database was used. Carbamazepine teratogenesis is not registered in this repository, however Fetal Valproate Syndrome (ORPHA 1906) and Fetal Hydantoin Syndrome (ORPHA 1912) phenotypes caused by valproic acid and phenytoin, respectively, were annotated.

A comparison between the genes associated for each phenotype was executed against the list of candidate genes obtained through the systems biology analyses here executed. Valproic acid evaluation provided 46 genes in common, between HPO and 278 epigenetic genes we selected, including chromatin remodeling genes *SMARCA2* and *SMARCA4*, and *CREBBP*, with the bigger value of betweenness centrality in the network statistical analysis (**Figure 4**). *KMT2A* and *SMC1A* were associated to five phenotypes, each. For phenytoin, of the 35 selected genes, nine were registered in the HPO database, including genes identified in the differential expression analyses, such as *PAX6*, *NEUROD1,* and *TSHZ1*. *PAX6* was associated to eight phenotypes. The complete list of genes associated to HPO phenotypes is available in **Supplementary Table 8**.

# DISCUSSION

The present study aimed to investigate the effect of valproic acid, carbamazepine, and phenytoin in the expression of genes with

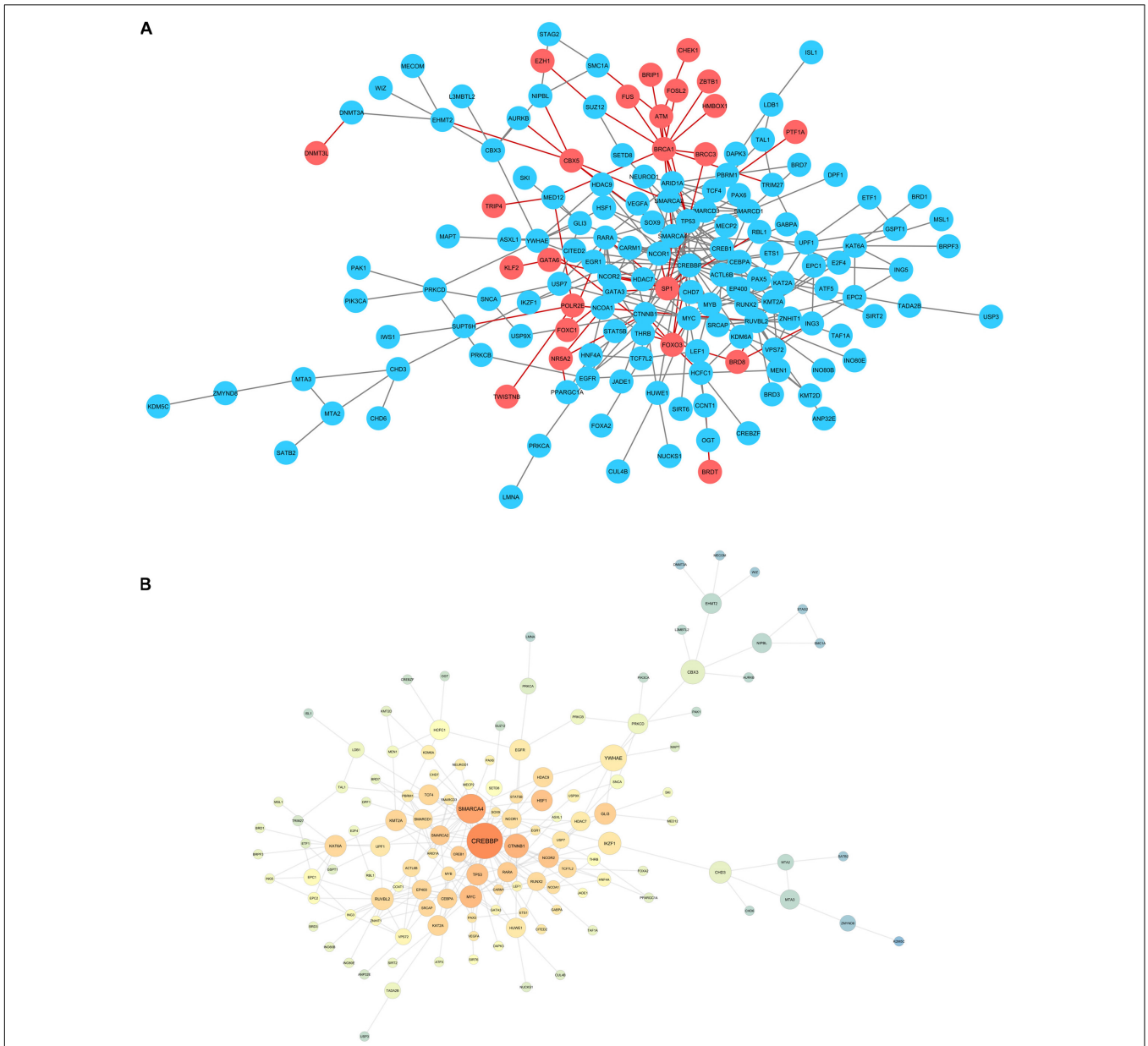



**FIGURE 2 | (A)** Network for the candidate genes encountered for valproic acid (blue), compared to the genes obtained in carbamazepine and phenytoin evaluations. **(B)** Network statistics for valproic acid selected genes. Warm colors: high closeness centrality score. Node size: big nodes for genes with high betweenness centrality score.

epigenetic-related mechanisms. This objective was accomplished by performing a careful systems biology evaluation using assays available in public genomic repositories and suggesting possible candidate genes for future researches.

With the differential gene expression analysis combined with WGCNA, 278 epigenetics genes were associated to valproic acid exposure, 40 to carbamazepine, and 35 to phenytoin. Combining HPO database evaluation, 46 epigenetics-related genes were associated to Fetal Valproate Syndrome and nine to Fetal Hydantoin Syndrome. The elevated number of "chromatin remodeling" enriched Gene Ontologies suggest this mechanism as a relevant mechanism for teratogenic disruption, which must be further evaluated in developmental toxicity assays.

The anticonvulsants here evaluated are known as neuroteratogens, because they might affect brain development, especially in second trimester, the period of continuous growth and maturation of the human brain (Ornoy, 2006; Tomson et al., 2019). Fetus exposed to these drugs—carbamazepine, phenytoin, and valproic acid—might present major congenital anomalies (especially craniofacial ones) and development delay related to this exposure, even in monotherapy (Ornoy, 2006; Tomson et al., 2019). Despite the similarities regarding the

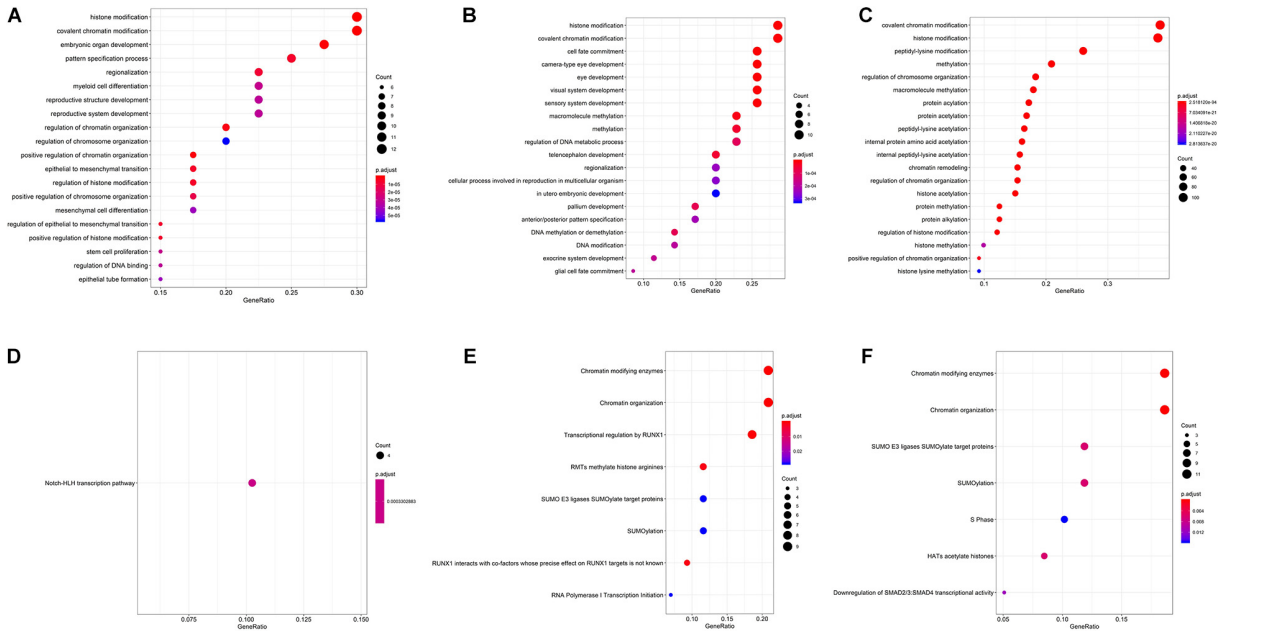



**FIGURE 3 |** Gene ontologies enrichment for carbamazepine **(A)**, phenytoin **(B)**, and valproic acid **(C)** selected genes, and Reactome database enriched pathways for carbamazepine **(D)**, phenytoin **(E)**, and valproic acid **(F)** drugs.

therapeutic effects, the dysmorphic features for each syndrome is very specific. This pattern might be explained by distinct molecular mechanisms for each drug, which may cause a dissimilar biological perturbation in the brain development. Hence, the main purpose of performing the WGCNA analysis was to identify these perturbations, by assessing potential biomarkers and candidate genes that might help in the comprehension of the phenotypical spectrum for each syndrome. This is the main goal of the WGCNA package, as proposed by its developers (Langfelder and Horvath, 2008).

It is well established that maternal exposure to different agents trigger epigenetic mechanisms, altering the gene expression and, consequently, impairing the embryofetal development (Salilew-Wondim et al., 2014). Despite that, few studies have evaluated the potential epigenetic disruption led by a teratogen exposure; recreational drugs such as alcohol and tobacco have been mainly assessed, as well as maternal infections (Banik et al., 2017; Chang et al., 2019). However, few drugs have been evaluated about these same mechanisms. It is estimated 90% of the women take at least one medication during pregnancy (Mitchell et al., 2011). Hence, the present study is an exploratory assay which attempts to fill these gaps in the understanding of teratogenesis and epigenetics linkage.

Together with antineoplastic agents, valproic acid has been one of the few drugs with a proposed epigenetic mechanism of teratogenesis (Mazzu-Nascimento et al., 2017). Valproic acid is a potent inhibitor of the histone deacetylase enzymes (HDAC), hence promoting an increased level of these proteins' acetylation. Other studies have demonstrated valproic acid also demethylates DNA (Milutinovic et al., 2007), what might be linked to its role as a folic acid antagonist. Here, we identified not only valproic acid association to DNA methylation and histone acetylation mechanisms, but also to chromatin remodeling genes. These candidate genes proposed are potentially important for the understanding not only of Fetal Valproate Syndrome, but also of other neurodevelopment disorders. Valproic acid is a known inducer of autism in rodent models (Nicolini and Fahnestock, 2018). Some researches point to histone acetylation alterations as associated to neurogenesis impairment, leading to postnatal autistic-like behaviors (Contestabile and Sintoni, 2013), although other epigenetics mechanisms must also be further investigated.

The ortholog analysis also brought potential candidates for Fetal Hydantoin Syndrome. Despite teratogenesis outcomes being variable between species, the mechanisms of many congenital anomalies have been suggested after extensive animal model assays (Shahbazi and Zernicka-Goetz, 2018). The candidates for phenytoin embryopathy, however, must be carefully evaluated before being extrapolated. Some of the genes we encountered have stablished roles in epigenetics mechanisms. *Prdm14* was the hub gene in WGCNA analysis in mESC cells with phenytoin exposure. The encoded protein is a transcription regulator with a consolidated role in pluripotency and epigenome establishment, especially wide DNA demethylation, in mESC. In hESC, its role is more associated with pluripotency regulation (Nakaki and Saitou, 2014). This gene has not been registered in HPO, neither in the Online Mendelian Inheritance in Man (OMIM) database; therefore, its role in genetic syndromes or teratogenic-induced embryopathies must be further investigated.

For the WGCNA analysis in hESC, one of the main hubs identified was *ZMYND11*. It is responsible for the reading of the histone H3K36 trimethylation, specific for H3.3, a H3 histone variant (Wen et al., 2014). This mechanism has

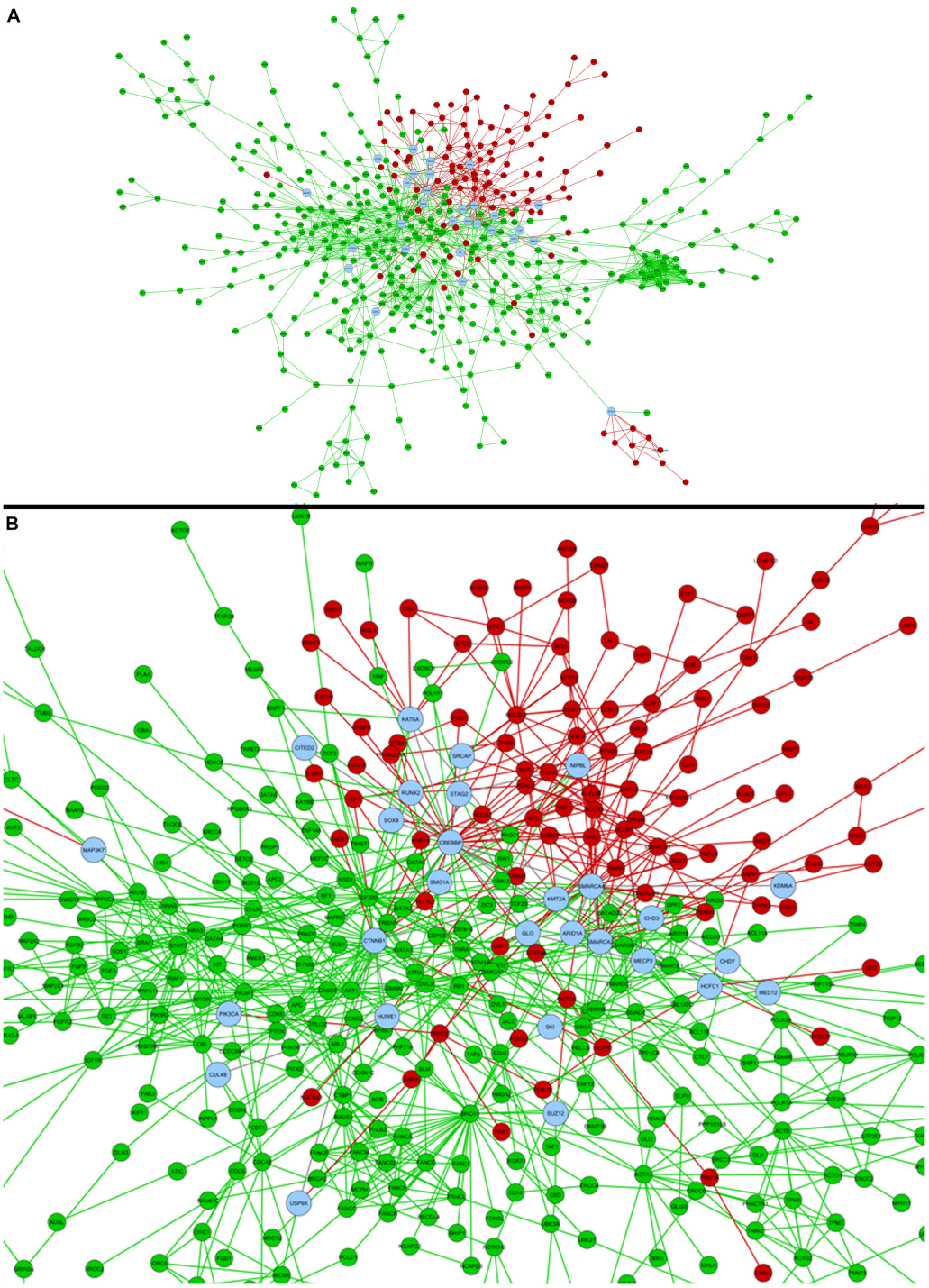



**FIGURE 4 | (A)** Comparison of valproic acid candidate genes obtained in the present study (red) and former HPO database registered genes for Fetal Valproate Syndrome (green). Common genes between both strategies are represented in blue, which can be better visualized in the zoom in **(B)**.

been evaluated in tumor suppression, but not in embryo development (Wen et al., 2014). Nevertheless, *ZMYND11* is registered in HPO, being associated with intellectual disability and facial dysmorphisms. These are common phenotypes in Mendelian disorders related to the epigenetics machinery genes (Bjornsson, 2015). Teratogenic-induced malformations are phenotypically similar to the ones caused by genetic syndromes, named phenocopies (Cassina et al., 2017). Therefore, genes like *ZMYND11*, already associated to genetic syndromes, are good candidates for the understanding of teratogenic embryopathies.

Our study lacks proper validation of the candidate genes proposed. Nevertheless, we highlight this as an exploratory research that used only previously validated experimental data for the systems biology and bioinformatics analysis. Much time and effort can be saved by conducting previous hypotheses-generator studies, targeting for biologically relevant genes or proteins (Kowalski et al., 2019). Systems biology is a feasible area for these strategies, due to its integrative and holistic characteristic (Hood et al., 2008; Le Novère, 2015). Notwithstanding, valproic acid was not only the target of our study, but also a marker of the analysis. The identification of histone acetylation genes included in significant modules for valproic acid exposure, was an incidental marker that the method was correctly applied in this investigation.

One of the strong points of our study is that many chromatin remodelers were indicated as good candidates for Fetal Valproate Syndrome understanding, including for genes of the *SMARCA* subgroup, part of the SWI1/SNF1 family (Pulice and Kadoch, 2016). These complexes enable chromatin accessibility by providing a dynamic control in an ATP-dependent mechanism (Clapier and Cairns, 2009). *SMARCA*-deficiencies are associated to several malignancies and birth defects; its homozygous loss lead to embryo lethality (Pulice and Kadoch, 2016). Valproic acid is known to alter the expression of *SMARCA4* and *SMARCD1* in neuroblastoma cells (Hu et al., 2020), and *SMARCA* genes are suggested as members of the neurogenic transcriptional network control (Higgins et al., 2019). Hence, valproate-induced perturbances in *SMARCA* genes might be sufficiently disruptive to explain Fetal Valproate Syndrome.

Finally, it has been hypothesized the understanding of epigenetic mechanisms of teratogenesis could be later used in primary prevention of congenital anomalies (Martínez-Frías, 2010). To its accomplishment, it is necessary to better comprehend these drugs' effects in chromatin during embryo development. This study was a first step in this investigation, which in future might help counseling many women who need to use these drugs during pregnancy.

## DATA AVAILABILITY STATEMENT

All data studied is available in the ArrayExpress database (European Bioinformatics Institute, EBI), under codes: E-MTAB-300, E-TABM-1205, and E-TABM-1216 (Theunissen et al., 2012, 2013), and in Gene Expression Omnibus (GEO) database (National Center of Biotechnology and Information, NCBI), series GSE64123 (Schulpen et al., 2015).

## ETHICS STATEMENT

All the genomic data used in the present research are publicly available in the ArrayExpress and Gene Expression Omnibus databases. There was no use of privileged information or data in this study.

## AUTHOR CONTRIBUTIONS

TK contributed in devising the concept, designing and conducting the analyses, and writing the manuscript. JG contributed in devising the concept, designing the experiment, and performing the analyses. MF and ÁD contributed in performing the analyses. MR-M contributed in devising and supervising the analyses. FV contributed in devising the concept, designing the experiments, supervising the analyses, and correcting the manuscript. All authors discussed the results and contributed scientifically to the manuscript.

## FUNDING

The authors would like to acknowledge the financial support: INAGEMP (National Institute of Population Medical Genetics; Grant CNPq 573993/2008-4, 465549/2014-4, FAPERGS 17/2551.0000521-0 and CAPES), FIPE/HCPA (GPPG #20170248, GPPG #20190792), CAPES (Coordination of Improvement of Higher Education Personnel, Grant 88881.132344/2016-01), and CNPq (National Council of Scientific and Technologic Development, Grant 312993/2017-0, and 156158/2018-3).

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fnins.2020.591196/full#supplementary-material

## REFERENCES

Banik, A., Kandilya, D., Ramya, S., Stünkel, W., Chong, Y. S., and Dheen, S. T. (2017). Maternal factors that induce epigenetic changes contribute to neurological disorders in offspring. *Genes* 8:150. doi: 10.3390/genes8060150

Bjornsson, H. T. (2015). The Mendelian disorders of the epigenetic machinery. *Genome Res.* 25, 1473–1481. doi: 10.1101/gr.190629.115

Cassina, M., Cagnoli, G. A., Zuccarello, D., Di Gianantonio, E., and Clementi, M. (2017). Human teratogens and genetic phenocopies. Understanding pathogenesis through human genes mutation. *Eur. J. Med. Genet.* 60, 22–31. doi: 10.1016/j.ejmg.2016.09.011

Chang, R. C., Wang, H., Bedi, Y., and Golding, M. C. (2019). Preconception paternal alcohol exposure exerts sex-specific effects on offspring growth and long-term metabolic programming. *Epigenet. Chromatin* 12:9.

Clapier, C. R., and Cairns, B. R. (2009). The biology of chromatin remodeling complexes. *Annu. Rev. Biochem.* 78, 273–304. doi: 10.1146/annurev.biochem.77.062706.153223

Contestabile, A., and Sintoni, S. (2013). Histone acetylation in neurodevelopment. *Curr. Pharm. Des.* 19, 5043–5050. doi: 10.2174/1381612811319280003

De Santis, M., Straface, G., Carducci, B., Cavaliere, A. F., De Santis, L., Lucchese, A., et al. (2004). Risk of drug-induced congenital defects. *Eur. J. Obstet. Gynecol. Reprod. Biol.* 117, 10–19.

Gilbert-Barness, E. (2010). Teratogenic causes of malformations. *Ann. Clin. Lab. Sci.* 40, 99–114.

Gupta, K. K., Gupta, V. K., and Shirasaka, T. (2016). An update on fetal alcohol syndrome-pathogenesis, risks, and treatment. *Alcohol. Clin. Exp. Res.* 40, 1594–1602. doi: 10.1111/acer.13135

Higgins, G. A., Williams, A. M., Ade, A. S., Alam, H. B., and Athey, B. D. (2019). Druggable transcriptional networks in the human neurogenic epigenome. *Pharmacol. Rev.* 71, 520–538. doi: 10.1124/pr.119.017681

Hood, L., Rowen, L., Galas, D. J., and Aitchison, J. D. (2008). Systems biology at the Institute for systems biology. *Brief Funct. Genomic Proteomic* 7, 239–248.

Hu, T. M., Chung, H. S., Ping, L. Y., Hsu, S. H., Tsai, H. Y., Chen, S. J., et al. (2020). Differential expression of multiple disease-related protein groups induced by valproic acid in human SH-SY5Y neuroblastoma cells. *Brain Sci.* 10:545. doi: 10.3390/brainsci10080545

Jambhekar, A., Dhall, A., and Shi, Y. (2019). Roles and regulation of histone methylation in animal development. *Nat. Rev. Mol. Cell Biol.* 20, 625–641. doi: 10.1038/s41580-019-0151-1

Kowalski, T. W., Dupont, Á, Rengel, B. D., Sgarioni, E., Gomes, J. D. A., Fraga, L. R., et al. (2019). Assembling systems biology, embryo development and teratogenesis: what do we know so far and where to go next? *Reprod. Toxicol.* 88, 67–75. doi: 10.1016/j.reprotox.2019.07.015

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

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

Leigh, R. S., Ruskoaho, H. J., and Kaynak, B. L. (2020). A novel dual reporter embryonic stem cell line for toxicological assessment of teratogen-induced perturbation of anterior-posterior patterning of the heart. *Arch. Toxicol.* 94, 631–645. doi: 10.1007/s00204-019-02632-1

Martínez-Frías, M. L. (2010). Can our understanding of epigenetics assist with primary prevention of congenital defects? *J. Med. Genet.* 47, 73–80. doi: 10.1136/jmg.2009.070466

Matok, I., Gorodischer, R., Koren, G., Landau, D., Wiznitzer, A., and Levy, A. (2009). Exposure to folic acid antagonists during the first trimester of pregnancy and the risk of major malformations. *Br. J. Clin. Pharmacol.* 68, 956–962. doi: 10.1111/j.1365-2125.2009.03544.x

Mazzu-Nascimento, T., Melo, D. G., Morbioli, G. G., Carrilho, E., Vianna, F. S. L., Silva, A. A., et al. (2017). Teratogens: a public health issue - a Brazilian overview. *Genet. Mol. Biol.* 40, 387–397. doi: 10.1590/1678-4685-gmb-2016-0179

Milutinovic, S., D'Alessio, A. C., Detich, N., and Szyf, M. (2007). Valproate induces widespread epigenetic reprogramming which involves demethylation of specific genes. *Carcinogenesis* 28, 560–571. doi: 10.1093/carcin/bgl167

Mitchell, A. A., Gilboa, S. M., Werler, M. M., Kelley, K. E., Louik, C., and Hernández-Díaz, S. (2011). Medication use during pregnancy, with particular focus on prescription drugs: 1976-2008. *Am. J. Obstet. Gynecol.* 205, 51–58.

Nakaki, F., and Saitou, M. (2014). PRDM14: a unique regulator for pluripotency and epigenetic reprogramming. *Trends Biochem. Sci.* 39, 289–298. doi: 10.1016/j.tibs.2014.04.003

Nicolini, C., and Fahnestock, M. (2018). The valproic acid-induced rodent model of autism. *Exp. Neurol.* 299, 217–227. doi: 10.1016/j.expneurol.2017.04.017

Ornoy, A. (2006). Neuroteratogens in man: an overview with special emphasis on the teratogenicity of antiepileptic drugs in pregnancy. *Reprod. Toxicol.* 22, 214–226. doi: 10.1016/j.reprotox.2006.03.014

Pulice, J. L., and Kadoch, C. (2016). Composition and Function of Mammalian SWI/SNF Chromatin Remodeling Complexes in Human Disease. *Cold Spring Harb. Symp. Quant. Biol.* 81, 53–60. doi: 10.1101/sqb.2016.81.031021

Rape, M. (2017). Ubiquitylation at the crossroads of development and disease. *Nat. Rev. Mol. Cell Biol.* 19:59. doi: 10.1038/nrm.2017.83

Salilew-Wondim, D., Tesfaye, D., Hoelker, M., and Schellander, K. (2014). Embryo transcriptome response to environmental factors: implication for its survival under suboptimal conditions. *Anim. Reprod. Sci.* 149, 30–38. doi: 10.1016/j.anireprosci.2014.05.015

Schulpen, S. H., Pennings, J. L., and Piersma, A. H. (2015). Gene expression regulation and pathway analysis after valproic acid and carbamazepine exposure in a human embryonic stem cell-based neurodevelopmental toxicity assay. *Toxicol. Sci.* 146, 311–320. doi: 10.1093/toxsci/kfv094

Shahbazi, M. N., and Zernicka-Goetz, M. (2018). Deconstructing and reconstructing the mouse and human early embryo. *Nat. Cell Biol.* 20, 878–887. doi: 10.1038/s41556-018-0144-x

Theunissen, P. T., Pennings, J. L., van Dartel, D. A., Robinson, J. F., Kleinjans, J. C., and Piersma, A. H. (2013). Complementary detection of embryotoxic properties of substances in the neural and cardiac embryonic stem cell tests. *Toxicol. Sci.* 132, 118–130. doi: 10.1093/toxsci/kfs333

Theunissen, P. T., Robinson, J. F., Pennings, J. L., van Herwijnen, M. H., Kleinjans, J. C., and Piersma, A. H. (2012). Compound-specific effects of diverse neurodevelopmental toxicants on global gene expression in the neural embryonic stem cell test (ESTn). *Toxicol. Appl. Pharmacol.* 262, 330–340. doi: 10.1016/j.taap.2012.05.011

Tomson, T., Battino, D., and Perucca, E. (2019). Teratogenicity of antiepileptic drugs. *Curr. Opin. Neurol.* 32, 246–252. doi: 10.1097/wco.0000000000000659

Tung, E. W., and Winn, L. M. (2010). Epigenetic modifications in valproic acid-induced teratogenesis. *Toxicol. Appl. Pharmacol.* 248, 201–209. doi: 10.1016/j.taap.2010.08.001

van Dartel, D. A., Pennings, J. L., de la Fonteyne, L. J., Brauers, K. J., Claessen, S., van Delft, J. H., et al. (2011). Evaluation of developmental toxicant identification using gene expression profiling in embryonic stem cell differentiation cultures. *Toxicol. Sci.* 119, 126–134. doi: 10.1093/toxsci/kfq291

von Mering, C., Jensen, L. J., Snel, B., Hooper, S. D., Krupp, M., Foglierini, M., et al. (2005). STRING: known and predicted protein-protein associations, integrated and transferred across organisms. *Nucleic Acids Res.* 33, D433–D437.

Wen, H., Li, Y., Xi, Y., Jiang, S., Stratton, S., Peng, D., et al. (2014). ZMYND11 links histone H3.3K36me3 to transcription elongation and tumour suppression. *Nature* 10, 263–268. doi: 10.1038/nature13045

Worley, K. E., Rico-Varela, J., Ho, D., and Wan, L. Q. (2018). Teratogen screening with human pluripotent stem cells. *Integr. Biol.* 10, 491–501. doi: 10.1039/c8ib00082d

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

frontiers in Genetics

TECHNOLOGY AND CODE
published: 08 December 2020
doi: 10.3389/fgene.2020.589231

# Analysis of Pan-omics Data in Human Interactome Network (APODHIN)

*Nupur Biswas, Krishna Kumar, Sarpita Bose, Raisa Bera and Saikat Chakrabarti**

*Structural Biology and Bioinformatics Division, CSIR-Indian Institute of Chemical Biology, Kolkata, India*

Analysis of Pan-omics Data in Human Interactome Network (APODHIN) is a platform for integrative analysis of transcriptomics, proteomics, genomics, and metabolomics data for identification of key molecular players and their interconnections exemplified in cancer scenario. APODHIN works on a meta-interactome network consisting of human protein–protein interactions (PPIs), miRNA-target gene regulatory interactions, and transcription factor-target gene regulatory relationships. In its first module, APODHIN maps proteins/genes/miRNAs from different omics data in its meta-interactome network and extracts the network of biomolecules that are differentially altered in the given scenario. Using this context specific, filtered interaction network, APODHIN identifies topologically important nodes (TINs) implementing graph theory based network topology analysis and further justifies their role via pathway and disease marker mapping. These TINs could be used as prospective diagnostic and/or prognostic biomarkers and/or potential therapeutic targets. In its second module, APODHIN attempts to identify cross pathway regulatory and PPI links connecting signaling proteins, transcription factors (TFs), and miRNAs to metabolic enzymes via utilization of single-omics and/or pan-omics data and implementation of mathematical modeling. Interconnections between regulatory components such as signaling proteins/TFs/miRNAs and metabolic pathways need to be elucidated more elaborately in order to understand the role of oncogene and tumor suppressors in regulation of metabolic reprogramming during cancer. APODHIN platform contains a web server component where users can upload single/multi omics data to identify TINs and cross-pathway links. Tabular, graphical and 3D network representations of the identified TINs and cross-pathway links are provided for better appreciation. Additionally, this platform also provides few example data analysis of cancer specific, single and/or multi omics dataset for cervical, ovarian, and breast cancers where meta-interactome networks, TINs, and cross-pathway links are provided. APODHIN platform is freely available at http://www.hpppi.iicb.res.in/APODHIN/home.html.

**Keywords: meta-interactome, network analysis, pan-omics, multi-omics analysis, pathway cross links**

OPEN ACCESS

***Edited by:***
*Amit Kumar Yadav, Translational Health Science and Technology Institute (THSTI), India*

***Reviewed by:***
*Bhanwar Lal Puniya, University of Nebraska–Lincoln, United States*
*Marco Vanoni, University of Milano-Bicocca, Italy*

***Correspondence:**
*Saikat Chakrabarti*
*saikat@iicb.res.in*

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

***Received:*** *30 July 2020*
***Accepted:*** *11 November 2020*
***Published:*** *08 December 2020*

***Citation:***
*Biswas N, Kumar K, Bose S, Bera R and Chakrabarti S (2020) Analysis of Pan-omics Data in Human Interactome Network (APODHIN). Front. Genet. 11:589231. doi: 10.3389/fgene.2020.589231*

## INTRODUCTION

Technological advances have made different types of omics data accessible in large scale. Different types of omics data are outcomes of profiling of different bio-entities, namely RNA (RNA transcriptomics), miRNA (miRNA transcriptomics), proteins (proteomics, phosphoproteomics), genes (genomics, epigenomics), metabolites (metabolomics), lipids (lipidomics), and pharmacogenomics. These bio-entities are functionally inter-related in a complex fashion. Extrapolation from single omics data of one type of bio-entity fails to provide the true biological status of various linked bio-entities (e.g., RNA, protein, metabolites). Hence, to inquire the causative phenomena underlying the genesis and progression of systemic/genetic diseases, an integrative analysis considering the profiles of above mentioned bio-entities appears as a requisite. Moreover, because of the heterogeneous nature of the diseases, even if patients having similar pathological features are treated similarly, the disease prognosis differs a lot. It shows the inadequacy of symptom-based diagnosis and demands patient-specific analysis of omics data. Collective analysis of these multi-dimensional omics data is referred to as "pan-omics" (Sandhu et al., 2018) which are also considered as "big" data in the context of biological data analysis. Pan-omics data enable us to predict novel functional interactions between molecular mediators at multiple levels. Also, these data have the potential to uncover crucial biological observations into hallmarks and pathways that would otherwise not be obvious through single-omics studies. Patient-specific pan-omics data analysis is going to disclose the genetic, epigenetic, and other functional profiles responsible for the disease of an individual which might eventually lead to the development of individualistic "precision medicine" and will provide right treatment to right patient at right time.

Cancer is a leading cause of death worldwide, being responsible for 9.6 million deaths in 2018 (Bray et al., 2018). Cancer is a heterogeneous disease caused by aberrations of genes and proteins. "Precision oncology" promises identification of disease subtypes, specific biomarkers and subsequently prediction and translation toward the development of treatment procedures. Pan-omics or multi-omics analysis in breast cancer has revealed significant differences in molecular subtype distribution (Kan et al., 2018). Genomics and transcriptomics analysis of breast cancer data of Korean and Caucasian cohorts showed underlying molecular differences, which are responsible for the occurrence of breast cancer at the younger age in the Asian population compared to the western population (Kan et al., 2018). Multi-omics analysis extended to different types of cancers confirms the existence of broadly two types of cancers, cancers caused by recurrent mutations and cancers caused by copy-number variations (Mcgrail et al., 2018). Computational methodologies like, artificial intelligence are being used widely to extract patient-specific information from these big data, discussed in a recent review (Biswas and Chakrabarti, 2020). Machine learning based pan-omics analysis of pan-cancer data shows the existence of clusters within different types of cancers (Ramazzotti et al., 2018), identifies cell-model selective anti-cancer drug targets for breast cancer (Gautam et al., 2019).

Multiple data portals like TCGA (TCGA, 2020) and ICGA (Zhang et al., 2011) have been developed to make multi-omics data conveniently accessible. LinkedOmics contains pan-omics data of several types of cancers (Vasaikar et al., 2018). Databases like, GliomaDB (Yang et al., 2019) and MOBCdb (Xie et al., 2018) are dedicated to integrate multi-omics data for specific type of cancers. Standalone software packages and web-servers are also being developed for the analysis pan-omics data. **Table 1** compares the analytical tools which are being used by researchers. R package mixOmics (Rohart et al., 2017), based on multi-variate analysis is available for the integration of multi-omics data. It finds subsets of important features but excludes network analysis. OmicsNet provides a web-based platform to create different types of interactive molecular interaction networks for single or multiple types of omics data (Zhou and Xia, 2018). Network-based integration of multi-omics data using iOmicsPASS, allows to predict subnetworks of molecular interactions within a single type or multiple types of omics data (Koh et al., 2019). R package Miodin (Ulfenborg, 2019) provides a software infrastructure for vertical and horizontal integration of multi-omics data but lacks a comprehensive network analysis and visualization. PaintOmics allows integrated visualization of multiple types of omics data in KEGG pathway diagrams (Hern et al., 2018). Software package, Multi-Omics Factor Analysis (MOFA) (Argelaguet et al., 2018) integrates omics data in an unsupervised approach implementing generalized principal component analysis (PCA). pathfindR (Ulgen et al., 2019) finds active sub networks for genes in omics data and perform pathway enrichment analysis. R package Mergeomics (Shu et al., 2016) provides a pipeline to identify important pathways and key drivers in biological systems. However, platforms required for systematic analysis of the landscape of genetic, epigenetic, and metabolomics alterations and biological and clinical relevance of multi-layer signature in cancers are still limited.

Different types of omics data carry information on different types of bio-entities, e.g., genes, proteins, miRNAs, metabolites, etc. Hence, integrative analysis of pan-omics data needs a meta-interactome consisting of a protein–protein interaction network (PPIN) as well as different regulatory networks. The web server for the Analysis of Pan-omics Data in Human Interactome Network (APODHIN) provides a unique platform where users can analyze different types of omics data using a human cellular meta-interactome network. Graph theory based network analysis has become an essential tool for analysis of PPIN for extracting proteins important in the construction and information flow of the network (Jeong et al., 2001; Barabási and Oltvai, 2004; Mistry et al., 2017; Ashtiani et al., 2018), APODHIN provides options to identify topologically important nodes (TINs) such as hubs, bottlenecks, and central nodes (CNs) and their subsequent modules via protein–protein interaction (PPI) and regulatory relationship network analyses and pathway enrichment analysis. TINs are also correlated as prospective diagnostic and/or prognostic biomarkers. APODHIN can also analyze and compare multiple omics data set for a single omics layer, such as transcriptomics, proteomics data collected from different patient cohorts and/or different stage/grade of the same cohort.

**TABLE 1 |** Comparison of APODHIN with other existing pan-omics data analysis tools.

| Feature | APODHIN | OmicsNet [14] | mixOmics [13] | iOmicsPASS [15] | Miodin [16] |
|---|---|---|---|---|---|
| Platform | Web | Web | Standalone | Standalone | Standalone |
| Programming language | Python, R, perl | R | R | C$^{++}$ | R |
| **Types of omics data as input** | | | | | |
| mRNA transcriptomics | Yes | Yes | Yes | Yes | Yes |
| miRNA transcriptomics | Yes | Yes | Yes | No | Yes |
| Proteomics | Yes | Yes | Yes | Yes | Yes |
| Phospho-proteomics | Yes | No | Yes | No | No |
| Genomics | Yes | Yes | Yes | Yes | Yes |
| Epi-genomics | Yes | No | No | No | Yes |
| Metabolomics | Yes | Yes | Yes | No | No |
| Multiple lists of same type of omics data | Yes | No | Yes | Yes | Yes |
| Finding deregulated proteins/genes/miRNAs | Yes | No | No | No | No |
| Map in meta-interactome | Yes | No | No | No | No |
| 3D interactive network | Yes | Yes | No | No | No |
| Network topology analysis | Yes | No | No | No | No |
| Prognostic status of proteins/genes (in cancer) | Yes | No | No | No | No |
| Pathway enrichment analysis | Yes | Yes | Yes | Yes | No |
| Analysis for regulatory network protein links | Yes | No | No | No | No |

Additionally, utilizing multi-omics data APODHIN calculates cross-pathway regulatory and PPI links connecting signaling proteins or transcription factors (TFs) or miRNAs to metabolic enzymes and their metabolites using network analysis and mathematical modeling. These cross-pathway links were shown to play important roles in metabolic reprogramming in cancer scenarios such as glioblastoma multiforme in a previous work (Bag et al., 2019).

In addition to the server part, APODHIN shares analysis of multi-omics data from various cancer cell lines where TINs and cross-pathway links were identified using publicly available omics datasets collected for various gynecological cancers. APODHIN platform is freely available at http://www.hpppi.iicb.res.in/APODHIN/home.html.

# MATERIALS AND METHODS

## Server Description

Analysis of Pan-omics Data in Human Interactome Network web server is dedicated for the integration and subsequent analysis using single or multiple types of omics data. For single type of omics data, APODHIN can analyze multiple datasets (up to 3) which may correspond to either different stages of a disease from a single cohort or from dataset collected from multiple patient cohorts and/or cell lines.

For multiple types of omics data, APODHIN allows single input data file for each type of omics data. Following sections briefly describe the various analytical part of the APODHIN server.

## Data Collection

Analysis of Pan-omics Data in Human Interactome Network web server is preloaded with a human cellular meta-interactome network. This meta-interactome consists of human protein–protein interaction network (HPPIN), network of human miRNAs and their target genes and network of human TFs and their target genes. The PPI data was collected from STRING (Szklarczyk et al., 2019) database (version 11). Interactions having a medium threshold of experimental score $\geq$700 were considered (Ferretti and Cortelezzi, 2011) for construction of the PPIN. Target gene information of miRNAs was collected from the TarBase (Vergoulis et al., 2012) and miRTarBase (Chou et al., 2016) databases. From the TarBase database (version 6) we have taken reliable interactions supported only by low-throughput experiments (e.g., reporter gene assay, western blot, qPCR, etc.) whereas miRNA target interactions with strong confidence (i.e., validated by either of report assay, western blot, qPCR experiments) from miRTarBase (version 6) were considered for APODHIN meta-interactome network. We trusted on the more reliable low-throughput experimental data to build the parent miRNA-target mRNA interactome network. We found 2492 target genes for 544 miRNAs creating 6917 interactions. TFs and their target genes were downloaded from Human Transcriptional Regulation Interactions database (HTRIdb) (Bovolenta et al., 2012). We found 11887 target genes for 284 TFs creating 18153 interactions. These three networks were merged together to form the APODHIN meta-interactome consisting of two types of biomolecular nodes i.e., proteins/genes and miRNAs along with three types of interactions, i.e., protein–protein, miRNA-target gene, and TF-target gene, respectively.

Additionally, we have also included a network of metabolites as substrate and product with their corresponding metabolic enzymes in the APODHIN server. For constructing this network, we downloaded metabolic reactions from MetaNetX database (Moretti et al., 2016) and extracted the metabolites along with the corresponding metabolic enzymes and further filtered those

enzymes and metabolites which have been listed in the Human Metabolome Database (HMDB) database (Wishart et al., 2018).

## Pan-omics Data Integration and Meta-Interaction Network Extraction

In APODHIN web server, user can upload single or multiple types of omics data. The server accepts RNA transcriptomics, miRNA transcriptomics, proteomics, phosphoproteomics, genomics, epigenomics, and metabolomics data. The current version of the server accept only processed format of the omics data where differential expression/abundance of corresponding biomolecules are provided with *log*FC for defining up and down regulation of genes/miRNAs/proteins and threshold probability or *p*-value. For RNA transcriptomics, miRNA transcriptomics and proteomics data user should select threshold values of *log*FC for defining up and down regulation of genes/miRNAs/proteins and corresponding adjusted *p*-value. Uploaded files should contain list of genes/miRNAs/proteins along with *log*FC and *p*-values. Sample file formats for different omics data are provided in the APODHIN help page. For genomics, epigenomics, and phosphoproteomics data, genes that are mutated and/or methylated and proteins, which are phosphorylated are considered, respectively. APODHIN help page also provides guidelines to process GEO (Barrett et al., 2013) transcriptomics data for using in APODHIN. Packages and tools for GEO series data are also enlisted in the APODHIN "Help" page. For other types, of omics data like, proteomics, genomics, metabolomics, useful links for data processing is provided in the APODHIN help page and it will be made more enriched gradually depending on the requirements from users.

Analysis of Pan-omics Data in Human Interactome Network web server extracts the interactome networks from the parent meta-interactome for the genes, mRNAs, miRNAs, proteins, and metabolites that are either deregulated or altered according to the user supplied single or multiple omics data. It creates a filtered meta-interactome network comprising of deregulated or altered nodes and their 1st or 2nd level (as chosen by user) interactors and/or regulators. For metabolomics data, the web server finds out the proteins linked with metabolites and constructs network. These single or multi omics data specific meta-interactome networks are subsequently displayed in an interactive three-dimensional (3D) network viewer within the APODHIN server. For creating omics data mapped network, and subsequently network analysis, APODHIN does not provide any special weight or scores to any type of omics data.

For the module "pathway connectivity analysis," RNA transcriptomics, miRNA transcriptomics, and proteomics data were considered as primary data and submission of at least one of them is mandatory to define deregulated miRNAs and/or genes/proteins. In case of "pathway connectivity analysis," the *log*FC values for each of the uploaded omics data is normalized in the scale of −1 to +1 following Eq. 1,

$$\log FC_{\text{normalized}} = \frac{\log FC}{\left|\log FC\right|_{\max}} \tag{1}$$

where positive and negative values indicate up and down regulated entities, respectively. If more than one primary omics data, for example, transcriptomics and proteomics are provided, APODHIN web server sums up the normalized *log*FC values from the different omics data for the same node (RNA/protein) and if the sum is non-zero, gene/protein is considered deregulated. Primary omics data determines whether the gene is deregulated or not. Also, if a gene is found not altered in supplied primary omics data, APODHIN does not consider this gene for further analysis, irrespective of its status in the supplied secondary omics data. Details of the utilization of the normalized omics values in mathematical modeling based pathway connectivity link identification are provided later. In this module, the information on metabolites for any enzyme can be obtained in the associated table on selection of enzyme.

## Network Analysis and Identification of TINs

Once the context specific meta-interactome network is formed via utilization of user supplied single or multiple omics data, APODHIN web server primarily finds three types of TINs, namely, hubs, CNs (Bhattacharyya and Chakrabarti, 2015) and bottlenecks (BNs) (Yu et al., 2007). To find the important nodes, network and node indices like degree, betweenness, closeness and clustering coefficients are calculated from the extracted meta-interactome network. These node parameters were calculated using previously reported methods and protocols (Bhattacharyya and Chakrabarti, 2015). For transcriptomics and proteomics data, TINs are identified from the expressed nodes only. For phosphoproteomics, genomics, epigenomics and metabolomics data, TINs are identified from phosphorylated, mutated, methylated proteins/genes and metabolic enzymes, respectively.

Hubs are nodes that have high degrees. Degree distribution is normalized following Eq. 2,

$$x_{i,normalized} = \frac{x_i}{x_{maximumn}} \tag{2}$$

where $x_i$ is degree value of a node $i$ and $x_{maximum}$ is the maximum degree of the network. APODHIN web server converts normalized degree distribution to corresponding z-score distribution. The plot of probability distribution function (PDF) of z-scores for all nodes in network is sent to the user by email. This email shares intermediate results only. From the plot of PDF, users are asked to provide the threshold value for hub identification. After receiving the threshold value, APODHIN initiates hub identification program. Nodes having degree greater than the threshold value are considered as hub. It is also mentioned in the help page. Scores concerning individual centrality parameters like, betweenness, closeness and clustering coefficients are calculated and the cumulative centrality scores (CCS) are estimated by summing over the combined scores for first layer interactors (Bhattacharyya and Chakrabarti, 2015). CCSs are normalized following Eq. 2 where $x$ is equal to CCS. Normalized CCS are converted into z-scores. The PDF of z-scores for all nodes of network are sent to the user by email and CNs are chosen based on the user provided threshold value of z-score following similar procedure as mentioned while identifying hubs.

Bottleneck nodes are characterized based on their betweenness values. Normalized betweenness values were obtained from Eq. 2 where $x$ is betweenness and subsequently, converted into z-scores. Similar to hubs, bottleneck nodes are also chosen based on the user provided threshold z-score, chosen from the PDF plot of z-score for all nodes.

Further, sub-network consisting of TINs and their first or second layer interactors are constructed and displayed in an interactive three-dimensional (3D) network viewer.

The overlap of TINs, as well as all nodes of the network, as prognostic cancer marker is checked after extraction of prognostic marker information from the Human Protein Atlas database (version 19) (Uhlen et al., 2017). The prognostic data was obtained from Kaplan-Meier survival analysis. The cancer type, for which prognostic status have minimum $p$-value, is shown in the "Node information" table in the page of "network view of identified important nodes." On mouse hover on the cancer type, more detail information for other cancer types, is available.

## Pathway Mapping and Network of Mapped Pathways

For each identified TIN, particularly for genes and proteins, APODHIN maps the corresponding pathways listed in the KEGG database (Kanehisa et al., 2017). APODHIN performs a hypergeometric Fishers Exact test and selects enriched pathways satisfying $p$-value ($p_{HGD}$) $\leq$0.05 using the following contingency table and formula.

$$\begin{bmatrix} a & b-a \\ c & d-c \end{bmatrix}$$

Where,

a = Number of genes in the pathway.
b = Number of genes in the gene list.
c = Total number of genes in the pathway.
d = Total number of genes in all pathways in KEGG.

$$p_{HGD} = \frac{\binom{b}{a}\binom{d}{c}}{\binom{b+d}{a+c}} \quad (3)$$

Further, a network representation of important nodes along with their enriched mapped pathways is displayed in an interactive three-dimensional (3D) network viewer. **Figure 1A** shows the flow chart of "pan-omics data mapping and network analysis" module of APODHIN.

## Pathway Connectivity Analysis and Cross-Pathway Links

This module of the APODHIN web server aims to construct regulatory interaction networks and subsequently identifies cross-pathway interaction links connecting different cellular pathway proteins [e.g., signaling proteins (S)], regulatory proteins [e.g., transcription factor (TF)] or miRNAs with metabolic pathway proteins (M).

For this purpose, APODHIN web server was preloaded with cross-pathway links or paths where protein–protein interactors (P) connect X nodes (X can be S or target gene of TF or target genes of miRNAs) with M (metabolic) proteins. We have limited the number ($n$) of protein–protein interactors (P) to a maximum value of three between X and M proteins. This limit provides four types of paths, XM ($n$ = 0), XPM ($n$ = 1), XPPM ($n$ = 2), XPPPM ($n$ = 3). These cross-pathway linking paths are filtered and selected based on expression and/or abundance status of the biomolecules supplied by user uploaded pan-omics data for a given disease or context. The filtering criteria for any given path is set when the terminal nodes are found to be deregulated and the remaining nodes are at least expressed within the user provided single or multi-omics datasets.

We implemented an established probabilistic approach based on the Hidden Markov Model (HMM) (Tuncbag et al., 2013; Vinayagam et al., 2014; Bag et al., 2019) utilizing the information of experimentally established PPIs and gene regulatory information to extract novel paths and interconnections between regulatory nodes such as signaling proteins, TFs and miRNAs and metabolic pathway proteins (M). Within these important X-M pairs, important cross-pathway connecting paths are again scored by considering all filtered paths between X-M pairs. To find important X-M pairs, weights are assigned on nodes and edges depending on network and biological properties. Edge weight is assigned in terms of normalized interaction probability which is proportional to the product of their expression scores.

Two types of node weights, network entropy, and effect-on-nodes are considered. Network entropy includes local entropy of the node. Another node weight parameter, "effect-on-node" considers the impact of interactors of a particular gene in the cross-connected network. The "effect-on-node" considers both biological and network properties of the node. Biological properties include deregulated gene, signaling crosstalk gene and rate limiting enzyme. Network properties include hubs, CNs, and bottlenecks.

Analysis of Pan-omics Data in Human Interactome Network web server allows the user to choose maximum four weight options out of the six weights. If a node satisfies any of the selected weight options, weight value 1 is assigned for each satisfied option. To identify important cross-connecting X-M pairs we have evaluated "path score" (PS) based on a HMM implemented within the core mathematical model that calculated the significant cross-pathway linking paths. "Path scores" are converted to z-scores and paths having z-score $\geq$1 are considered as important cross-connecting paths. A detailed description of the mathematical models and path calculation is available in our previous publication (Bag et al., 2019). **Figure 1B** shows the flow chart of "pathway connectivity analysis" module of APODHIN.

## APODHIN Architecture

Analysis of Pan-omics Data in Human Interactome Network web server is created using HTML, PHP, PYTHON, and JAVA scripts. Client/user side scripts are written in HTML, PHP and JAVA scripts. User uploaded data is analyzed using PYTHON scripts. For network analysis, PYTHON package networkX (version

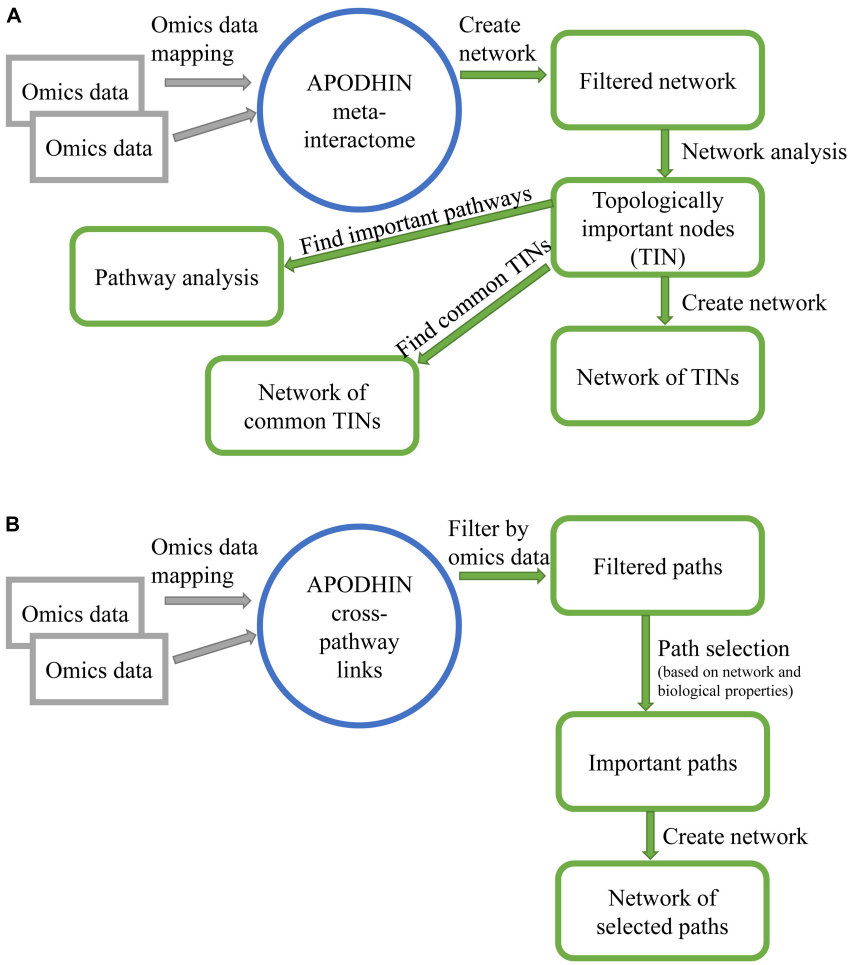



**FIGURE 1 |** Flow charts showing work flow in APODHIN web-server for module **(A)** data mapping and network analysis and **(B)** pathway connectivity analysis.

1.8.1) is used. For visualization of 3D presentation of networks JAVA scripts based open source technologies (*three.js* and *3d-force-graph.js*) were utilized.

Analysis of Pan-omics Data in Human Interactome Network has two separate parts A. APODHIN server and B. APODHIN example data analysis.

## APODHIN Server

Analysis of Pan-omics Data in Human Interactome Network web server is preloaded with human interactome network containing PPIN, target gene network of miRNAs and target gene network of TFs. Proteins participating in signaling and metabolic pathways are also marked separately. Metabolites along with their target enzymes are also included within APODHIN. This meta-interactome network is used as framework of cellular interactions and is further used to map user supplied single or multiple types of "omics" data to perform the following analyses.

- Omics data mapping and network analysis: This module has two sub-modules. On clicking first submit button, this web server provides meta-interactome network filtered by uploaded omics data where deregulated and/or altered nodes along with their interactors are included. Users can further proceed for finding important interacting nodes from the "pan-omics" data mapped interaction network by clicking second submit button. Tabular, graphical and 3D network representations of the identified TINs are provided for better appreciation. Overlap of the TINs is shown both in tabular and interactive 3D network visualization. Additionally, TINs and their enriched pathways are also shown in tabular and interactive 3D network visualization manner.

  Sample input files for each omics data type and example analysis output are provided for the ease of use and apprehension.
- Pathway connectivity analysis: As mentioned before, this sub-module highlights significant PPI and regulatory paths connecting signaling proteins/TF/miRNAs to metabolic proteins. These cross-pathway links are thought to be supra-molecular regulatory links/signatures connected with metabolic rearrangement or reprogramming events that are observed during cancer. In APODHIN, these

cross-pathway regulatory links can be constructed from three types of interaction networks.

1. Integrated network where signaling (S) and metabolic (M) pathway proteins are connected through protein–protein interactors (P).
2. Integrated network where target genes of TFs and metabolic (M) pathway proteins are connected through protein–protein interactors (P).
3. Integrated network where miRNA target genes and metabolic (M) pathway proteins are connected through protein–protein interactors (P).

Cross-pathway linking paths are filtered and selected based on expression and/or abundance status of the biomolecules supplied by user uploaded single or pan-omics data for a given disease or context. These paths are shown both in tabular and interactive 3D network visualization.

## APODHIN Example Data Analysis

Analysis of Pan-omics Data in Human Interactome Network example data analysis page showcase few example analysis of multi-omics data for different cancer cell lines. We have used the APODHIN web server to construct individual cancer and dataset centric meta-interactome network using cell line specific single and/or multi-omics data collected from various resources such as GEO (Barrett et al., 2013), PRIDE (Perez-Riverol et al., 2019), publication reports and data sources for cervical, ovarian, and breast cancers, respectively. Further, these cancer and dataset specific meta-interactome networks were analyzed and important interacting nodes and cross-pathway links were identified and provided within the APODHIN example data analysis module. We have used cancer cell line derived omics data freely available from different public resources. Options are provided for the users to select single and/or multi-omics data to construct the meta-interactome networks and further analyze them to identify and important interacting nodes and cross-pathway links specific for the selected dataset.

# RESULTS

## Input Options

Analysis of Pan-omics Data in Human Interactome Network server provides two different but linked analysis options for the users who would like to utilize single or multiple types of omics data for a given context. APODHIN web server provides options to upload seven types of "omics" data comprising of mRNA transcriptomics, miRNA transcriptomics, proteomics, phosphoproteomics, genomics, epigenomics, and metabolomics. The file formats for each data type is specified in the "Help" page and sample input files are also available in the server input page. Information on preparing input files for using in APODHIN is also shared in the "Help" page. For transcriptomics and proteomics data, maximum and minimum threshold values for the differential expression/abundance (*log*FC) and statistical significance of that (*p*-values) need to be provided. As the calculations are computation intensive, results are sent via email.

Similarly, for cross-pathway connectivity analysis users need to upload single or multiple types of "omics" data for a given context. At least one "primary" type (see Methods) of omics data need to be uploaded. Now, in this case, users also need to specify the type of connectivity they would like to explore, for example, signaling to metabolic proteins, TFs to metabolic proteins, or miRNAs to metabolic proteins. Only one type of pathway connectivity can be explored at a time for a given set of "omics" data. Additionally, users also need to select the kind of weights (see section "Materials and Methods") that would be applied while calculating the scores of the selected cross-pathway regulatory and PPI paths. E-mail address needs to be supplied for APODHIN server to send the result link of the identified cross-pathway connections.

## Output Options

Output option for the "Data mapping and network analysis" module has two stages. At first stage (**Figure 2A**), the context specific meta-interactome network ("filtered network") can be visualized via a user interactive 3D network viewer where information regarding each node and edge are provided in graphical as well as tabular view (**Figure 2B**). Status of the "omics" data mapping is shown in various color codes for the nodes whereas different relationship like PPI, miRNA-target gene interaction, and TF and target in connections are shown varied color codes. Additional details about the protein nodes can be obtained via GeneCards (Stelzer et al., 2016) link while miRNA details can be found via miRTarBase (Chou et al., 2016) link. List of metabolites mapped onto the protein nodes are also provided both in the network viewer as well as in the adjacent tabular format. If network analysis is opted, along with filtered network, APODHIN provides the PDFs for the opted TINs (**Figure 2C**). Filtered nodes (genes/proteins/miRNAs) that satisfied the selected threshold criteria are characterized as TINs and further utilized for meta-interactome network construction. If multiple files of single type of omics data is uploaded, users can see the number of TINs (as hub, bottlenecks, and CNs) and their mutual overlap using interactive Venn diagram by clicking the "link for analysis" option for single or combination of "omics" data (**Figure 2D**). Combined analysis of multiple types of omics data files is shown if multiple types of omics data files are provided. Here also, the resultant page (**Figure 2E**) provides three output options. First, the regulatory and PPI connectivity specific to the hubs, bottleneck and CNs can be seen via corresponding link where networks of deregulated hubs, bottleneck, and CNs can be seen separately and saved accordingly (**Figure 2F**). Association to various kinds of cancers for the identified TINs as favorable/unfavorable prognostic markers are also provided here after mapping the TINs (see Methods) to the data provided in Human Protein Atlas (Uhlen et al., 2017). Another option provides the network of common TINs (**Figure 2G**) whereas a separate link provides network of enriched pathways with the identified TINs (**Figure 2H**). Enriched pathway networks of deregulated hubs, bottleneck, and CNs can be seen separately and saved accordingly. In all these three network output options, data can be downloaded in text format for further analysis.

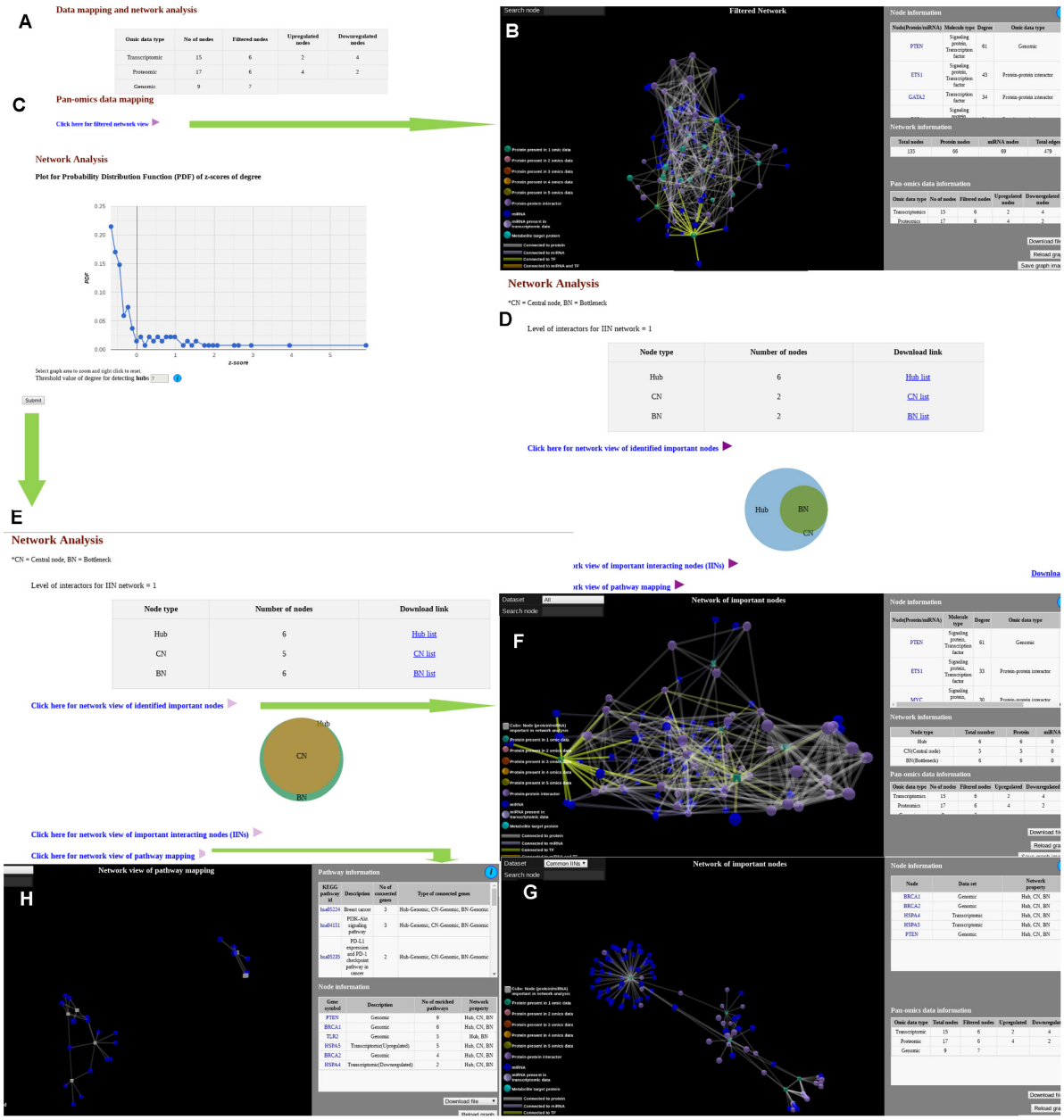



**FIGURE 2 |** Snapshots of outputs of module "data mapping and network analysis." **(A)** Page showing link for filtered network and probability distribution function. **(B)** Filtered network. **(C)** Probability distribution function for network analysis. **(D)** Output page of a single omics data. **(E)** Network analysis page for multi-omics data. **(F)** Network of important interacting nodes. **(G)** Network of important nodes. **(H)** Network of pathway mapping.

Similar to "Data mapping and network analysis," "Pathway connectivity analysis" module also provides a tabular result with a summary of the user uploaded data (**Figure 3A**). Users can see the cross-pathway links for multiple types of omics data (**Figure 3B**). For multiple types of files with single type of omics data (**Figure 3C**), the comparison (**Figure 3D**) is shown in Venn diagram as well as in network visualization. In the 3D network visualization window, significant PPI and regulatory paths connecting signaling proteins/TFs/miRNAs to metabolic proteins are shown in color coded fashion. As described before, these cross-pathway links or paths connect X nodes (X can be S or target gene of TF or target genes of miRNAs) with metabolic (M) proteins. These linking paths are filtered and selected based on expression and/or abundance status of the biomolecules supplied by the users where for any given path the terminal nodes are found to be deregulated and the remaining nodes are at least expressed. The corresponding pathways and biological functions of the proteins are also provided in tabular format adjacent to the network viewer. Additionally, the metabolites connected to the metabolic proteins that are

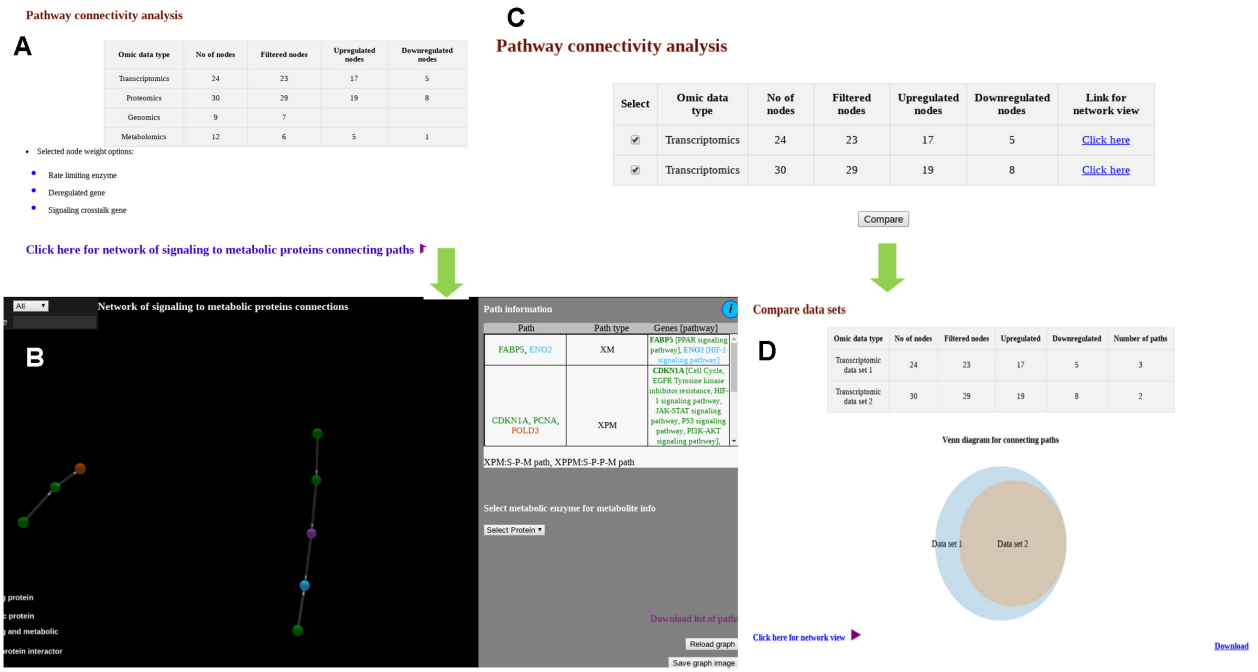



**FIGURE 3 |** Snapshots of outputs of module "pathway connectivity analysis." **(A)** Output page shows user provided data in tabular form along with link for network view. **(B)** Output page showing network of signaling to metabolic proteins connecting paths for multiple types of omics data. **(C)** Output page when multiple files for single type of omics data is provided. **(D)** Venn diagram shows overlap of signaling to metabolic proteins connecting paths for different omics data set.

part of the selected cross-pathway links are also provided in the same page.

## Example Data Analysis Option

Analysis of Pan-omics Data in Human Interactome Network example data analysis page contains important nodes (genes/proteins/miRNAs), pathways, and their networks with interacting partners specific for cancers affecting women such as cervical, ovarian, and breast cancer. This section also contains important paths linking signaling proteins/TFs/miRNAs to metabolic enzymes, which could perhaps be responsible for metabolic reprogramming in cancer. The example content is produced by APODHIN web server using publicly available cervical, ovarian, and breast cancer specific cell line based omics data. **Figure 4** briefs the statistics derived from APODHIN example analyses for mRNA transcriptomics data of different cell lines of cervical, ovarian, and breast cancer. **Figure 4A** shows the overlap of deregulated genes. It reveals lesser overlap among deregulated genes across cell lines for all cancers. However, there is almost complete overlap of pathways mapped by deregulated genes (**Figure 4B**). Nodes satisfying any two types of TINs are considered as important interacting nodes (IINs). **Figure 4C** shows overlap for common IINs between cell lines across cancer types are observed. Similarly, **Figure 4D** shows much higher overlap of common pathways mapped by IINs. This demonstrates that IINs and their pathways represent the common core genes and processes related to a cancer type in a better way than that achieved by the initial deregulated genes obtained from the omics data. We also checked whether the mapped pathways are related to cancer pathways enlisted in KEGG database (Kanehisa et al., 2017). **Figure 5A** shows that pathways mapped by IINs are more cancer specific compared to the pathways mapped by deregulated genes for all cell lines. **Figures 4E,F** show the number and overlap of deregulated genes and IINs as prognostic markers of respective cancer type. **Figure 5B** shows that compared to the deregulated genes, IINs possess higher fractions of prognostic markers for all cancer cell lines, except MDAMB231. This advocates the usefulness of the IINs over deregulated genes. Moreover, as the number of IINs is much smaller than that of deregulated genes the false discovery rate is also expected to be lower.

**Figure 6** shows overlap of cross-pathway links or paths connecting signaling (S) proteins, TFs, and miRNAs to metabolic (M) proteins identified using omics data derived from the cell lines of three types of cancers. For signaling to metabolic connection, four common paths for three cervical cell lines were observed. However, no such overlap was found for breast and ovarian cancer cell lines.

Analysis of pan-omics data considering transcriptomics, genomics, epigenomics, metabolomics data in different combinations are available for different cell lines in the example data analysis section of APODHIN.

## DISCUSSION

Large-scale genomics, transcriptomics and proteomics approaches have made it possible to characterize different

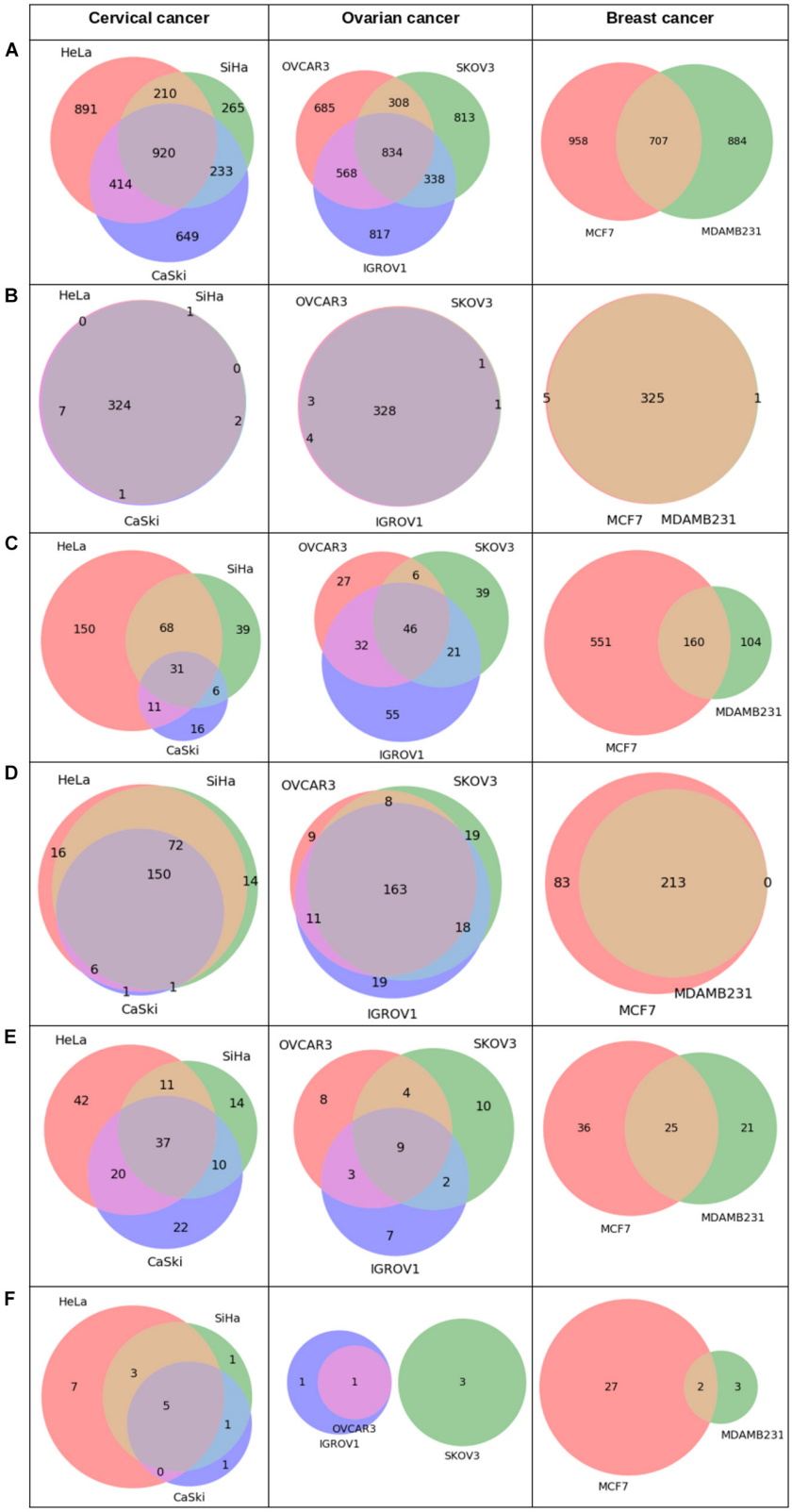



**FIGURE 4 |** Statistics derived from APODHIN database for mRNA transcriptomics data derived from different cell lines of cervical (HeLa, SiHa, and CaSki), ovarian (IGROV1, SKOV3, OVCAR3), and breast cancer (MCF7 and MDAMB231). Transcriptomics data was derived from the GEO datasets GSE9750, GSE19352, and GSE71363, respectively. **(A)** Deregulated genes, **(B)** Overlap of pathways mapped by deregulated genes, **(C)** Overlap of IINs, **(D)** Overlap of pathways mapped by IINs, **(E)** Deregulated genes as prognostic marker, and **(F)** IINs as prognostic marker.

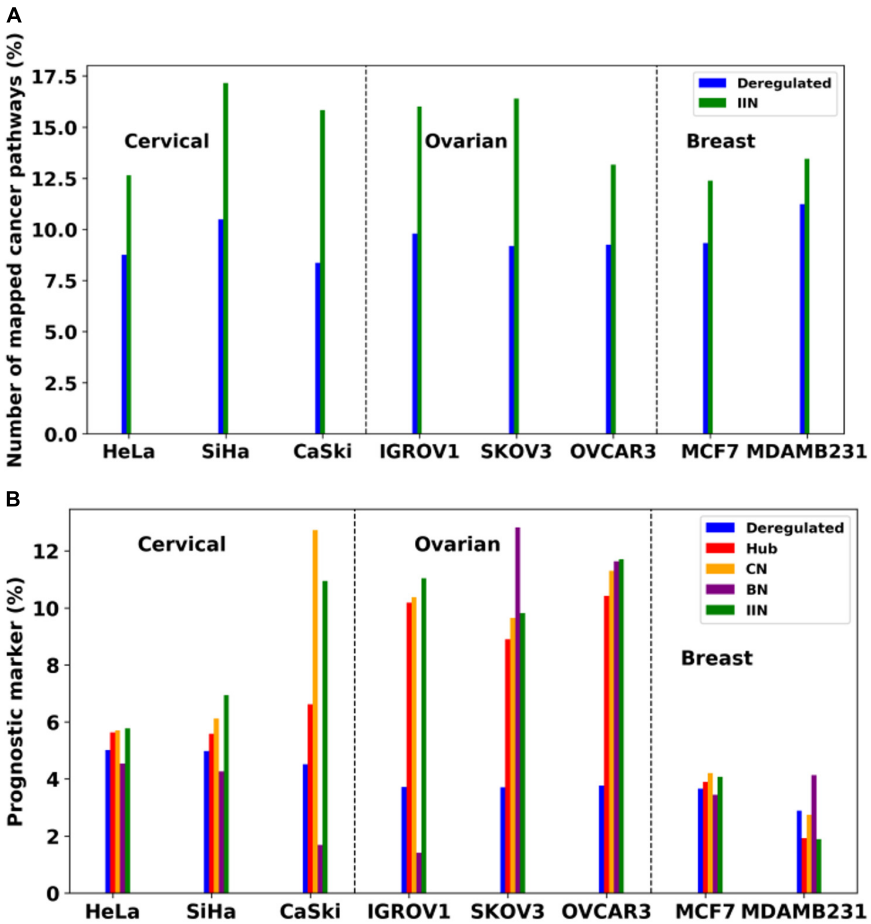



**FIGURE 5 | (A)** Comparison of number of cancer specific pathways mapped by deregulated genes and IINs. **(B)** Comparison of fraction of prognostic markers within the deregulated genes and network analysis derived important nodes, such as IIN and various TIN (e.g., Hubs, CN, and BN, respectively). Dashed lines are drawn to separate cell lines of different cancer types.

clinical spectra associated with cancers. Use of pan-omics platforms and approaches in the analysis of systemic disease like cancer will not only help to identify numerous useful biomarkers but also will expose areas for further improvement in therapeutic intervention. Here, we present APODHIN web server, which extracts cellular interactome networks from the parent meta-interactome for the genes, mRNAs, miRNA, proteins, and metabolites that are either deregulated or altered according to the user supplied single or multiple omics data. These single or multi-omics data specific meta-interactome networks are utilized to identify TINs and their sub-modules enriched with PPI and regulatory relationship via utilization of graph theory based network analyses and biological pathway enrichment analysis. Important interacting nodes (proteins and miRNAs), IINs are identified based on the overlap of key nodes such as hubs and bottlenecks. Using data from The Human Protein Atlas database, APODHIN provides the probable prognostic status of the IINs. Also, as observed in our earlier works (Bhattacharyya and Chakrabarti, 2015), IINs extracted from network topology, could correlate to be prospective diagnostic and/or prognostic biomarkers or even turn out to be potential therapeutic targets.

Molecular mechanisms for cancer progression and development of potential therapeutics to inhibit these complex diseases are difficult from the independent knowledge of signaling, TFs, miRNAs, and metabolic pathways. Metabolic reprogramming is an essential hallmark of cancer (Hanahan and Weinberg, 2011). Understanding the coordination among various cellular pathways, such as gene-regulatory, signaling and metabolic pathways is crucial and may provide clues into the molecular mechanism of metabolic adaptation in cancer and associated cells. Therefore, there is an urgent need for systems biology model, which can coordinate among signaling-induced proliferation of tumor cells/growth, transcription factor/miRNA based gene regulation and metabolic processes. Hence, we emphasized to design a mathematical approach to identify significant proteins forming interconnections between signaling, regulatory and metabolic pathways. We have constructed an integrated network where signaling (S), regulatory (TFs and miRNAs), and metabolic (M) pathway entities are connected

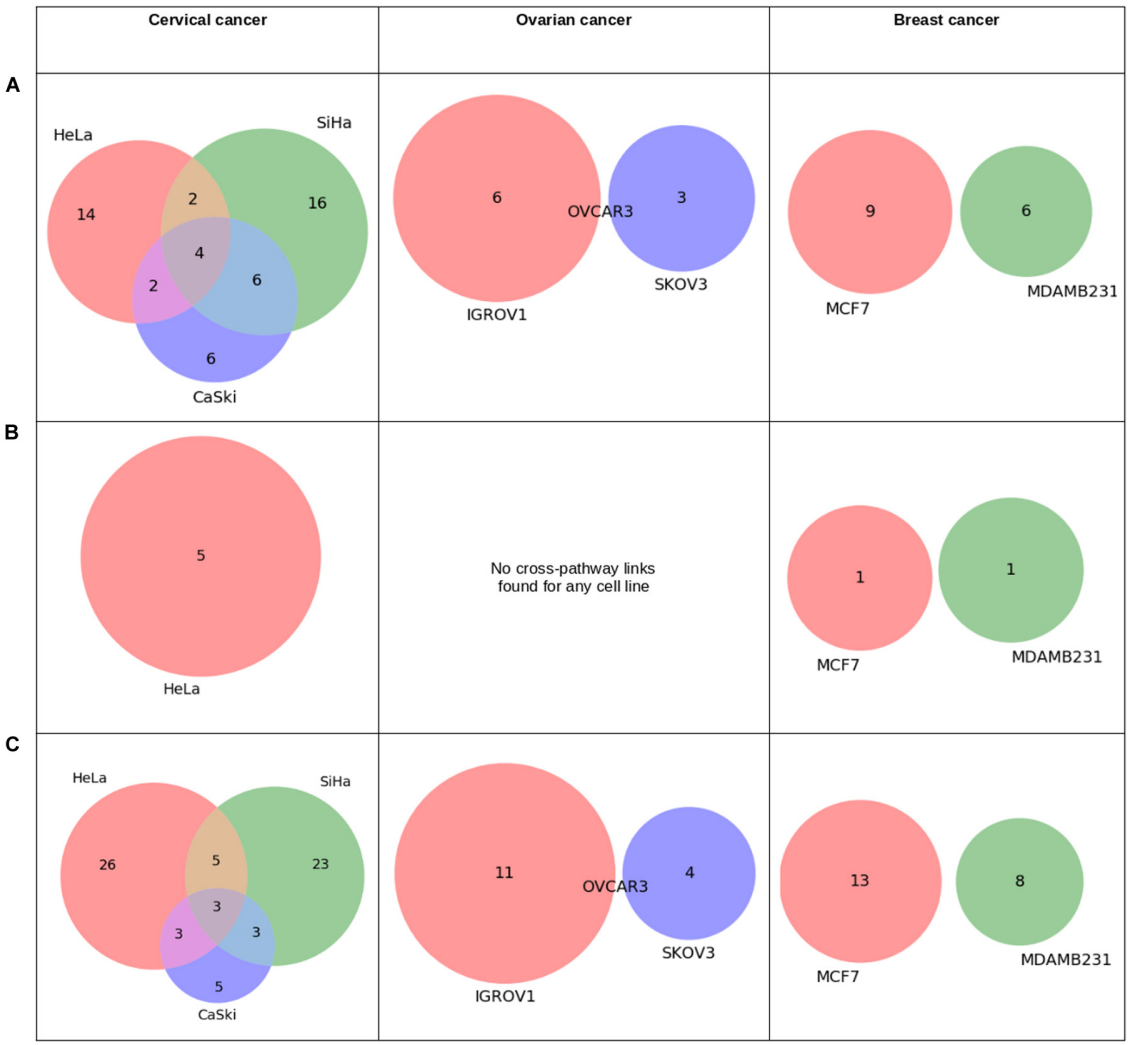



**FIGURE 6 |** Statistics derived from "pathway connectivity analysis" module of APODHIN database for mRNA transcriptomics data derived from different cell lines of cervical (HeLa, SiHa, and CaSki), ovarian (IGROV1, SKOV3, OVCAR3), and breast cancer (MCF7 and MDAMB231). Transcriptomics data was derived from the GEO datasets GSE9750, GSE19352, and GSE71363, respectively. **(A)** Overlap of cross-pathway links connecting signalling (S) to metabolic (M) proteins, **(B)** Overlap of pathway links connecting target genes of TFs to metabolic (M) proteins, and **(C)** Overlap of cross-pathway links connecting target genes of miRNAs to metabolic (M) proteins.

through protein–protein and gene regulatory interactions. Interconnections between regulatory components such as signaling proteins/TFs/miRNAs and metabolic pathways need to be elucidated rigorously to understand the role of oncogene and tumor suppressors in regulation of metabolism alongside their normal functions. Analyses of such cross-connected network and linking paths will facilitate probable way(s) to inhibit cancer progression in a more specific manner.

Considering the growing demand of multi-omics data integration followed by systems biology based analytical interpretation of the large-scale "omics" data, implementation of a robust and user-friendly web-based platform is very much due. In order to make better sense out of the various "omics" data, it is imperative to utilize them in a way so that the global scenario of the complex and multi-layer cellular interactome can be recapitulated. Several data portals have been coming up to make multi-omics data accessible, visible and more importantly, interpretable. Various programs and web portals are being made to interpret omics data in different perspectives. Each of these tools has its own merits and limitations also. **Table 1** provides a qualitative comparison of features and functionalities of APODHIN with respect to existing omics data analysis tools. Web servers like OmicsNet (Zhou and Xia, 2018) is a technically powerful web based platform specifically meant for better visualization of molecular networks. It mainly provides varied and efficient ways of network visualization including different components. However, it provides minimal emphasis on networks analysis and identification and interpretation of

important interacting nodes and cross-pathway links. Similarly, this server only accept differential omics data for genes/proteins and metabolites, it does not have the option to include the epigenetic modification, miRNA expression data, and phosphoproteomics data. mixOmics (Rohart et al., 2017) is a software package which is based on multi-variate analysis. It performs data reduction, and then identifies combination of biomarkers. It offers a network visualization but does not consider network topology. It does not consider any meta-interactome. Software package iOmicsPASS (Koh et al., 2019) considers a meta-interactome by including PPIN and TF regulatory network within omics data. But it excludes miRNA-mRNA regulatory network. It considers only three types of omics data, transcriptomics, proteomics, DNA copy number data, thus limiting its applicability. Another R package Miodin (Ulfenborg, 2019) provides opportunity of creating a workflow of data analysis. It considers different omics data, but not metabolomics data. It requires pre-installation of several R packages. Miodin provides Venn diagram of differentially expressed genes, overlapped within different datasets. However, Miodin does not consider any meta-interactome and does not construct any network. None of these tools perform network topology analysis and provide cross-pathway connectivity information of proteins. APODHIN is perhaps the only available web based platform that offers to (a) integrate multi-omics data onto an exhaustive multi layered cellular meta-interactome network, (b) extract and analyze the context specific networks and sub-networks to identify TINs that could serve as potential biomarkers and/or therapeutic targets (c) rationalize the role of the identified TINs to the given context via pathway enrichment and prognostic marker correlation, and (d) identify cross-pathway interconnections between regulatory components such as signaling proteins/TFs/miRNAs and metabolic pathways for better understanding the role of oncogenes and tumor suppressors in regulation of metabolic reprogramming during cancer. Additionally, being a web based tool, APODHIN requires no installation of software, good computing systems, and technical expertise. We believe these features make APODHIN useful as well as a user-friendly application.

However, there is still scope for improvement for the APODHIN server. The example data analysis part can be enriched to upgrade as a database. For example, in future we would like to equip the server to accept and process raw "omics" data directly and further create the processed data for genetic or epigenetic alterations, differential expression and abundance, respectively. We would also like to add components for handling large number of datasets which will be able to analyze cohort data. Current version is mostly aimed to patient-specific personalized data. Similarly, the server and along with a database should be enriched in such a way that it could be utilized for deep learning and artificial intelligence based tools to predict the disease outcome, recurrence and drug resistance, respectively.

## 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.

## AUTHOR CONTRIBUTIONS

NB and SC designed the web server. NB created the web server. KK, SB, and RB provided the data for meta-interactome network. NB and SC analyzed the data and drafted the manuscript. All authors contributed to the article and approved the submitted version.

## FUNDING

The authors acknowledge CSIR-Indian Institute of Chemical Biology for infrastructural support. SC acknowledges the Systems Medicine Cluster (SyMeC) grant (GAP357), Department of Biotechnology (DBT) for funding. NB acknowledges the Systems Medicine Cluster (SyMeC) grant (GAP357), Department of Biotechnology (DBT) for fellowship. KK, SB, and RB acknowledge Department of Biotechnology (DBT), Council of Scientific and Industrial Research (CSIR), respectively for their fellowships. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. This manuscript has been released as a pre-print at bioRxiv (Biswas et al., 2020).

## REFERENCES

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.

Ashtiani, M., Salehzadeh-yazdi, A., Razaghi-moghadam, Z., Hennig, H., and Wolkenhauer, O. (2018). A systematic survey of centrality measures for protein-protein interaction networks. *BMC Syst. Biol.* 12:80. doi: 10.1186/s12918-018-0598-2

Bag, A. K., Mandloi, S., Jarmalavicius, S., and Mondal, S. (2019). Connecting signaling and metabolic pathways in EGF receptor-mediated oncogenesis of glioblastoma. *PLoS Comput. Biol.* 15:e1007090. doi: 10.1371/journal.pcbi.1007090

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

Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., et al. (2013). NCBI GEO: Archive for functional genomics data sets - Update. *Nucleic Acids Res.* 41, 991–995.

Bhattacharyya, M., and Chakrabarti, S. (2015). Identification of important interacting proteins (IIPs) in *Plasmodium falciparum* using large-scale interaction network analysis and in-silico knock-out studies. *Malar J.* 14, 1–17.

Biswas, N., and Chakrabarti, S. (2020). Artificial intelligence (AI) based systems biology approaches in multi-omics data analysis of cancer. *Front. Oncol.* 10:588221. doi: 10.3389/fonc.2020.588221

Biswas, N., Kumar, K., Bose, S., Bera, R., and Chakrabarti, S. (2020). Analysis of pan-omics data in human interactome network (APODHIN). *bioRxiv* [Preprint], doi: 10.1101/2020.04.18.048207

Bovolenta, L. A., Acencio, M. L., and Lemke, N. (2012). HTRIdb: an open-access database for experimentally verified human transcriptional regulation interactions. *BMC Genom.* 13:405. doi: 10.1186/1471-2164-13-405

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

Chou, C.-H., Chang, N.-W., Shrestha, S., Hsu, S.-D., Lin, Y.-L., Lee, W.-H., et al. (2016). miRTarBase 2016: updates to the experimentally validated miRNA-target interactions database. *Nucleic Acids Res.* 44, 239–247.

Ferretti, L., and Cortelezzi, M. (2011). Preferential attachment in growing spatial networks. *Phys. Rev. E* 84:016103.

Gautam, P., Jaiswal, A., Aittokallio, T., Al-ali, H., and Wennerberg, K. (2019). Phenotypic screening combined with machine learning for efficient identification of breast cancer-selective therapeutic targets. *Cell Chem. Biol.* 26, 970–979. doi: 10.1016/j.chembiol.2019.03.011

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

Hern, R., Tarazona, S., Mart, C., Balzano-nogueira, L., Furi, P., Pappas, G. J., et al. (2018). PaintOmics 3: a web resource for the pathway analysis and visualization of multi-omics data. *Nucleic Acids Res.* 46, W503–W509.

Jeong, H., Mason, S. P., Barabási, A.-L., and Oltvai, Z. N. (2001). Lethality and centrality in protein networks. *Nature* 411, 41–42. doi: 10.1038/3507 5138

Kan, Z., Ding, Y., Kim, J., Jung, H. H., Chung, W., Lal, S., et al. (2018). Multi-omics profiling of younger Asian breast cancers reveals distinctive molecular signatures. *Nat. Commun.* 9:1725.

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

Koh, H. W. L., Damian, F., Vogel, C., Choi, K. P., Ewing, R. M., and Choi, H. (2019). iOmicsPASS: network-based integration of multiomics data for predictive subnetwork discovery. *NPJ Syst. Biol. Appl.* 5:22.

Mcgrail, D. J., Federico, L., Li, Y., Dai, H., Lu, Y., Mills, G. B., et al. (2018). Multi-omics analysis reveals neoantigen- independent immune cell infiltration in copy-number driven cancers. *Nat. Commun.* 9:1317.

Mistry, D., Wise, R. P., and Dickerson, J. A. (2017). DiffSLC: a graph centrality method to detect essential proteins of a protein-protein interaction network. *PLoS One* 12:e0187091. doi: 10.1371/journal.pcbi.0187091

Moretti, S., Martin, O., Van Du Tran, T., Bridge, A., Morgat, A., and Pagni, M. (2016). MetaNetX/MNXref - Reconciliation of metabolites and biochemical reactions to bring together genome-scale metabolic networks. *Nucleic Acids Res.* 44, D523–D526.

Perez-Riverol, Y., Csordas, A., Bai, J., Bernal-Llinares, M., Hewapathirana, S., Kundu, D. J., et al. (2019). The PRIDE database and related tools and resources in: Improving support for quantification data. *Nucleic Acids Res.* 47, D442–D450.

Ramazzotti, D., Lal, A., Wang, B., and Batzoglou, K. (2018). Serafim, Sidow A. Multi-omic tumor data reveal diversity of molecular mechanisms that correlate with survival. *Nat. Commun.* 9:4453.

Rohart, F., Gautier, B., Singh, A., and Cao, K. A. L. (2017). mixOmics: an R package for 'omics feature selection and multiple data integration. *PLoS Comput. Biol.* 13:e1005752. doi: 10.1371/journal.pcbi.1005752

Sandhu, C., Qureshi, A., and Emili, A. (2018). Panomics for precision medicine. *Trends Mol. Med.* 24, 85–101. doi: 10.1016/j.molmed.2017.11.001

Shu, L., Zhao, Y., Kurt, Z., Byars, S. G., Tukiainen, T., Kettunen, J., et al. (2016). Mergeomics: multidimensional data integration to identify pathogenic perturbations to biological systems. *BMC Genom.* 17:874. doi: 10.1186/s12864-016-3198-9

Stelzer, G., Rosen, N., Plaschkes, I., Zimmerman, S., Twik, M., Fishilevich, S., et al. (2016). The GeneCards suite: from gene data mining to disease genome sequence analyses. *Curr. Protoc. Bioinform.* 54, 1.30.1–1.30.33.

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, 607–613.

TCGA (2020). Available at: https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga (accessed 22 July 2020).

Tuncbag, N., Braunstein, A., Pagnani, A., Huang, S. S. C., Chayes, J., Borgs, C., et al. (2013). Simultaneous reconstruction of multiple signaling pathways via the prize-collecting steiner forest problem. *J. Comput. Biol.* 20, 124–136. doi: 10.1089/cmb.2012.0092

Uhlen, M., Zhang, C., Lee, S., Sjöstedt, E., Fagerberg, L., Bidkhori, G., et al. (2017). A pathology atlas of the human cancer transcriptome. *Science* 357:660.

Ulfenborg, B. (2019). Vertical and horizontal integration of multi- omics data with miodin. *BMC Bioinform.* 20:649. doi: 10.1186/s12859-019-3224-4

Ulgen, E., Ozisik, O., and Sezerman, O. U. (2019). pathfindR: an R package for comprehensive identification of enriched pathways in omics data through active subnetworks. *Front. Genet.* 10:858. doi: 10.3389/fgene.2019.00858

Vasaikar, S. V., Straub, P., Wang, J., and Zhang, B. (2018). LinkedOmics: analyzing multi-omics data within and across 32 cancer types. *Nucleic Acids Res.* 46, D956–D963.

Vergoulis, T., Vlachos, I. S., Alexiou, P., Georgakilas, G., Maragkakis, M., Reczko, M., et al. (2012). TarBase 6.0: capturing the exponential growth of miRNA targets with experimental support. *Nucleic Acids Res.* 40, D222–D229.

Vinayagam, A., Zirin, J., Roesel, C., Hu, Y., Yilmazel, B., Samsonova, A. A., et al. (2014). Integrating protein-protein interaction networks with phenotypes reveals signs of interactions. *Nat. Methods* 11, 94–99. doi: 10.1038/nmeth.2733

Wishart, D. S., Feunang, Y. D., Marcu, A., Guo, A. C., Liang, K., Azquez-Fresno, R. V., et al. (2018). HMDB 4.0: the human metabolome database for 2018. *Nucleic Acids Res.* 46, D608–D617.

Xie, B., Yuan, Z., Yang, Y., Sun, Z., Zhou, S., and Fang, X. (2018). MOBCdb?: a comprehensive database integrating multi - omics data on breast cancer for precision medicine. *Breast Cancer Res. Treat.* 169, 625–632. doi: 10.1007/s10549-018-4708-z

Yang, Y., Sui, Y., Xie, B., Qu, H., and Fang, X. (2019). GliomaDB: a web server for integrating glioma omics data and interactive analysis. *Genom. Proteom. Bioinform.* 17, 465–471. doi: 10.1016/j.gpb.2018.03.008

Yu, H., Kim, P. M., Sprecher, E., Trifonov, V., and Gerstein, M. (2007). The importance of bottlenecks in protein networks: Correlation with gene essentiality and expression dynamics. *PLoS Comput. Biol.* 3:e59. doi: 10.1371/journal.pcbi.0030059

Zhang, J., Baran, J., Cros, A., Guberman, J. M., Haider, S., Hsu, J., et al. (2011). International cancer genome consortium data portal-a one-stop shop for cancer genomics data. *Database* 2011:bar026. doi: 10.1093/database/bar026

Zhou, G., and Xia, J. (2018). OmicsNet: a web-based tool for creation and visual analysis of biological networks in 3D space. *Nucleic Acids Res.* 46, W514–W522.

**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 December 2020
doi: 10.3389/fgene.2020.595242

# Mechanistic Modeling of Gene Regulation and Metabolism Identifies Potential Targets for Hepatocellular Carcinoma

*Renliang Sun, Yizhou Xu, Hang Zhang, Qiangzhen Yang, Ke Wang, Yongyong Shi* and Zhuo Wang**

*Bio-X Institutes, Key Laboratory for the Genetics of Developmental Neuropsychiatric Disorders (Ministry of Education), Shanghai Jiao Tong University, Shanghai, China*

OPEN ACCESS

**Edited by:**
*Bhabatosh Das, Translational Health Science and Technology Institute (THSTI), India*

**Reviewed by:**
*Somsubhra Nath, Saroj Gupta Cancer Centre and Research Institute, Kolkata, India*
*Kalyan C. Vinnakota, Gilbert Family Foundation, United States*
*Lu Xie, Shanghai Center For Bioinformation Technology, China*

***Correspondence:**
*Yongyong Shi*
*shiyongyong@gmail.com*
*Zhuo Wang*
*zhuowang@sjtu.edu.cn*

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

**Received:** *15 August 2020*
**Accepted:** *30 November 2020*
**Published:** *23 December 2020*

**Citation:**
*Sun R, Xu Y, Zhang H, Yang Q, Wang K, Shi Y and Wang Z (2020) Mechanistic Modeling of Gene Regulation and Metabolism Identifies Potential Targets for Hepatocellular Carcinoma. Front. Genet. 11:595242. doi: 10.3389/fgene.2020.595242*

Hepatocellular carcinoma (HCC) is the predominant form of liver cancer and has long been among the top three cancers that cause the most deaths worldwide. Therapeutic options for HCC are limited due to the pronounced tumor heterogeneity. Thus, there is a critical need to study HCC from a systems point of view to discover effective therapeutic targets, such as through the systematic study of disease perturbation in both regulation and metabolism using a unified model. Such integration makes sense for cancers as it links one of the dominant physiological features of cancers (growth, which is driven by metabolic networks) with the primary available omics data source, transcriptomics (which is systematically integrated with metabolism through the regulatory-metabolic network model). Here, we developed an integrated transcriptional regulatory-metabolic model for HCC molecular stratification and the prediction of potential therapeutic targets. To predict transcription factors (TFs) and target genes affecting tumorigenesis, we used two algorithms to reconstruct the genome-scale transcriptional regulatory networks for HCC and normal liver tissue. which were then integrated with corresponding constraint-based metabolic models. Five key TFs affecting cancer cell growth were identified. They included the regulator *CREB3L3*, which has been associated with poor prognosis. Comprehensive personalized metabolic analysis based on models generated from data of liver HCC in The Cancer Genome Atlas revealed 18 genes essential for tumorigenesis in all three subtypes of patients stratified based on the non-negative matrix factorization method and two other genes (*ACADSB* and *CMPK1*) that have been strongly correlated with lower overall survival subtype. Among these 20 genes, 11 are targeted by approved drugs for cancers or cancer-related diseases, and six other genes have corresponding drugs being evaluated experimentally or investigationally. The remaining three genes represent potential targets. We also validated the stratification and prognosis results by an independent dataset of HCC cohort samples (LIRI-JP) from the International Cancer Genome Consortium database. In addition, microRNAs targeting key TFs and

genes were also involved in established cancer-related pathways. Taken together, the multi-scale regulatory-metabolic model provided a new approach to assess key mechanisms of HCC cell proliferation in the context of systems and suggested potential targets.

**Keywords: regulatory-metabolic integration, metabolic model, molecular stratification, potential therapeutic target, hepatocellular carcinoma, metabolic reprogramming**

## INTRODUCTION

Hepatocellular carcinoma (HCC) is the most common type of primary liver cancer and is the third leading cause of cancer-related death (Ferlay et al., 2015). Obesity, diabetes, fatty liver, virus infection, and many other diseases can lead to HCC. Treatment of HCC largely depends on surgery. Radiochemotherapy is unsatisfactory in part because of the current difficulty in early diagnosis. Furthermore, although drugs like Sorafenib and Lenvatinib had been approved by the Food and Drug Administration (FDA), the drug–response rates are relatively low probably due to the pronounced tumor heterogeneity. For example, in one trial the median survival was only 2–3 months longer compared to the placebo arm in Asians and Caucasians (Cheng et al., 2009). More precise patient stratification and discovery of novel drug targets are necessary to improve treatment outcomes of HCC.

Several recent studies classified the molecular subtypes of HCC based on proteomic data. In one study, the classification of early-stage Chinese HCC samples revealed the mechanism of early tumor cell development (Jiang et al., 2019). In the other study, the classification of hepatitis B virus (HBV)-related HCC samples identified three subgroups with distinct features in metabolic reprogramming, microenvironment dysregulation, and cell proliferation (Gao et al., 2019).

Metabolic reprogramming is an important characteristic and driver of cancer. Genome-scale metabolic models (GEMs) have been successfully used to characterize cancer metabolism and to identify drug targets for cancer treatment. GEMs are a powerful framework to mechanistically represent the relationship between genotype and phenotype by computationally modeling the biochemical constraints imposed on the phenotype. The models are capable of simulating various biological tasks under given conditions (Mardinoglu and Nielsen, 2012, 2015). This allows the identification of essential genes or reactions for a particular objective function. Many disease-related genes and metabolites have been experimentally validated by comparing the altered metabolism between normal and tumor tissue models. Folger et al. (2011) used microarray data to identify key genes for non-small-cell lung cancer. Mardinoglu et al. (2014) utilized data from the Human Protein Atlas Database with the INIT algorithm to successfully construct 69 cell-specific models and 16 cancer-specific models. More recently, Uhle et al. (2017) employed RNA-Seq data from The Cancer Genome Atlas (TCGA) database together with the INIT algorithm to reconstruct 6753 patient-specific metabolic models for various cancers.

Although many anti-cancer drugs developed by target-based approaches have been approved by the FDA (Assoun et al., 2017; Howie et al., 2018), there are still few effective therapeutic targets for HCC. Bidkhori et al. (2018) recently addressed this by utilizing metabolic network topology analysis to divide 179 liver HCC (LIHC) samples from the TCGA-LIHC database into three subtypes and identify potential subtype-specific therapeutic targets. However, metabolic networks are dramatically affected by complex transcriptional regulatory networks, while the changes in transcriptional regulation can lead to changes in enzyme abundance or activity, which in turn lead to changes in physiological states (e.g., cancer cell growth). The close crosstalk between metabolic and regulatory mechanisms during the complex tumor development necessitates the investigation of multi-level mechanisms by integrating both regulation and metabolism. Since the regulatory role of miRNA in liver cancer remains largely in the work-in-progress phase, it is hard to get the full spectrum of dysregulated miRNA in HCC (Sartorius et al., 2019), we focused on the genome-scale transcriptional regulatory network between TFs and genes, which was then mechanistically combined with genome-scale liver metabolic model. Several studies are constructing global transcriptional regulatory networks for liver tissue or HCC tissue (Zhu et al., 2012; Chen et al., 2017), but to our knowledge, no computational studies have integrated regulation and metabolism into a unified genome-scale model in studying HCC.

In this study, schematically summarized in **Figure 1**, we used integrated regulatory-metabolic modeling to investigate the possible mechanism of HCC using all TCGA-LIHC samples. We have previously developed the Integrated Deduced Regulation And Metabolism (IDREAM) algorithm (Wang Z. et al., 2017), which uses a bootstrapping linear regression model on large-scale gene expression datasets (e.g., 2,929 microarray for *Saccharomyces cerevisiae*) to predict TF regulation on enzyme-encoding genes, followed by a probabilistic regulation of metabolism approach to apply regulatory constraints to the metabolic network. The integrated model can predict the influence of each TF knockout on certain objective functions, such as cell growth. The model has been successfully applied in *S. cerevisiae* to effectively predict the influence of transcriptional regulation on the metabolic phenotype. It also can reveal novel synthetic lethal pairs of TFs and metabolic genes with an important interaction mechanism. But IDREAM requires a large-scale expression dataset to infer regulatory network, which is limited for HCC, so we modified it extensively for the application in liver cancer study herein. We inferred the tumor/normal regulatory networks using two independent algorithms, MERLIN and CMIP. Then The regulatory relationships deduced by both algorithms were regarded as "high confidence" regulations and were tagged in the transcriptional regulatory networks

for further integration with the metabolic model. Using the integrated model, we classified HCC patients into different subgroups by expression data of transcription factors (TFs) and genes in the integrated network. The classification results were evaluated by overall survival (OS) outcomes. The integrated regulatory-metabolic model allows the identification of the mechanisms of HCC tumor cell progression, the genes associated with poor prognosis, and potential therapeutic targets. In addition, microRNAs (miRNAs) regulating the influential TFs and metabolic genes were incorporated to validate whether the genes identified by the integrated model were important for HCC tumorigenesis and their value as targets for clinical treatment. The results were consistent with previous *in-silico* and experimental studies.

# MATERIALS AND METHODS

## HCC Gene Expression Data

RNA-Seq expression data were obtained from 315 HCC samples with clinical outcomes from the TCGA-LIHC Project, 232 HCC samples with clinical outcomes from the International Cancer Genome Consortium-Liver Cancer RIKEN (ICGC-LIRI) Project, and 50 HCC paired tumor-normal samples from the Gene Expression Omnibus (GEO) database (GSE77314) (Liu G. et al., 2016). The three gene expression datasets were, respectively, employed to construct the integrated regulatory-metabolic network model. The GSE77314 dataset was also used to infer tumor and normal liver regulatory networks.

## Metabolic Network Models

The genome-scale metabolic model of liver tissue used for integration was retrieved from the Human metabolic Atlas (HMA) Database (the[1]). It was built based on the combination of the HMR2 model with RNA-Seq data of liver tissue to provide an approach to explore metabolic and proteomic functions in cancer (Uhlén et al., 2015). The patient-specific GEMs of HCC used for metabolic analyses were retrieved from the BioModels Database[2]. Uhle et al. (2017) utilized the tINIT algorithm to perform the reconstruction. The characteristics of the metabolic pathways in each model were determined by the protein-coding genes expression level detected from individual patient RNA-Seq data in the TCGA-LIHC Project. Biomass representing cell growth (whose formula was also obtained from Uhle et al., 2017) was set to be the objective function. We selected 315 of 338 HCC individual models with clear clinical stage information (excluding "not reported") for metabolic reprogramming analysis.

## Construction of Regulatory Networks

Two independent algorithms—the modular regulatory network learning with per-gene information (MERLIN) (Roy et al., 2013) and conditional mutual information measurement using a parallel computing framework CMIP (Zheng et al., 2016)—were used to construct the tumor/normal regulatory networks from the expression data (GSE77314), which were implemented using the Part 1 script in **Supplementary File 1**. MERLIN combines the per-gene method and per-module concept based on a probabilistic graphical model to infer regulatory network. Thus, MERLIN cannot include only memberships deduced from individual genes. The algorithm must also take the similarity within a group of genes into consideration. The algorithm is effective in predicting transcriptional changes in human differentiation neural progenitor cells (Roy et al., 2013). In addition, MERLIN outperforms several other state-of-the-art algorithms. We used default settings, except for the use of five-fold cross-validation.

The CMIP algorithm quantifies the interactions between genes on the basis of conditional mutual information measurement to avoid neglecting subtle relations under certain conditions. For example, if both A and B are strongly connected to C, then the actual relationship between A and B may be confusing because of the interference of C. The performance was evaluated by the average Area Under Curve (AUC) of 10 benchmark datasets provided by the DREAM3 algorithm. CMIP performed better than the other algorithms. Additionally, parallelized computation enabled it to handle genome-scale datasets and to complete tasks within a relatively short time compared to other popular methods presented in DREAM3 Projects (Marbach et al., 2009). CMIP was run using default parameters to let the algorithm automatically decide the threshold of the dynamic removal of gene-pairs.

The regulatory relationships deduced by both algorithms were regarded as "high confidence" regulations and were tagged in the regulatory networks for further integration with the metabolic model.

## Metabolic Analysis

The COBRA Toolbox incorporated in MATLAB was used for the metabolic analysis (Heirendt et al., 2019). Flux Balance Analysis predicts feasible phenotypic states by setting appropriate constraints gained from prior knowledge or assigned conditions. By identifying the metabolic task to be studied, the flux distribution of all reactions in the model can be calculated and solved as follows:

$$\begin{aligned} &\text{maximum}: && \text{Cell growth} \\ &\text{subject to}: && S \cdot v = 0 \\ & && a_j \leq v \leq b_j \end{aligned}$$

where v is a flux vector representing a particular flux configuration, S is the stoichiometric matrix, and $a_j$ and $b_j$ are the minimum and maximum fluxes, respectively, through reaction *j*.

We mainly used the "SingleGeneDeletion" function to find metabolic genes whose knockout led to decreased cell growth. The "OptimizeCbModel" function was used to calculate the optimal growth rate and corresponding flux distribution.

## Integration of Regulatory Network and Metabolic Model for HCC

Modeling the regulatory networks of HCC and normal liver tissue required the determination of TFs functioning in liver tissue. To do this, we used liver regulatory network information

[1]https://metabolicatlas.org/gems/repository

[2]https://www.ebi.ac.uk/biomodels/

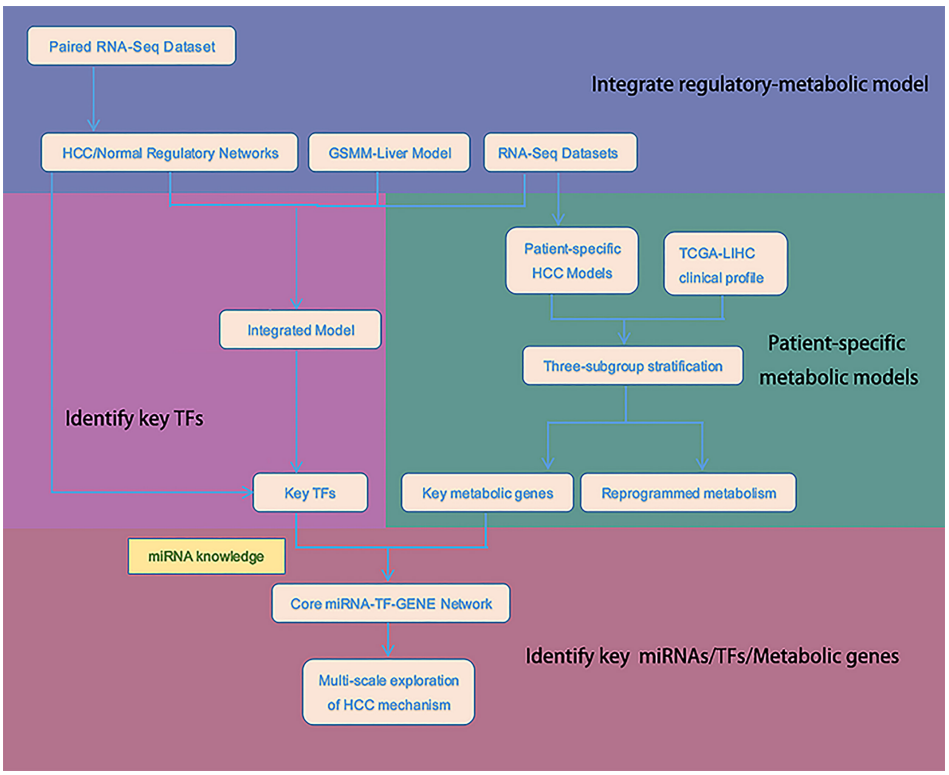



**FIGURE 1 |** The schema of the integrated model for stratification and key targets discovery.

from RegulatoryCircuits (Daniel et al., 2016)[3], which was inferred based on the interactions of TFs-promoters, TFs-enhancers, promoters-genes, and enhancers-genes. The data were validated by introducing ChIP-Seq, expression quantitative trait loci (eQTL), and RNA-Seq data. We also used the human regulatory network from the RegNetwork (Zhi-Ping et al., 2015)[4], which was constructed by considering prior knowledge of TF binding sites and post-transcriptional regulation by miRNAs. In addition, convincing published results were also included.

The union of these two public human regulatory networks yielded 1,366 TFs. We used these 1,366 TFs along with the 2,456 metabolic genes contained in the liver tissue model in the HMA database together with GSE77314 RNA-Seq expression data to determine the regulatory associations in the HCC and normal liver metabolic models. Different from the bootstrapping linear regression model used for regulatory associations inference in IDREAM, here we applied two independent algorithms, MERLIN and CMIP to calculate the interactions. The union of the results predicted by the two methods represented the regulatory network. The overlapping interactions represented 'high confidence' interactions. Then we used the probabilistic regulation of metabolism approach to build the integrated regulatory-metabolic model and predicted TFs affecting cell growth in tumor and normal liver. We first calculated the probability of a target gene being ON when TF was OFF, designated as Prob(Gene = ON| Factor = OFF). The constraints on the corresponding reaction flux were $V_{max}$ × Prob, where Vmax was derived by flux variability analyses. We then simulated the changes in cell growth and each reaction flux. The implementation of the integrated model construction code by MATLAB is provided in Part 2 of **Supplementary File 1**.

## Stratification, Survival, and Analysis of Differentially Expressed Genes (DEGs)

In total, there are 3,492 expressed genes in the integrated regulatory-metabolic network (1,366 TFs and 2,456 metabolic genes), excluding overlapping genes and those with no expression data. The expression data of these 3,492 genes were used to stratify 315 TCGA-LIHC samples using the non-negative matrix factorization (NMF) consensus clustering method from the "NMF" R package (Attila and Mattias, 2008). This machine-learning algorithm aims to distinguish different molecular patterns in high-throughput genomic data. We used 200 iterations to determine the best clustering number between two and 10? and selected the three best-value clusters according to the cophenetic correlation coefficient and average silhouette width.

Clinical outcomes of the TCGA samples were used to evaluate the clustering results. The Kaplan-Meier survival curve implemented in the "survival" R package was applied to assess the OS rate. The NMF clustering subtypes showed significant differences in survival outcomes.

[3]http://regulatorycircuits.org/

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

For the analysis of DEGs, we used a linear model and moderated *t*-statistics based algorithm implemented in the "Limma" package, with absolute value log2(fold change) $\geq$1 and $P \leq 0.05$. We compared the three clusters in pairs and selected the intersection of DEGs between Class2:Class1 and Class2:Class3 as the significantly upregulated/downregulated genes of the subtype with the worst prognosis.

Functional enrichment analyses of the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways were performed using Database for Annotation, Visualization, and Integrated Discovery (DAVID;[5]). Adjusted $P \leq 0.05$ indicated significant enrichment.

## Network Topology Analysis

Cytoscape software was used for topology explorations (Su et al., 2014). The "Tools"–"Merge"–"networks" function with the optional parameter "difference" was used to detect differences between tumor and normal liver networks. The principle was to remove all identical nodes to identify TFs/metabolic genes that were present only in HCC or the normal regulatory network. We highlighted the hub genes being responsible for the abnormity on the topological structure.

# RESULTS

## Differences of Regulatory Networks Between Tumor and Normal Liver Cells

There are many algorithms designed to infer regulatory networks from transcriptome profiles. The results have been validated in model organisms that include *S. cerevisiae* and *Escherichia coli*. We used the MERLIN and CMIP algorithms together with paired RNA-Seq data obtained from the GEO database (GSE77314) (Wang Z. et al., 2017) to construct the regulatory networks of HCC and paired normal tissue, implemented by the Part1 script in **Supplementary File 1**. There were a total of 15,143 pairs and 29,127 pairs of regulation between TFs and target genes deduced from tumor and normal samples (**Supplementary Table 1**). Of these, 1,654 pairs were the same. Cytoscape was used to visualize the topology difference between these two networks. After removing the nodes that had little influence, the core structure was obtained (**Figure 2**). In the core structure, *NME2* and *NFKBIA* were the hub TFs that were important in normal liver models (**Figure 2A**). These two TFs were absent in the HCC tumor model (**Figure 2B**). Nuclear factor κB (NF-κB) affects multiple biological processes by regulating the immune response and inflammation. NF-κB is a hallmark in cancer progression (Fengting et al., 2014). *NFKBIA* is a member of a cellular protein family that can mask the nuclear localization signals of NF-κB and block its binding to DNA. Because of this inhibition ability, *NFKBIA* has long been considered as a tumor suppressor (Laos et al., 2006). *NME*, which is located on chromosome 17q21, is a gene family associated with the suppression of cancer metastasis and invasion (Steeg et al., 1988). In particular, the *NME2* product inhibits metastasis of breast cancer and lung cancer (Hennessy et al., 1991; Krishna et al., 2014). Therefore, the reconstructed regulatory networks effectively revealed the critical known differences between liver cancer and normal tissues. *NME2* and *NFKBIA* represent putative tumor suppressor factors for future studies.

[5]https://david.ncifcrf.gov/

## Integrative Regulatory-Metabolic Network Identified Abnormality of Hippo Signaling as Key Misregulation in HCC

We integrated the regulatory network with metabolic models to identify potential TFs vital to the growth of HCC cells using the source code of Part2 in **Supplementary File 1**. The compositions of the integrated models for HCC and normal liver tissue are listed in **Supplementary Table 1**. The basic metabolism was

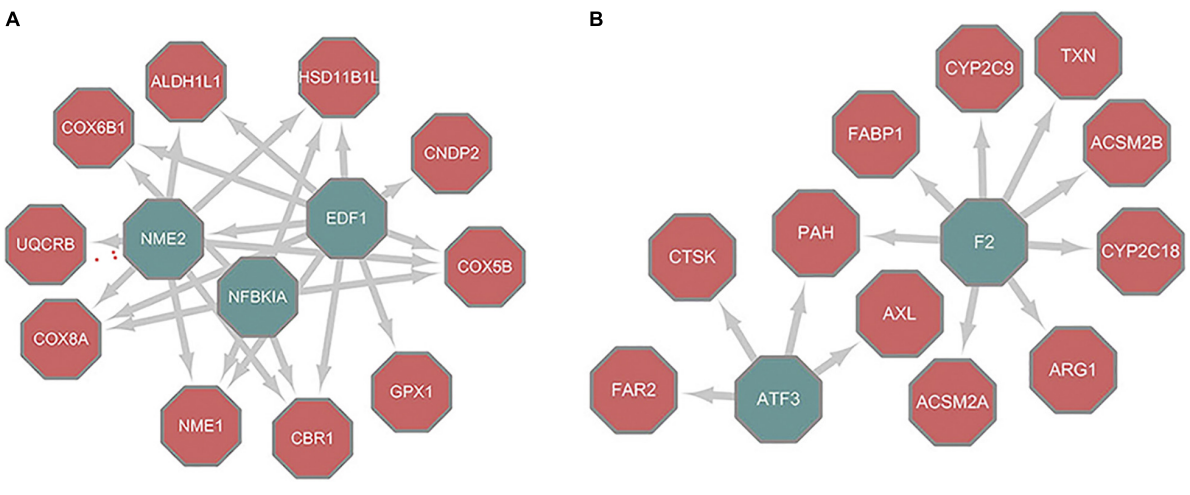



**FIGURE 2 |** Core structure of different nodes between topology of normal/tumor regulatory networks. **(A)** Differences between normal and tumor networks. **(B)** Differences between tumor and normal networks; Nodes filled with green are TFs while nodes filled with red are metabolic genes.

consistent, while the TFs and target genes differed. There were 1313 and 1312 TFs in the HCC and normal model, respectively. These TFs included 33 that were HCC-specific and 32 that were specific for a normal liver.

Reactome database analysis of the 32 specific TFs (including *NME2* and *NFKBIA*) not involved in the HCC regulatory network revealed that they were enriched for the YAP1- and WWTR1 (TAZ)-stimulated gene expression pathways. They are transcriptional co-activators interacting with TEAD family genes to promote the expression of TFs critical to cell proliferation and apoptosis through the Hippo signaling pathway (Lehmann et al., 2016). The findings suggested that depletion of these 32 TFs might lead to abnormal Hippo signaling and might induce a wide range of cancers. In addition, the 33 specific TFs in the HCC integrated model were mainly enriched in cancer metabolism and transcriptional misregulation pathways.

For each TF knockout simulation, we changed the constraints on corresponding reactions according to activation/inhibition interactions and then simulated the cell growth rate to calculate the growth ratio relative to wildtype, as implemented in Part3 script in **Supplementary File S1**. We found TFs affecting both tumor and normal cell growth, as well as TFs that only reduced the growth of tumor cells (**Supplementary Table 2**). For example, disruption of SMAD2, HEY2, ELK1, and CREB3L3 was predicted to lead to >80% reduction in tumor cell growth while having no effect on normal cells. In particular, the involvement of HEY2 and SMAD2 in HBV induced HCC development was evident. The important TFs are likely to be effective targets for the inhibition of tumor cells of HCC.

## Precise Stratification of TCGA-LIHC Samples Based on Metabolic and Transcriptional Gene Expression

The identification of genes or pathways that could be valuable as targets for treatment has been a goal for a long time. Precise clinical diagnosis has been hindered by the pronounced heterogeneity of HCC. This heterogeneity partly reflects the inefficient current TNM stage classification. Molecular stratification of HCC patients and identification of corresponding therapeutic targets are current research goals. Bidkhori et al. (2018) utilized a metabolic network-based method to divide 179 TCGA-LIHC samples into three subtypes and identified their specific characteristics. Jiang et al. (2019) used proteomic data to classify HCC patients and explored the mechanism of an early-stage HCC tumor cell. Here, we used the expression data of 3,492 genes in the integrated model to stratify all the HCC samples with actual clinical survival information from the TCGA-LIHC dataset and to identify altered metabolism among different subgroups and specific characteristics of the poor prognosis subgroup.

Using an NMF consensus clustering analysis, three major classes were identified in the TCGA-LIHC cohort: Class 1 ($n$ = 130), Class 2 ($n$ = 127), and Class 3 ($n$ = 58). The survival curves (**Figure 3A**) revealed a significantly lower OS rate for Class 2 ($P$ = 0.00049). Comparison of the 159 overlapped samples in a previous study (Bidkhori et al., 2018) and this study revealed the relatively good agreement in identifying the lowest OS subgroup: 91% (48 of 53; former results that are also in ours) and 70% (49 of 70, our results that are also in the former findings). Consequently, we focused on determining the characteristics of the Class 2 poor prognosis subgroup at the transcriptome and metabolism levels.

A supervised analysis using Limma (Matthew et al., 2015) revealed 399 differentially expressed genes having distinguishable pattern in Class 2 compared to Class 1 and 3, as shown in the heatmap in **Figure 3B**, comprising 287 upregulated genes [including three potential therapeutic targets: *ALDOA*, *G6PD*, and *ACSS1* specific to the lowest OS subgroup identified by Bidkhori et al. (2018)] and 112 downregulated genes (**Supplementary Table 3**) enriched in 17 and 13 non-overlapping KEGG pathways, respectively (**Figure 3C**).

To validate the effectiveness of our stratification strategy, we applied the same strategy for the LIRI-JP dataset in the ICGC database to form three subgroups with significant prognosis differences (**Figure 4A**; $P$ = 0.0018). We found 332 differentially expressed genes revealed distinguishable pattern between the poor prognosis subgroup and other two subgroups, as shown in heatmap of **Figure 4B**, and the pathways enriched for DEGs were very consistent with DEG-enriched pathways of TCGA-LIHC data (**Figure 4C**). Specifically, the upregulated genes were mainly enriched in established cancer-related pathways involved in improved cell proliferation. Notably, viral carcinogenesis and HBV pathways were upregulated and could be directly linked with HCC development. Another example is increased glucose uptake as a principal nutrient source in central carbon metabolism of cancer, cell cycle, and fructose metabolism. We also found that hypoxia-inducible factor signaling was upregulated. This signaling consists of master regulators of oxygen homeostasis that allow tumor cells to adapt to a hypoxic environment by enhancing oxygen delivery and also affect important growth factors like the *vascular endothelial growth factor* gene. In contrast, the downregulated genes were generally found in pathways contributing to drug metabolism. An example is the peroxisome proliferator-activated receptor

**TABLE 1 |** HCC Cell growth ratio by influential TFs knockouts.

| Ratio after knockout of common TFs in all three classes of TCGA-LIHC | | | |
|---|---|---|---|
| **TF** | **Class 1** | **Class 2** | **Class 3** |
| CTBP1 | 0.926 | 0.926 | 0.926 |
| HTATIP2 | 0.926 | 0.926 | 0.926 |
| ETV7 | 0.234 | 0.12 | 0.09 |

| Ratio after knockout of specific TFs in lowest survival class | | |
|---|---|---|
| **TF** | **TCGA-LIHC** | **LIRI-JP** |
| NR1I3 | 0.978 | 0.978 |
| HNF4A | 0.969 | 0.984 |
| RORC | 0.935 | 0.888 |
| F2 | 0.975 | 0.967 |
| CREB3L3 | 0.856 | 0.876 |

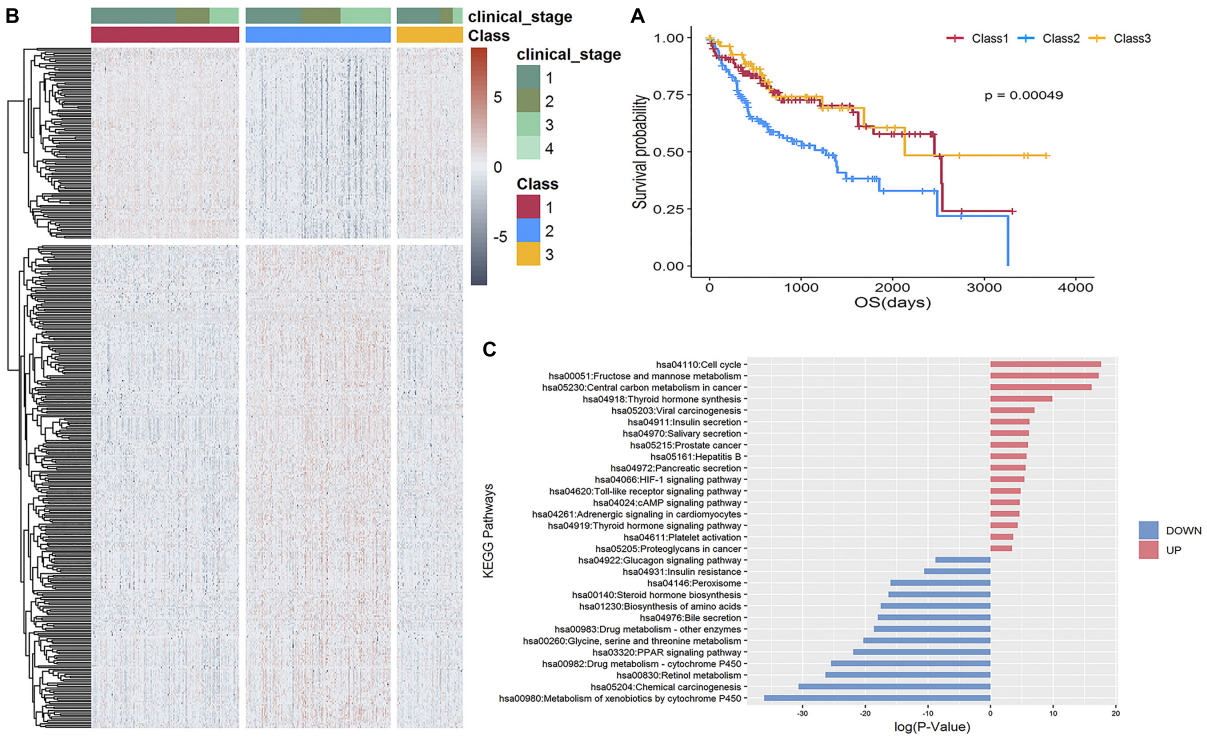



**FIGURE 3 |** Stratification of 315 TCGA-LIHC samples. **(A)** Kaplan-Meier curve shows the survival outcomes of Class 1 (red), Class 2 (blue), and Class 3 (yellow); the *P*-value is 0.00049, which is significant. **(B)** Heatmap of 399 differentially expressed genes revealed a distinguishable pattern between Class 2 and Class 1 and 3. Red color represents upregulated while dark gray represents downregulated. **(C)** Enrichment analysis of KEGG pathways of 287 upregulated and 112 downregulated genes, respectively.

signaling pathway, which has also been identified in less aggressive HCC subtypes through proteomics analysis, as well as drug cytochrome P450 metabolism, which is reduced in advanced cancer patients (Rivory et al., 2002).

## Phosphoinositide 3-Kinase (PI3K)-Akt and Mammalian Target of Rapamycin (mTOR) Signaling Pathways Are Critical to HCC Tumor Cell Growth

By using IDREAM, eight, 13, and five TFs were identified as being vital for HCC cell growth in Class 1, Class 2, and Class 3, respectively, of TCGA-LIHC, after excluding TFs that also affected normal tissue. Three TFs were common in all three classes (**Table 1**).

The knockout of *ETV7* produced the greatest decrease in growth rate in all three classes, as shown in **Table 1**. *ETV7* is a TF belonging to the ETS family, which is responsible for the development of different tissues as well as the progression of several cancers, such as HCC (Peeters et al., 1997; Matos et al., 2009). Due to its translocation function, the overexpression of *ETV7* has been associated with tumorigenic transformation and restriction of apoptosis by blocking the Mys-induced apoptosis pathway (Cardone et al., 2005; Carella et al., 2006; Federica et al., 2018). Accumulating experimental evidence indicates that *ETV7* also plays a significant role in the mTOR signaling pathway by assembling the mTOR3 complex, which can stimulate cell proliferation and is not sensitive to rapamycin, a common anti-tumor agent, unlike mTOR1/2 (Harwood et al., 2018). Therefore, *ETV7* depletion may cause the inactivation of mTOR3 and lead to tumor cell death after treatment.

*CTBP1* is a well-known cancer hallmark. The gene is linked with a pro-tumorigenic process and can affect the regulatory network (Blevins et al., 2017). It can bind to the C-terminus of the adenovirus protein E1A to promote cell proliferation and invasion (Hildebrand and Soriano, 2002). In addition, the characteristic elevated NADH level of cancer cells makes it possible for *CTBP1* to bind to NADP with a high affinity, thus triggering a conformational change that leads to hyper-activity of both tumorigenesis and tumor progression.

To explore the characteristics of TFs leading to low survival rate and poor prognosis, we selected TFs whose knockout only influenced samples in Class 2 of the TCGA-LIHC dataset. Five TFs were specific for Class 2 (threshold: ratio <0.98). Four of the five TFs were also associated with the lowest survival subgroup (Class 3) of the LIRI-JP dataset (the ratio of HNF4A somewhat exceeded the threshold), as shown in **Table 1**.

The knockout of *CREB3L3*, which was predicted to decrease the growth rate of tumor cells by over 15% but which had no effect on normal tissues, is reportedly activated with PPARα for lipid metabolism in liver-specific tissue (Vecchi et al., 2013). They both play important roles in the utilization of fatty acid

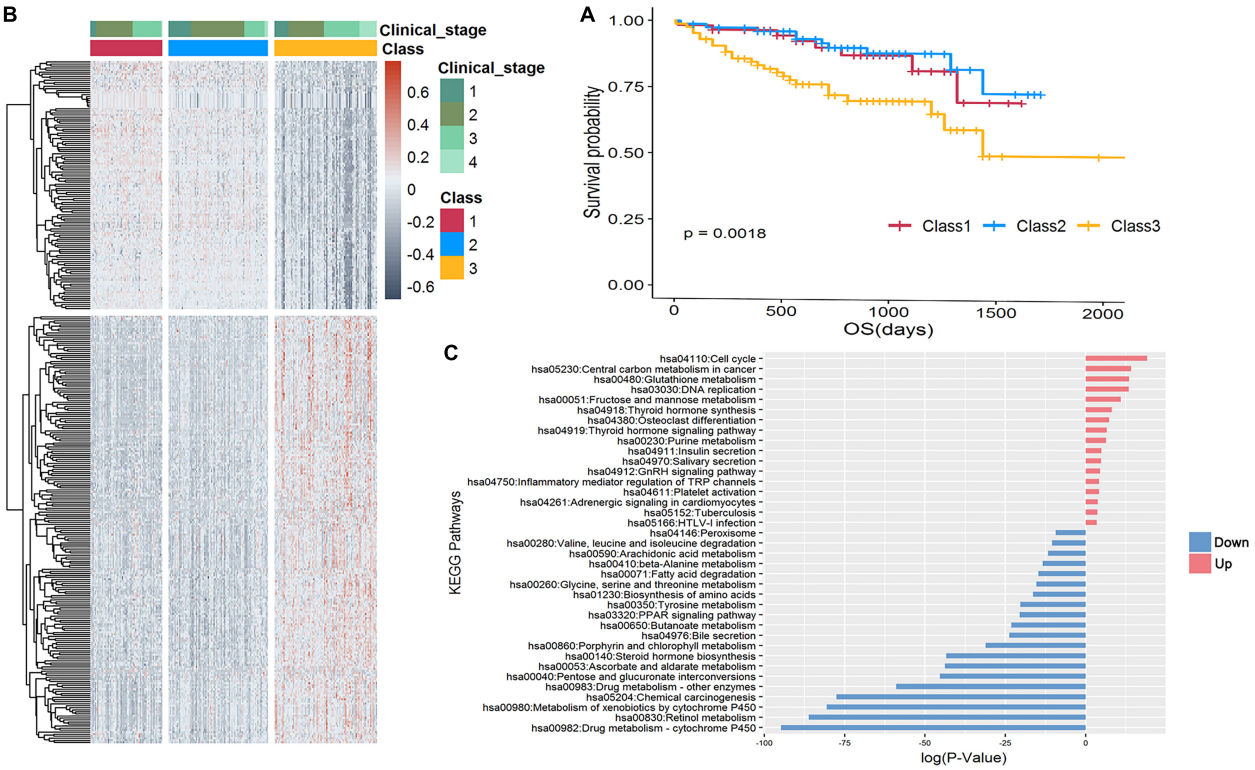



**FIGURE 4 |** Stratification of 232 LIRI-JP samples. **(A)** Kaplan-Meier curve shows the survival outcomes of Class 1 (red), Class 2 (blue), and Class 3 (yellow); the *P*-value is 0.0018, which is significant. **(B)** Heatmap of 332 differentially expressed genes revealed a distinguishable pattern between Class 3 and Class 1 and 2. Red color represents upregulated while dark gray represents downregulated. **(C)** Enrichment analysis of KEGG pathways of 210 upregulated and 122 downregulated genes, respectively.

for energy in a fasting state and in cell proliferation. Thus, it was not surprising that its absence was predicted to result in a decreased growth rate in tumor cells. The expression of *CREB3L3* is linked with restricted apoptosis, cell survival, and HBV-associated HCC development by regulating hepatic genes in the PI3K-Akt and AMPK signaling pathways. The alignment of *in-silico* analyses and biological knowledge suggests that *CREB3L3* is a potential therapeutic target, especially for advanced-stage HCC patients.

## Metabolic Genes in Cholesterol Biosynthesis Are Druggable Targets in HCC Treatment

We incorporated patient-specific models established by Uhle et al. (2017) to do metabolic analyses, including identification of metabolic genes essential for tumor cell growth and annotation of the specific reactions altered during tumor development. All 315 metabolic models were built to represent tumor growth. Using the genetic human metabolic model HMR2 and RNA-Seq expression data from TCGA-LIHC, a task-driven model reconstruction algorithm called tINIT was employed to construct all models.

We performed a single gene deletion simulation using a function provided in the COBRA Toolbox. The total gene number of each model ranged from 1,106 to 2,169. We first identified the essential genes in the three subtypes of TCGA-LIHC samples calculated by the NMF stratification strategy. We then collected genes that were essential in at least half of the samples in each class. Nineteen, 20, and 18 genes remained for Class 1, Class 2, and Class 3, respectively, after filtering those having no influence on tumor cell growth. Of these, the 18 genes found in Class 3 (relatively high OS rate) were are also found in the other two classes. *ACADSB* was shared by Class 1 and Class 2, and *CMPK1* was only identified in Class 2. We assessed prior knowledge about the therapeutic potential of these 20 genes in DrugBank[6]. The findings are summarized in **Table 2**.

The DrugBank analysis identified 11 genes (*CRAT, EBP, ACADSB, CMPK1, SLC22A5, HMGCR, HSD17B7, NSDHL, DHCR7,FDPS*, and *CYP51A1*) that have already been targeted by approved drugs in the treatment of cancer or relative diseases. Six other genes have corresponding drugs being evaluated experimentally or investigationally. Both *CMPK1* and *ACADSB* seem to be vital to tumor cell growth in HCC models with a lower survival rate. These genes have been implicated as prognosis biomarkers associated with worse survival in multiple tumors for a long time (Ryu et al., 2011; Ohmine et al., 2015;

[6]https://www.drugbank.ca/

**TABLE 2 |** Lethal metabolic genes as potential targets and corresponding drugs in DrugBank.

| Lethal gene | Target drug | Drug state | Brief description of drug |
|---|---|---|---|
| IDI1 | Dimethylallyl diphosphate | Experimental | Unknown |
| SQLE | Ellagic acid | Investigational | Antioxidant and anti-proliferative/anti-cancer effects |
| FDFT1 | TAK-475 | Investigational | Target rate-limiting enzyme in the hepatic biosynthesis of cholesterol |
| CRAT | Levocarnitine | Approved | Treatment of primary systemic carnitine deficiency |
| EBP | Tamoxifen | Approved | Treatment of metastatic breast cancer |
| ACADSB | Isoleucine | Approved | Anti-proliferative effects useful in cancer therapy |
| | Valproic Acid | Approved | |
| SLC22A5 | Levocarnitine | Approved | Treatment of primary systemic carnitine deficiency, affect fatty acid synthesis |
| HMGCR | Lovastatin | Approved | Lowering LDL cholesterol and triglycerides, hypercholesterolemia; |
| | Cerivastatin | Approved | Target rate-limiting enzyme in the hepatic biosynthesis of cholesterol |
| | Simvastatin | Approved | |
| | Atorvastatin | Approved | |
| | Rosuvastatin | Approved | |
| | Meglutol | Experimental | |
| CMPK1 | Gemcitabine | Approved | Various advanced cancers |
| | Lamivudine | Approved | Treatment of HBV |
| | Sofosbuvir | Approved | Treatment of HCV |
| | | | Reduce incidence of HCC |
| MVK | Farnesyl thiopyrophosphate | Experimental | Unknown |
| HSD17B7 | NADH | Approved | Treating Parkinson's disease, chronic fatigue syndrome, Alzheimer's disease and cardiovascular disease |
| NSDHL | NADH | Approved | Treating Parkinson's disease, chronic fatigue syndrome, Alzheimer's disease and cardiovascular disease |
| DHCR7 | NADH | Approved | Treating Parkinson's disease, chronic fatigue syndrome, Alzheimer's disease and cardiovascular disease |
| ACACB | Soraphen A | Experimental | Anti-HCV viral activity |
| LSS | R048-8071 | Experimental | Unknown |
| | Lanosterol | Experimental | |
| FDPS | Pamidronic acid | Approved | Treating severe hypercalcemia of malignancy |
| | Zoledronic acid | Approved | Treating Paget's disease of bone |
| | Alendronic acid | Approved | Treating bone metastases from solid tumors |
| | Ibandronate | Approved, investigational | Treating osteolytic lesions of multiple myeloma |
| | Risedronic acid | Approved, investigational | Experimental drugs' targets are still unknown |
| | ISOPENTENYL PYROPHOSPHATE | Experimental | |
| | Dimethylallyl Diphosphate | Experimental | |
| | Farnesyl diphosphate | Experimental | |
| | Geranyl Diphosphate | Experimental | |
| | Geranylgeranyl diphosphate | Experimental | |
| | Isopentyl Pyrophosphate | Experimental | |
| | Incadronic acid | Experimental | |
| CYP51A1 | Levoketoconazole | Investigational | Treating fungal infections in immunocompromised and non-immunocompromised patients |
| | (S)-econazole | Experimental | Treating diabetes mellitus type 2. |
| | Miconazole | Approved, investigational, vet_approved | |
| | Itraconazole | Approved, investigational | |
| | Tioconazole | Approved | |
| PMVK | Unknown | Unknown | Unknown |
| MVD | Unknown | Unknown | Unknown |
| SC5D | Unknown | Unknown | Unknown |

Liu N.Q. et al., 2016; Zhou et al., 2017a,b; Zhang B. et al., 2019). *CMPK1* is also the target of three FDA approved cancer drugs (Gemcitabine, Lamivudine, and Sofosbuvir) for the treatment of diseases induced by a virus infection, such as HCC caused by HBV/hepatitis C virus. Li et al. (2019) recently reported that in Kaposi's sarcoma, a common acquired-immunodeficiency-syndrome-related malignancy caused by infection of Kaposi's sarcoma-associated herpesvirus, the invasiveness and motility of cells can be increased by overexpression of *CMPK*. This effect has also been validated by the knockout experiments carried out in cell lines. *FDPS* has been targeted by 11 drugs, among which five types of drugs are approved for mainly treating osteoporosis as well as bone metastases from solid tumors. *CYP51A1* has been the targets of three approved drugs, which are mainly used for treating fungal infections.

Among the 18 genes lethal in all three classes, 15 genes participate in cholesterol biosynthesis via the desmosterol (DESMOL) pathway, which is the dominant form of liver cholesterol biosynthesis (Song et al., 2005). The *HMGCR, MVK, PMVK, MVD*, and *IDI1* genes involving in the mevalonate pathway that converts acetyl-CoA to dimethylallyl pyrophosphate (DMAPP). The enzyme encoded by *FDPS* aids DMAPP in synthesizing farnesyl pyrophosphate (FAPP). *FDFT1* catalyzes the dimerization of two FAPP into squalene (SQNE). In the next step, *SQLE* and *LSS* play important rate-limiting roles in cholesterol biosynthesis by catalyzing the conversion of SQNE to lanosterol (LNSOL). LNSOL then goes through demethylation, oxidation, and reduction steps catalyzed by *CYP51A1*, *NSDHL*, and *HSD17B7* to form zymosterol (ZYMOL), the precursor in the DESMOL pathway. The *EBP, SC5D,* and *DHCR7* gene catalyze the conversion of ZYMOL to DESMOL. Finally, DESMOL is reduced by *DHCR24* to produce cholesterol. Knockout of any of these genes will disrupt cholesterol biosynthesis and lead to the depletion of cholesterol, which is disastrous for tumor cell growth.

There are only three predicted essential genes that have not been recorded in DrugBank. The high hit rate of drug targets (17/20) suggested that those three metabolic genes are potential targets and worthy of exploration in future studies. As mentioned above, *PMVK* and *MVD* are involved in the mevalonate pathway that converts acetyl-CoA to DMAPP. It has been reported that a key enzyme *HMGCR*, in the mevalonate pathway was confirmed to be closely related to cancer (Jiang et al., 2019). These three genes together help the transformation from Mevalonic acid to Isopentenyl diphosphate (IPP). *SC5D* catalyzes a dehydrogenation to introduce C5-6 double bond into lathosterol in cholesterol biosynthesis. Krakowiak et al. (2013) found that the mouse with *SC5D* disruption had elevated lathosterol, decreased cholesterol levels, and abnormal hedgehog signaling, which is considered to be related to tumorigenesis (Patrycja et al., 2003). Furthermore, *SC5D* regulates the enzyme converting lathosterol to 7-Dehydrocholesterol. And the downstream gene *DHCR7*, which converts 7-Dehydrocholesterol to cholesterol, has been targeted by drugs inhibiting HBV infection (Xiao et al., 2020). Therefore, although the three genes are currently not targets of existing drugs, they are all related to the main-effect pathway cholesterol biosynthesis and important for tumorigenesis of HCC.

## Enhancement of Glutathione and Fatty Acid Biosynthesis Are Important Metabolic Reprogramming Events Associated With Poor Prognosis

It is widely accepted that tumor cells reprogram metabolic pathways to enable unlimited cell proliferation, aggressive invasiveness, and restricted apoptosis. We investigated 1,329 reactions in all 315 patient-specific models to identify flux patterns and enzymes that differed between the poor prognosis subgroup (Class 2) and the other two classes. We conducted flux balance analysis with cell growth as the objective function to calculate the flux distribution for each patient, and selected candidate reactions having similar flux changes in over half samples of each subgroup. Four flux alteration patterns were evident. The first was from negative flux value in Class 1 and Class 3 to positive flux in Class 2. The second was from positive flux value in Class 1 and Class 3 to negative flux in Class 2. The third was from a non-zero flux in Class 1 and Class 3 to zero flux in Class 2. The fourth was from zero flux in Class 1 and Class 3 to non-zero flux in Class 2. The altered reactions, formulas, enzymes, and corresponding types of flux patterns are shown in **Supplementary Table 4**.

Two reactions simulated type 1 and type 2 flux change, respectively. According to these four reactions, the production of glutathione (GSH) was suspected to increase in Class 2 samples due to the enhancement of AKG biosynthesis and cysteine accumulation in the cytosol. GSH is a key member of the cell immune response system. The lack of GSH can easily lead to cell death. Several labs have confirmed its common occurrence in all cancers (Mehrmohamadi et al., 2014) and it is considered a potential therapeutic target. Additionally, loss of the enzyme catalyzing these reactions (encoded by *SLC25A11*) inhibits tumor cell growth in non-small cell lung cancer (Lee et al., 2019). Baulies et al. (2018) suggested that the overexpression of *SLC25A11* works as an adaptive mechanism of HCC to provide enough GSH for abundant cell growth, while *SLC25A11* induces the export of AKG to the cytosol to activate the mTOR pathway to promote cell growth and anabolism through egl-9 family hypoxia-inducible factors (EGLNs) (Villar et al., 2015).

Eleven reactions displayed no flux in Class 2 but a positive flux in Class 1 and 3 (type 3). Four of these reactions are part of porphyrin metabolism. The enzyme encoded by *UROD* is involved in this pathway and was recently identified as a potential anti-cancer target due to its ability to convert uroporphyrinogen to coproporphyrinogen (Yip et al., 2014). Another enzyme encoded by *ALAD* is overexpressed in breast cancer patients with a favorable clinical outcome. Its upregulation can suppress cell proliferation and invasion (Ge et al., 2017). In addition, a set of enzymes responsible for carnitine shuttling, which are encoded by *SLC22A1, SLC25A20, SLC25A29*, and *CPT2*, are downregulated in HCC tumor cells. These enzymes play rate-limiting roles in controlling fatty acid oxidation (Meihua et al., 2018). Their low expression has been significantly

associated with worse patient survival (Heise et al., 2012) and differentiation state by impairing production of nitric oxide and the mTOR signaling pathway mediated by arginine. In some situations, this can lead to severe autophagy (Lifeng et al., 2016; Keshet and Erez, 2018).

Three reactions displayed non-zero flux in Class 2 but zero flux in Class 1 and 3 (type 4). These involved fatty acid activation responsible for providing adequate ATP and CoA for tumor cell growth; glycine, serine, threonine metabolism, which helps reduce reactive oxygen species pressure through the serine–glycine-one-carbon metabolic network during tumor metastasis (Amelio et al., 2014); and arginine/proline metabolism, which can regulate response to nutrient and oxygen deprivation in oncogenesis, thus avoiding cell apoptosis (Phang et al., 2015). Furthermore, exploration of enzymes revealed that *ACADSB* (which was also highlighted by previous analyses), *ACSL3*, and *ACSL4* regulate proteins that stimulate tumor cell proliferation, including p-AKT, LSD1, and β-catenin (Wu et al., 2015).

The altered reactions specific to Class2 samples promote tumor cell growth and decrease sensitivity towards normal apoptosis signals. Several key enzymes have already been implicated as biomarkers in cancers. Metabolic reprogramming accounting for poor prognosis also supports our stratification of the HCC patients.

## miRNAs Regulating Influential Genes for HCC Cell Proliferation

To investigate the interplay between regulation and metabolism of HCC further, we retrieved miRNAs regulating the influential genes highlighted in our previous analyses. These include the three common TFs that were influential in all three classes, the five overlapping TFs that specifically affected the lowest survival subgroup of TCGA-LIHC (Class 2) and LIRI-JP (Class 3), and the 20 metabolic genes revealed by single-gene deletion result (**Supplementary Table 5**). Evaluation of the MIRNET database identified the miRNAs functioning in liver tissue with higher connections to target TFs/genes. We found six miRNAs connected to the 28 genes of interest (**Supplementary Table 5**). Three of these were directly linked with HCC. MiR-124-3p and miR-1-3p have been reported to be downregulated in HCC patients compared to normal subjects (Lang and Ling, 2012; KöBerle et al., 2013). MiR-24-3p is involved in an HCC diagnosis panel because of its abnormal overexpression.

The specific mechanism concerning how the loss-of-function or gain-of-function of these miRNAs contribute to tumorigenesis remains unclear. However, there are some experiment-based hypotheses. **Figure 5A** depicts the core network comprising miRNAs, TFs, and genes involved in HCC. The data will inform further studies in HCC development.

In particular, miRNA-124-3p appears to be the key miRNA during oncogenesis in many cancers (Murakami et al., 2006; Dai et al., 2009; Vlierberghe et al., 2010). Zheng et al. (2012) opined that miR-124-3p participates in reducing tumor cell motility and invasion by controlling epithelial–mesenchymal cell transition as well as cytoskeletal events through a cpG-island methylation (Furuta et al., 2010).

Zhang H. et al. (2019) suggested that miR-1-3p overexpression can inhibit cell proliferation and induce apoptosis by targeting the PI3K-Akt and mTOR pathways through ETV7. The downregulation of mir-24-3p can assist this process by deactivating the Fas receptor in the NOTCH pathway and inhibiting *HNF4A* to drive a feedback loop that leads to cancer-related inflammatory reaction (Salam et al., 2016). Additionally, Wang G. et al. (2017) and Chen et al. (2016) indicated that the regulatory impact of miR-24-3p includes an altered cell cycle by inducing p53 mutation as well as the avoidance of cell death by targeting the Fas receptor in the NOTCH pathway (Nicolas et al., 2003).

In addition, miR-26b-5p, which was connected to nine of the 28 genes, has been experimentally validated to be under-expressed in HCC patients with a worse prognosis. It can suppress tumor invasion as well as inducing apoptosis by targeting *SMAD1* (Wang et al., 2016), which is consistent with our conclusion about the *SMAD* gene. Three of the genes obtained by our integrated regulatory-metabolic analysis (*CMPK1, ACADSB*, and *RORC*) are directly regulated by miR-26b-5p. The fact that they are all specific genes for Class 2 (the class with the worst OS rate) substantiates the previous association.

Ingenuity Pathway Analysis (Krämer et al., 2013) of the 28 candidate genes and six top-connected miRNAs was performed to explore the biological connection among them. As shown in **Figure 5B**, *EGFR* was inferred and linked with our core gene set. *EGFR* is one of the most crucial genes responsible for cancer cell growth. Its overexpression can lead to unlimited cell proliferation, just like that in tumor cells. The gene is a potential therapeutic target in cancer therapy. Multiple FDA approved drugs, such as Gefitinib and Lapatinib, are effective in EGFR-related non-small-cell lung cancer and several other cancers (Rawluk and Waller, 2018; Voigtlaender et al., 2018).

# DISCUSSION

## Integrated Regulatory-Metabolic Network Differences Between HCC and Normal Liver Cells

The curated information linking the reactions of genes and proteins in GEMs has enabled the identification of many potential disease-related biomarkers by metabolic analyses. The interconnectedness between metabolism and regulation permits the integration of regulatory with metabolic models, which in turn allows the more precise description of the phenotypic impact of mutations and environmental perturbations. This integration has proven effective in model organisms, including *S. cerevisiae* and *E. coli*, but has not yet been applied to the study of human diseases.

Here we leveraged the mechanistic modeling of transcriptional regulatory network and metabolic network for HCC study, by extensively improving our IDREAM framework. We used two different approaches to construct transcriptional regulatory networks for HCC and normal liver tissue samples. Through

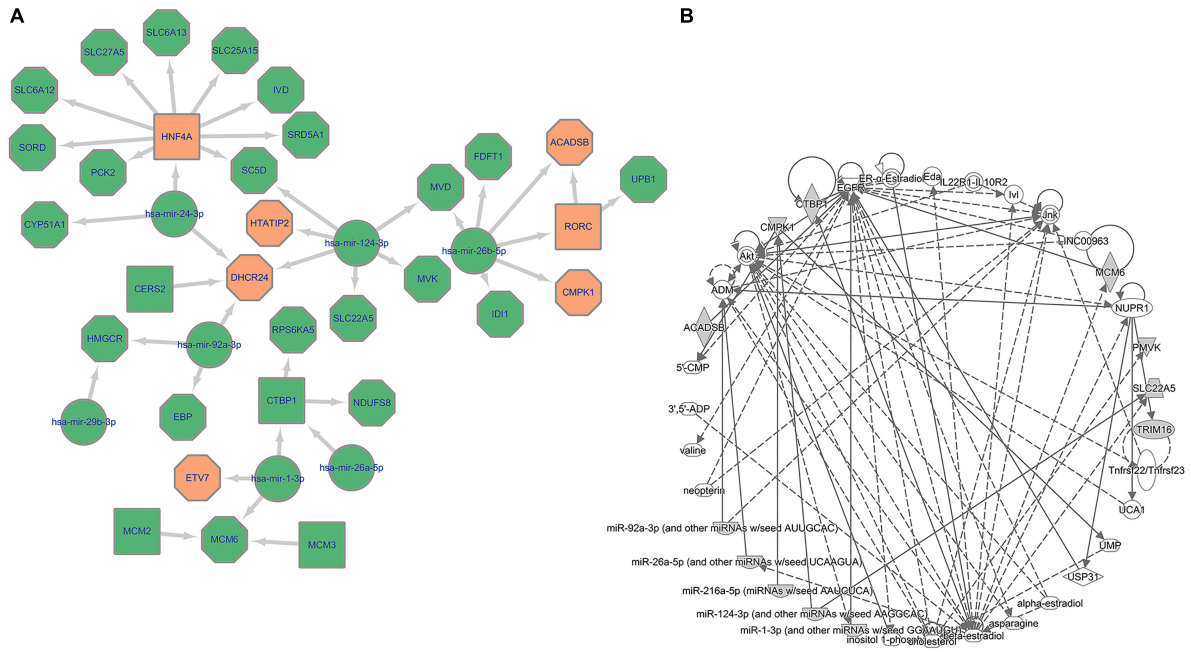



**FIGURE 5 |** Multi-scale network exploration of HCC mechanism. **(A)** Core miRNA relation network of potential therapeutic targets. The octagon represents metabolic genes while the square represents TFs and the circle represents miRNAs; The orange color represents genes/TFs that have been experimentally validated as crucial genes in HCC tumor cells. **(B)** Biological connection of 34 genes (28 candidate genes and top 6 miRNAs).

topology analysis, *NME2*, and *NFBIKA* were implicated as tumor suppressor TFs because of their absence in a tumor regulatory network and high connectivity in a non-tumor network. We integrated the regulatory networks with a human liver metabolic model, and compared the effects of TFs on cell growth in tumor and normal models. TFs that only reduced the growth of tumor cells were predicted to be potential targets. These included the *SMAD2, HEY2, ELK1*, and *CREB3L3* genes.

## Three Subtypes of HCC Samples Demonstrate Significantly Different Prognosis

By allocating TCGA-LIHC samples using pre-filtered 3,492 genes, we defined three patient subgroups distinguished by the OS rate. Patients in Class 2 displayed the worst survival. We identified three essential TFs for HCC tumor cell growth that were common in all three groups. Among these, *ETV7* displayed the greatest impact, decreasing cell growth rate by approximately 88% in Class 2. *ETV7* is a TF belonging to the ETS family. It is responsible for the progression of several cancers, including HCC. Because of its translocation function, the overexpression of *ETV7* has been associated with tumorigenic transformation and restricted apoptosis by blocking the Mys-induced apoptosis pathway. There is growing evidence of a significant role of *ETV7* in the mTOR signaling pathway, which involves the assembly of the mTOR3 complex to stimulate cell proliferation and prevent cell damage by rapamycin, a common anti-tumor agent.

In addition, we identified potential TFs related to poor prognosis based on the simulated knockouts of five TFs, which were predicted to specifically affect patients in Class 2. Among these five TFs, *CREB3L3* was also predicted as being influential for advanced-stage HCC samples by the TF knockout simulation in the generic integrated regulatory-metabolic model. It has been reported that the expression of *CREB3L3* is linked with cell survival and HBV-associated HCC development by regulating hepatic genes in the PI3K-Akt and AMPK signaling pathways (Vecchi et al., 2013).

The poor prognosis group (Class 2) also exhibited a specific pattern of altered metabolism. Flux alterations in Class 2 samples included the accumulation of both AKG and cysteine, which indicated the over-production of GSH, a key member of the cellular immune response system that improves cell proliferation and avoids apoptosis. Besides the biosynthesis of fatty acids, mTOR signaling was also hyper-activated, and pathways that included those of glycine, serine, and threonine metabolism reduce reactive oxygen species stress during tumor homeostasis.

We used the same stratification strategy for the LIRI-JP dataset. Survival outcomes likewise displayed significant differences among the subgroups. The predicted outcomes of TFs affecting the lowest survival subgroup were consistent with that of the TCGA-LIHC dataset.

## Key Metabolic Genes in Cholesterol Biosynthesis Identified by Patient-Specific Models Are Potential Targets

The metabolic analyses based on patient-specific models revealed 20 metabolic genes with important roles in HCC tumor

cell growth by participating in the cholesterol biosynthesis pathway. Recent research uncovered that cholesterol biosynthesis supports the growth of hepatocarcinoma lesions depleted of fatty acid synthase, concomitant targeting de novo lipogenesis and cholesterol biosynthesis are highly detrimental for the growth of human HCC cells (Che et al., 2019)

According to DrugBank, eleven genes have already been therapeutically targeted in various cancers or cancer-related diseases, and six other genes have corresponding drugs being evaluated experimentally or investigationally. Although the remaining three genes, *PMVK*, *MVD*, and *SC5D* are currently not targets of existing drugs, they are all related to the main-effect pathway cholesterol biosynthesis and important for tumorigenesis of HCC, which might become novel potential therapeutic targets and worthy of exploration in future studies. We further found that *ACADSB* and *CMPK1* appeared to be specifically essential in Class 2. These two genes could be associated with poor prognosis and may be the targets for the treatment of more serious HCC patients.

### Multi-Scale Regulatory-Metabolic Network Reveals a Critical Mechanism of HCC Cell Proliferation

In addition to the integration of transcriptional regulation with metabolism, it is well known that dysregulated miRNAs also played an important regulatory role in tumorigenesis. We incorporated the miRNAs regulating the identified influential TFs and metabolic genes generated from an integrated transcriptional regulatory-metabolic network model. Based on the highlighted genes (total of 28 key genes), we predicted miRNAs regulating these candidates using MIRNET. Three miRNAs (miR-124-3p, miR-1-3p, and miR-24-3p) have been described as important factors associated with HCC tumorigenesis and function in established cancer-related pathways, including NOTCH, PI3K-Akt, and mTOR. We illustrated the core network of HCC cell proliferation involving interactions between miRNAs-TFs, miRNAs-Targets, and TFs-Targets (**Figure 5A**), and emphasized the targets that were highlighted in the combined analyses. In general, the inhibition of miRNAs on overexpressed genes in HCC were consistent with their validated function such as suppressing tumorigenesis. The findings suggest potential mechanisms associating the key genes predicted from our regulatory-metabolic network analysis with cancer cell growth outcomes. Notably, the direct regulation of miR-26b-3p on *ACADSB* and *CMPK1* provides experimental evidence to support the idea that these two metabolic genes are linked with lower OS in HCC. Moreover, the biological connection inferred by the Ingenuity Pathway Analysis indicated these highlighted genes are closely connected to *EGFR*, which plays a significant role in cancer cell proliferation, providing evidence for our comprehensive analyses.

## DATA AVAILABILITY STATEMENT

The expression data that support the findings of this study are available in GEO Datasets with the identifier (doi: 10.18632/oncotarget.8927) (Liu G. et al., 2016). The expression and clinical data that support the findings of this study are available in TCGA Database, LIHC section (https://portal.gdc.cancer.gov/). The expression and clinical data that support the findings of this study are available in ICGC database, JIRI-JP section (https://icgc.org/). The genome-scale metabolic liver model that support the findings of this study is available in HMA database (https://metabolicatlas.org/gems/repository). The patient-specific metabolic models that support the findings of this study are available in BioModels Database (https://www.ebi.ac.uk/biomodels/) with the identifier (doi: 10.1126/science.aan2507) (Uhle et al., 2017).

## AUTHOR CONTRIBUTIONS

RS was responsible for the data gathering, model integration, metabolic analysis, and manuscript writing. YX was responsible for regulatory network inferring, statistical analysis, network topology analysis, and literature mining. HZ was responsible for the TCGA and ICGC RNA-Seq the data gathering and the data preprocessing. QY was responsible for the IPA analysis. KW was responsible for the survival analysis. YS involved in design and management of the project. ZW was responsible for the design of the integrated model, explanation of results, and manuscript writing. All authors contributed to the article and approved the submitted version.

## FUNDING

This work was supported by the National Key Research and Development Program (2019YFA0905400), National Natural Science Foundation of China (32070679), Shanghai Municipal Science and Technology Major Project (2017SHZDZX01), the National Key Research and Development Program (2017YFC0908105), the Natural Science Foundation of China (U1804284, 81421061, 81701321, 31571012, and 81501154), the Shanghai Natural Science Funding (16ZR1449700), the Shanghai Hospital Development Center (SHDC12016115), the Shanghai Science and Technology Committee (17JC1402900 and 17490712200), and the Shanghai municipal health commission (ZK2015B01 and 201540114). The funder had no role in study design, the data collection and analysis, decision to publish, or preparation of the manuscript.

## ACKNOWLEDGMENTS

We appreciate Dr. Shuyi Ma for the valuable discussion and editing of the manuscript. This manuscript has been released as a pre-print at ResearchSquare (Xu et al., 2020).

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.595242/full#supplementary-material

**Supplementary File 1 |** The scripts for the integrated regulatory-metabolic model construction and prediction of TF/Gene knockout effects on cancer cell growth.

**Supplementary Table 1 |** Composition of normal/HCC liver model: The regulatory network of normal/HCC liver and the composition of integrated models.

**Supplementary Table 2 |** TFs affecting normal/HCC liver cell growth and corresponding growth ratio after TFs knocking down.

**Supplementary Table 3 |** Differentially expressed genes in Class2 compared to Class 1 and Class 3 in TCGA-LIHC samples.

**Supplementary Table 4 |** Altered metabolic reactions in Class2 compared to Class 1 and Class 3 in TCGA-LIHC samples.

**Supplementary Table 5 |** 28 core genes filtered according to our research.

## REFERENCES


Amelio, I., Cutruzzolá, F., Antonov, A., Agostini, M., and Melino, G. (2014). Serine and glycine metabolism in cancer. *Trends Biochem. Sci.* 39, 191–198. doi: 10.1016/j.tibs.2014.02.004

Assoun, S., Brosseau, S., Steinmetz, C., Gounant, V., and Zalcman, G. (2017). Bevacizumab in advanced lung cancer: state of the art. *Future Oncol.* 13, 2515–2535. doi: 10.2217/fon-2017-0302

Attila, F., and Mattias, H. (2008). Non-negative matrix factorization for the analysis of complex gene expression data: identification of clinically relevant tumor subtypes. *Cancer Inf.* 6, 275–292.

Baulies, A., Montero, J., Insausti, N., Terrones, O., Vallejo, C., Martinez, L., et al. (2018). The 2-oxoglutarate carrier promotes liver cancer by sustaining mitochondrial GSH despite cholesterol loading. *Redox Biol.* 14, 164–177. doi: 10.1016/j.redox.2017.08.022

Bidkhori, G., Benfeitas, R., Klevstig, M., Zhang, C., Nielsen, J., and Uhlen, M. (2018). Metabolic network-based stratification of hepatocellular carcinoma reveals three distinct tumor subtypes. *Proc. Natl. Acad. Sci. U.S.A.* 115, E11874–E11883.

Blevins, M. A., Huang, M., and Zhao, R. (2017). The role of CtBP1 in oncogenic processes and its potential as a therapeutic target. *Mol. Cancer Ther.* 16:981. doi: 10.1158/1535-7163.mct-16-0592

Cardone, M., Kandilci, A., Carella, C., Nilsson, J. A., Brennan, J. A., and Sirma, S. (2005). The Novel ETS factor TEL2 cooperates with Myc in B lymphomagenesis. *Mol. Cell. Biol.* 25, 2395–2405. doi: 10.1128/mcb.25.6.2395-2405.2005

Carella, C., Potter, M., Bonten, J., Rehg, J. E., Neale, G., and Grosveld, G. C. (2006). The ETS factor TEL2 is a Hematopoietic oncoprotein. *Blood* 107, 1124–1132. doi: 10.1182/blood-2005-03-1196

Che, L., Chi, W., Qiao, Y., Zhang, J., Song, X., Liu, Y., et al. (2019). Cholesterol biosynthesis supports the growth of hepatocarcinoma lesions depleted of fatty acid synthase in mice and humans. *Gut* 69, 177–186. doi: 10.1136/gutjnl-2018-317581

Chen, J., Qian, Z., Li, F., Li, J., and Lu, Y. (2017). Integrative analysis of microarray data to reveal regulation patterns in the pathogenesis of hepatocellular carcinoma. *Gut Liver* 11, 112–120. doi: 10.5009/gnl16063

Chen, L., Luo, L., Chen, W., Xu, H. X., Chen, F., Chen, L. Z., et al. (2016). MicroRNA-24 increases hepatocellular carcinoma cell metastasis and invasion by targeting p53: miR-24 targeted p53. *Biomed. Pharmacother.* 84, 1113–1118. doi: 10.1016/j.biopha.2016.10.051

Cheng, A. L., Kang, Y. K., Chen, Z., Tsao, C. J., Qin, S., Kim, J. S., et al. (2009). Efficacy and safety of sorafenib in patients in the Asia-Pacific region with advanced hepatocellular carcinoma: a phase III randomised, double-blind, placebo-controlled trial. *Lancet Oncol.* 10, 25–34.

Dai, Y., Sui, W., Lan, H., Yan, Q., Huang, H., and Huang, Y. (2009). Comprehensive analysis of microRNA expression patterns in renal biopsies of lupus nephritis patients. *Rheumatol. Intern.* 29, 749–754. doi: 10.1007/s00296-008-0758-6

Daniel, M., David, L., Gerald, Q., Manolis, K., Zoltán, K., and Sven, B. (2016). Tissue-specific regulatory circuits reveal variable modular perturbations across complex diseases. *Nat. Methods* 13, 366–370. doi: 10.1038/nmeth.3799

Federica, A., Laura, P., Daniel, M., Resnick, M. A., and Yari, C. (2018). ETV7-mediated DNAJC15 repression leads to doxorubicin resistance in breast cancer cells. *Neoplasia* 20:857. doi: 10.1016/j.neo.2018.06.008

Fengting, H., Jian, T., Xiaohong, Z., Yanyan, Z., Wenjie, C., Wenbo, C., et al. (2014). MiR-196a promotes pancreatic cancer progression by targeting nuclear factor kappa-B-inhibitor alpha. *PLoS One* 9:e87897. doi: 10.1371/journal.pone.0087897

Ferlay, J., Soerjomataram, I., Dikshit, R., Eser, S., Mathers, C., and Rebelo, M. (2015). Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. *Int. J. Cancer* 136, E359–E386.

Folger, O., Jerby, L., Frezza, C., Gottlieb, E., Ruppin, E., and Shlomi, T. (2011). Predicting selective drug targets in cancer through metabolic networks. *Mol. Syst. Biol.* 7:517. doi: 10.1038/msb.2011.51

Furuta, M., Kozaki, K. I., Tanaka, S., Arii, S., Imoto, I., and Inazawa, J. (2010). miR-124 and miR-203 are epigenetically silenced tumor-suppressive microRNAs in hepatocellular carcinoma. *Carcinogenesis* 31, 766–776. doi: 10.1093/carcin/bgp250

Gao, Q., Zhu, H., Dong, L., Shi, W., Chen, R., and Song, Z. (2019). Integrated proteogenomic characterization of HBV-related hepatocellular carcinoma. *Cell* 179, 561–577.e22.

Ge, J., Yu, Y., Xin, F., Yang, Z. J., Zhao, H. M., Wang, X., et al. (2017). Downregulation of delta-aminolevulinate dehydratase is associated with poor prognosis inpatients with breast cancer. *Cancer Ence* 108, 604–611. doi: 10.1111/cas.13180

Harwood, F. C., Monica, C., Laura, J., David, F., Igor, E., Leena, P., et al. (2018). ETV7 is an essential component of a rapamycin-insensitive mTOR complex in cancer. *Sci. Adv.* 4:eaar3938. doi: 10.1126/sciadv.aar3938

Heirendt, L., Arreckx, S., Pfau, T., Mendoza, S. N., Richelle, A., and Heinken, A. (2019). Creation and analysis of biochemical constraint-based models using the COBRA Toolbox v.3.0. *Nat. Protoc.* 14, 639–702.

Heise, M., Lautem, A., Knapstein, J., Hoppe-Lotichius, M., Foltys, D., Weiler, N., et al. (2012). Downregulation of organic cation transporters OCT1 (SLC22A1) and OCT3 (SLC22A3) in human hepatocellular carcinoma and their prognostic significance. *BMC Cancer* 12:109. doi: 10.1186/1471-2407-12-109

Hennessy, C., Henry, J. A., May, F. E., Westely, B. R., Angus, B., and Lennard, T. W. (1991). Expression of the antimetastatic gene nm23 in human breast cancer: an association WIth good prognosis. *JNCI J. Nat. Cancer Instit.* 83, 281–285. doi: 10.1093/jnci/83.4.281

Hildebrand, J. D., and Soriano, P. (2002). Overlapping and unique roles for C-terminal binding protein 1 (CtBP1) and CtBP2 during mouse development. *Mol. Cell. Biol.* 22, 5296–5307. doi: 10.1128/mcb.22.15.5296-5307.2002

Howie, L. J., Scher, N. S., Amiri-Kordestani, L., Zhang, L., and Beaver, J. A. (2018). FDA approval summary: pertuzumab for adjuvant treatment of HER2-positive early breast cancer. *Clin. Cancer Res.* 25:Clincanres.3003.2018.

Jiang, Y., Sun, A., Zhao, Y., Ying, W., Sun, H., Yang, X., et al. (2019). Proteomics identifies new therapeutic targets of early-stage hepatocellular carcinoma. *Nature* 567, 257–261.

Keshet, R., and Erez, A. (2018). Arginine and the metabolic regulation of nitric oxide synthesis in cancer. *Dis. Models Mechan.* 11:dmm033332. doi: 10.1242/dmm.033332

KöBerle, V., Kronenberger, B., Pleli, T., Trojan, J. R., Imelmann, E., Welker, M. W., et al. (2013). Serum microRNA-1 and microRNA-122 are prognostic markers in patients with hepatocellular carcinoma. *Eur. J. Cancer* 49, 3442–3449. doi: 10.1016/j.ejca.2013.06.002

Krakowiak, P. A., Wassif, C. A., Kratz, L., Cozma, D., Kovářová, M., Harris, G., et al. (2003). Lathosterolosis: an inborn error of human and murine cholesterol synthesis due to lathosterol 5-desaturase deficiency. *Hum. Mol. Genet.* 12, 1631–1641. doi: 10.1093/hmg/ddg172

Krämer, A., Green, J., Pollard, J. Jr., and Tugendreich, S. (2013). Causal analysis approaches in ingenuity pathway analysis. *Bioinformatics* 30, 523–530. doi: 10.1093/bioinformatics/btt703

Krishna, T. R., Kumar, Y. V., Akinchan, K., Ankita, S., Krishnendu, P., and Luke, H. (2014). Non-metastatic 2 (NME2)-mediated suppression of lung cancer metastasis involves transcriptional regulation of key cell adhesion factor vinculin. *Nuclc Acids Res.* 42, 11589–11600. doi: 10.1093/nar/gku860

Lang, Q., and Ling, C. (2012). MiR-124 suppresses cell proliferation in hepatocellular carcinoma by targeting PIK3CA. *Biochem. Biophys. Res. Commun.* 426, 247–252. doi: 10.1016/j.bbrc.2012.08.075

Laos, S., Baeckstrm, D., and Hansson, G. C. (2006). Inhibition of NF-kappaB activation and chemokine expression by the leukocyte glycoprotein, CD43, in colon cancer cells. *Int. J. Oncol.* 28, 695–704.

Lee, J. S., Lee, H., Lee, S., Kang, J. H., Lee, S. H., Kim, S. G., et al. (2019). Loss of SLC25A11 causes suppression of NSCLC and melanoma tumor formation. *EBiomedicine* 40, 184–197. doi: 10.1016/j.ebiom.2019.01.036

Lehmann, W., Mossmann, D., Kleemann, J., Mock, K., Meisinger, C., Brummer, T., et al. (2016). ZEB1 turns into a transcriptional activator by interacting with YAP1 in aggressive cancer types. *Nat. Commun.* 7:10498.

Li, W., Wang, Q., Feng, Q., Wang, F., Yan, Q., Gao, S. J., et al. (2019). Oncogenic KSHV-encoded interferon regulatory factor upregulates HMGB2 and CMPK1 expression to promote cell invasion by disrupting a complex lncRNA-OIP5-AS1/miR-218-5p network. *PLoS Pathog.* 15:e1007578. doi: 10.1371/journal.ppat.1007578

Lifeng, X., Jade, T., Michael, B., Regina, L., Susanna, L., and Patrick, W. (2016). Arginine metabolism in bacterial pathogenesis and cancer therapy. *Intern. J. Mol. Sci.* 17:363. doi: 10.3390/ijms17030363

Liu, G., Hou, G., Li, L., Li, Y., and Liu, L. (2016). Potential diagnostic and prognostic marker dimethylglycine dehydrogenase (DMGDH) suppresses hepatocellular carcinoma metastasis in vitro and in vivo. *Oncotarget* 7, 32607–32616. doi: 10.18632/oncotarget.8927

Liu, N. Q., De Marchi, T., Timmermans, A., Trapman-Jansen, A. M. A. C., Foekens, R., Look, M. P., et al. (2016). Umar, prognostic significance of nuclear expression of UMP-CMP kinase in triple negative breast cancer patients. *Sci. Rep.* 6:32027.

Marbach, D., Schaffter, T., Mattiussi, C., and Floreano, D. (2009). Generating realistic in silico gene networks for performance assessment of reverse engineering methods. *J. Comput. Biol.* 16, 229–239. doi: 10.1089/cmb.2008.09tt

Mardinoglu, A., Agren, R., Kampf, C., Asplund, A., and Nielsen, J. (2014). Genome-scale metabolic modelling of hepatocytes reveals serine deficiency in patients with non-alcoholic fatty liver disease. *Nat. Commun.* 5:3083.

Mardinoglu, A., and Nielsen, J. (2012). Systems medicine and metabolic modelling. *J. Intern. Med.* 271, 142–154. doi: 10.1111/j.1365-2796.2011.02493.x

Mardinoglu, A., and Nielsen, J. (2015). New paradigms for metabolic modeling of human cells. *Curr. Opin. Biotechnol.* 34, 91–97. doi: 10.1016/j.copbio.2014.12.013

Matos, J. M., Witzmann, F. A., Cummings, O. W., and Schmidt, C. M. (2009). A pilot study of proteomic profiles of human hepatocellular carcinoma in the United States. *J. Surg. Res.* 155, 237–243. doi: 10.1016/j.jss.2008.06.008

Matthew, R. E., Belinda, P., Di, W., Yifang, H., Charity, W. L., Wei, S., et al. (2015). Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* 7, e47. doi: 10.1093/nar/gkv007

Mehrmohamadi, M., Liu, X., Shestov, A. A., and Locasale, J. W. (2014). Characterization of the usage of the serine metabolic network in human cancer. *Cell Rep.* 9, 1507–1519. doi: 10.1016/j.celrep.2014.10.026

Meihua, L., Duo, L., Yunliang, Z., Minglan, W., Chang, X., Qiao, Z., et al. (2018). Downregulation of CPT2 promotes tumorigenesis and chemoresistance to cisplatin in hepatocellular carcinoma. *Oncotarg. Therapy* 11, 3101–3110. doi: 10.2147/ott.s163266

Murakami, Y., Yasuda, T., Saigo, K., Urashima, T., Toyoda, H., and Okanoue, T. (2006). Comprehensive analysis of microRNA expression patterns in hepatocellular carcinoma and non-tumorous tissues. *Oncogene* 25, 2537–2545. doi: 10.1038/sj.onc.1209283

Nicolas, M., Wolfer, A., Raj, K., Kummer, J. A., Clevers, H., Dotto, G. P., et al. (2003). Notch1 functions as a tumor suppressor in mouse skin. *Nat. Genet.* 33, 416–421. doi: 10.1038/ng1099

Ohmine, K., Kawaguchi, K., Ohtsuki, S., Motoi, F., Ohtsuka, H., Kamiie, J., et al. (2015). Quantitative targeted proteomics of pancreatic cancer: deoxycytidine kinase protein level correlates to progression-free survival of patients receiving gemcitabine treatment. *Mol. Pharm.* 12, 3282–3291. doi: 10.1021/acs.molpharmaceut.5b00282

Patrycja, A. K., Christopher, A. W., Lisa, K., Diana, C., Martina, K., Ginny, H., et al. (2003). Lathosterolosis: an inborn error of human and murine cholesterol synthesis due to lathosterol 5-desaturase deficiency. *Hum. Mol. Genet.* 12, 1631–1641. doi: 10.1093/hmg/ddg172

Peeters, P., Raynaud, S. D., Cools, J., Wlodarska, I., and Marynen, P. (1997). Fusion of TEL, the ETS-variant gene 6 (ETV6), to the receptor-associated kinase JAK2 as a result of t(9; 12) in a Lymphoid and t(9; 15; 12) in a myeloid leukemia. *Blood* 90, 2535–2540. doi: 10.1182/blood.v90.7.2535

Phang, J. M., Liu, W., Hancock, C. N., and Fischer, J. W. (2015). Proline metabolism and cancer: emerging links to glutamine and collagen. *Curr. Opin. Clin. Nutr. Metab. Care* 18, 71–77. doi: 10.1097/mco.0000000000000121

Rawluk, J., and Waller, C. F. (2018). Gefitinib. *Recent Results Cancer Res.* 211, 235–246.

Rivory, L. P., Slaviero, K. A., and Clarke, S. J. (2002). Hepatic cytochrome P450 3A drug metabolism is reduced in cancer patients who have an acute-phase response. *Br. J. Cancer* 87:277. doi: 10.1038/sj.bjc.6600448

Roy, S., Lagree, S., Hou, Z., Thomson, J. A., Stewart, R., and Gasch, A. P. (2013). Integrated module and gene-specific regulatory inference implicates upstream signaling networks. *PLoS Comput. Biol.* 9:e1003252. doi: 10.1371/journal.pcbi.1003252

Ryu, J. S., Shin, E. S., Nam, H. S., Yi, H. G., Cho, J. H., Kim, C. S., et al. (2011). Differential effect of polymorphisms of CMPK1 and RRM1 on survival in advanced non-small cell lung cancer patients treated with gemcitabine or taxane/cisplatinum. *J. Thorac. Oncol.* 6, 1320–1329. doi: 10.1097/jto.0b013e3182208e26

Salam, S. A., Arroyo, A. B., Raúl, T. M., Nuria, G. B., Vanessa, R., Vicente, V., et al. (2016). MiRNA-based regulation of hemostatic factors through hepatic nuclear factor-4 alpha. *PLoS One* 11:e0154751. doi: 10.1371/journal.ppat.0154751 doi: 10.1371/journal.pone.0154751

Sartorius, K., Makarova, J., Sartorius, B., An, P., Winkler, C., Chuturgoon, A., et al. (2019). The regulatory role of MicroRNA in hepatitis-B virus-associated hepatocellular carcinoma (HBV-HCC) pathogenesis. *Cells* 8:1504. doi: 10.3390/cells8121504

Song, B. L., Javitt, N. B., and Debose-Boyd, R. A. (2005). Insig-mediated degradation of HMG CoA reductase stimulated by lanosterol, an intermediate in the synthesis of cholesterol. *Cell Metab.* 1, 179–189. doi: 10.1016/j.cmet.2005.01.001

Steeg, P. S., Bevilacqua, G., Kopper, L., Thorgeirsson, U. P., and Sobel, M. E. (1988). Evidence for a novel gene associated with low tumor metastatic potential. *J. Natl. Cancer Inst.* 80, 200–204. doi: 10.1093/jnci/80.3.200

Su, G., Morris, J. H., Demchak, B., and Bader, G. D. (2014). Biological network exploration with cytoscape 3. *Curr. Protoc. Bioinform.* 47, 8.13.1–8.13.24.

Uhle, M., Zhan, C., Le, S., Sjösted, E., Fagerber, L., and Bidkhor, G. (2017). A pathology atlas of the human cancer transcriptome. *Science* 357:eaan2507.

Uhlén, M., Fagerberg, L., Hallström, B. M., Lindskog, C., Oksvold, P., and Mardinoglu, A. (2015). Tissue-based map of the human proteome. *Science* 347:1260419.

Vecchi, C., Montosi, G., Garuti, C., Corradini, E., Sabelli, M., Canali, S., et al. (2013). Gluconeogenic signals regulate iron homeostasis via Hepcidin in mice. *Gastroenterology* 146, 1060–1069. doi: 10.1053/j.gastro.2013.12.016

Villar, V. H., Merhi, F., Djavaheri-Mergny, M., and Durán, R. V. (2015). Glutaminolysis and autophagy in cancer. *Autophagy* 11, 1198–1208. doi: 10.1080/15548627.2015.1053680

Vlierberghe, P. V., Weer, A. D., Mestdagh, P., Feys, T., and Speleman, F. (2010). Comparison of miRNA profiles of microdissected Hodgkin/Reed-sternberg cells and Hodgkin cell lines versus CD77+ B-cells reveals a distinct subset of differentially expressed miRNAs. *Br. J. Haematol.* 147, 686–690. doi: 10.1111/j.1365-2141.2009.07909.x

Voigtlaender, M., Schneider-Merck, T., and Trepel, M. (2018). Lapatinib. *Recent Results Cancer Res.* 211, 19–44.

Wang, G., Dong, F., Xu, Z., Sharma, S., Hu, X., Chen, D., et al. (2017). MicroRNA profile in HBV-induced infection and hepatocellular carcinoma. *BMC Cancer* 17:805. doi: 10.1186/s12885-017-3816-1

Wang, Z., Danziger, S. A., Heavner, B. D., Ma, S., Smith, J. J., and Li, S. (2017). Combining inferred regulatory and reconstructed metabolic networks enhances phenotype prediction in yeast. *PLoS Comput. Biol.* 13:e1005489. doi: 10.1371/journal.pcbi.1005489

Wang, Y., Sun, B., Zhao, X., Zhao, N., Sun, R., Zhu, D., et al. (2016). Twist1-related miR-26b-5p suppresses epithelial-mesenchymal transition, migration

and invasion by targeting SMAD1 in hepatocellular carcinoma. *Oncotarget* 7, 24383–24401. doi: 10.18632/oncotarget.8328

Wu, X., Deng, F., Li, Y., Daniels, G., Du, X., Ren, Q., et al. (2015). ACSL4 promotes prostate cancer growth, invasion and hormonal resistance. *Oncotarget*. 6, 44849–44863. doi: 10.18632/oncotarget.6438

Xiao, J., Li, W., Zheng, X., Qi, L., Wang, H., Zhang, C., et al. (2020). Targeting 7-dehydrocholesterol reductase integrates cholesterol metabolism and IRF3 activation to eliminate infection. *Immunity* 52, 109–122.e6.

Xu, Y. Z., Hang, Z., Qiang-Zhen, Y., Ren-Liang, S., Ke, W., Yong-Yong, S., et al. (2020). Integrated regulatory-metabolic network model reveals critical mechanism and potential targets for Hepatocellular Carcinoma. *ResearchSquare* [Preprint], doi: 10.21203/rs.3.rs-21615/v1

Yip, K. W., Zhang, Z., Huang, J. W., Vu, N. M., Chiang, Y. K., Lin, C. L., et al. (2014). A porphodimethene chemical inhibitor of uroporphyrinogen decarboxylase. *PLoS One* 9:e89889. doi: 10.1371/journal.ppat.1089889

Zhang, B., Wu, Q., Wang, Z., Xu, R., Hu, X., Sun, Y., et al. (2019). The promising novel biomarkers and candidate small molecule drugs in kidney renal clear cell carcinoma: Evidence from bioinformatics analysis of high-throughput data. *Mol. Genet. Genom. Med.* 7:e607. doi: 10.1002/mgg3.607

Zhang, H., Zhang, Z., Gao, L., Qiao, Z., and Yang, T. (2019). miR-1-3p suppresses proliferation of hepatocellular carcinoma through targeting SOX9. *Oncotargets Therapy* 12, 2149–2157. doi: 10.2147/ott.s197326

Zheng, F., Liao, Y. J., Cai, M. Y., Liu, Y. H., Liu, T. H., Chen, S. P., et al. (2012). The putative tumour suppressor microRNA-124 modulates hepatocellular carcinoma cell aggressiveness by repressing ROCK2 and EZH2. *Gut* 61, 278–289. doi: 10.1136/gut.2011.239145

Zheng, G., Xu, Y., Zhang, X., Liu, Z. P., Wang, Z., Chen, L., et al. (2016). CMIP: a software package capable of reconstructing genome-wide regulatory networks using gene expression data. *BMC Bioinf.* 17:535. doi: 10.1186/s12859-016-1324-y

Zhi-Ping, L., Canglin, W., Hongyu, M., and Hulin, W. (2015). RegNetwork: an integrated database of transcriptional and post-transcriptional regulatory networks in human and mouse. *Database J. Biol. Databases Curation* 2015:bav095. doi: 10.1093/database/bav095

Zhou, D., Zhang, L., Lin, Q., Ren, W., and Xu, G. (2017a). Data on the association of CMPK1 with clinicopathological features and biological effect in human epithelial ovarian cancer. *Data Brief* 13, 77–84. doi: 10.1016/j.dib.2017.05.022

Zhou, D., Zhang, L., Sun, W., Guan, W., Lin, Q., Ren, W., et al. (2017b). Cytidine monophosphate kinase is inhibited by the TGF-β signalling pathway through the upregulation of miR-130b-3p in human epithelial ovarian cancer. *Cell. Signal.* 35:197. doi: 10.1016/j.cellsig.2017.04.009

Zhu, H., Rao, S. P. R., Zeng, T., and Chen, L. (2012). Reconstructing dynamic gene regulatory networks from sample-based transcriptional data. *Nucleic Acids Res.* 40, 10657–10667. doi: 10.1093/nar/gks860

**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 Neuroscience

ORIGINAL RESEARCH
published: 11 January 2021
doi: 10.3389/fnins.2020.580929

# Changes of Metabolites in Acute Ischemic Stroke and Its Subtypes

*Xin Wang[1,2], Luyang Zhang[1,2], Wenxian Sun[1,2], Lu-lu Pei[1,2], Mengke Tian[1,2], Jing Liang[1,2], Xinjing Liu[1,2], Rui Zhang[1,2], Hui Fang[1,2], Jun Wu[1,2], Shilei Sun[1,2], Yuming Xu[1,2]*, Jian-Sheng Kang[1]* and Bo Song[1,2]**

[1] Department of Neurology, The First Affiliated Hospital of Zhengzhou University, Zhengzhou, China, [2] Henan Key Laboratory of Cerebrovascular Diseases, Zhengzhou, China

Existing techniques have many limitations in the diagnosis and classification of ischemic stroke (IS). Considering this, we used metabolomics to screen for potential biomarkers of IS and its subtypes and to explore the underlying related pathophysiological mechanisms. Serum samples from 99 patients with acute ischemic stroke (AIS) [the AIS subtypes included 49 patients with large artery atherosclerosis (LAA) and 50 patients with small artery occlusion (SAO)] and 50 matched healthy controls (HCs) were analyzed by non-targeted metabolomics based on liquid chromatography–mass spectrometry. A multivariate statistical analysis was performed to identify potential biomarkers. There were 18 significantly different metabolites, such as oleic acid, linoleic acid, arachidonic acid, L-glutamine, L-arginine, and L-proline, between patients with AIS and HCs. These different metabolites are closely related to many metabolic pathways, such as fatty acid metabolism and amino acid metabolism. There were also differences in metabolic profiling between the LAA and SAO groups. There were eight different metabolites, including L-pipecolic acid, 1-Methylhistidine, PE, LysoPE, and LysoPC, which affected glycerophospholipid metabolism, glycosylphosphatidylinositol-anchor biosynthesis, histidine metabolism, and lysine degradation. Our study effectively identified the metabolic profiles of IS and its subtypes. The different metabolites between LAA and SAO may be potential biomarkers in the context of clinical diagnosis. These results highlight the potential of metabolomics to reveal new pathways for IS subtypes and provide a new avenue to explore the pathophysiological mechanisms underlying IS and its subtypes.

**Keywords: ischemic stroke, metabolites, non-targeted metabolomics, TOAST, biomarkers**

OPEN ACCESS

***Edited by:***
*Bhabatosh Das, Translational Health Science and Technology Institute (THSTI), India*

***Reviewed by:***
*Abhinav Achreja, University of Michigan, United States*
*Rashmi Kumari, Pennsylvania State University, United States*

***Correspondence:**
*Yuming Xu*
*xuyuming@zzu.edu.cn*
*Jian-Sheng Kang*
*kjs@zzu.edu.cn*
*Bo Song*
*fccsongb@zzu.edu.cn*

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

***Received:*** *07 July 2020*
***Accepted:*** *26 November 2020*
***Published:*** *11 January 2021*

***Citation:***
*Wang X, Zhang L, Sun W, Pei L-l, Tian M, Liang J, Liu X, Zhang R, Fang H, Wu J, Sun S, Xu Y, Kang J-S and Song B (2021) Changes of Metabolites in Acute Ischemic Stroke and Its Subtypes. Front. Neurosci. 14:580929. doi: 10.3389/fnins.2020.580929*

## INTRODUCTION

Stroke is one of the main causes of human death and disability (Wang et al., 2014) and is associated with a high rate of disability and recurrence. According to population-based studies (Benjamin et al., 2017), ischemic stroke (IS) accounts for more than 80% of all strokes. According to its etiology and imaging, IS can be categorized into five subtypes (Adams et al., 1993; Chen et al., 2012), including large artery atherosclerosis (LAA), small artery occlusion (SAO), cardioembolism, stroke of other determined cause, and stroke of undetermined cause. The classification of IS can help with the early treatment and prevention of long-term recurrence in patients (Montaner et al., 2008). However, the diagnosis and classification of IS mainly rely on neuroimaging techniques,

which are scarce, expensive, and time-consuming (Latchaw et al., 2009). Therefore, new biomarkers for the rapid and accurate prediction, diagnosis, and classification of IS might play a positive role in clarifying the pathophysiological mechanism of IS and promoting the secondary prevention and management of patients with IS.

It is difficult to release macromolecules from the brain into the blood due to the presence of the blood–brain barrier (Jickling and Sharp, 2015). Some conventional detection methods make it difficult to detect specific sensitive different metabolites in patients with IS. However, with the development of the emerging science of metabolomics, it may be possible to identify specific small molecular biomarkers in patients with IS and determine the underlying etiology. Metabolomics is an effective method to reveal biomolecules' phenotypes. This method enables the identification of changes in small molecular metabolites in various diseases, which can greatly help in understanding and diagnosing diseases. Many studies using metabolomics have revealed the differences in metabolites between patients with acute ischemic stroke (AIS) and healthy controls (HCs) (Jung et al., 2011; Kimberly et al., 2013). To date, few studies have explored the differences in metabolites between the LAA and SAO subtypes of IS. In this study, non-targeted metabolites based on liquid chromatography–mass spectrometry (LC–MS) were used to study the different metabolites between patients with AIS and the HCs and between patients with the LAA and SAO subtypes of IS. The proposed method offers important advantages over traditional alternatives, ensuring that it is feasible to screen potential biomarkers and further explore the relevant underlying pathophysiological mechanisms.

# MATERIALS AND METHODS

## Study Population

In this study, 99 patients with AIS within 7 days of onset were included in the AIS group, including 49 patients with LAA and 50 patients with SAO. Among the 99 AIS patients, the time from onset to blood withdrawal was within 24 h among 45 patients and within 72 h among 38 patients. The remaining 16 patients showed transient ischemic attack symptoms within 7 days, but the blood samples were only collected at around 72 h after the symptoms begin to persist. A total of 50 HCs with age, sex, and risk factors matched with the AIS group were recruited. All patients needed to meet the following conditions: (1) no history of stroke or coronary heart disease, (2) no history of malignant tumor or autoimmune disease, (3) blood samples can be obtained within 24 h of enrollment, and (4) head magnetic resonance imaging and angiography were completed during hospitalization. The patients in the AIS group, who were hospitalized in the Department of Neurology from October 2015 to December 2016, and the samples of the AIS group were acquired from the ischemic cerebrovascular disease database and blood database of The First Affiliated Hospital of Zhengzhou University. The details of the database and related articles have been published elsewhere (Song et al., 2013; Kelly et al., 2016; Wang et al., 2020). All patients with AIS were diagnosed according to the diagnostic criteria of the World Health Organization [WHO] (1989). The TOAST classification was evaluated back to back by two professional neurologists. Written informed consent was obtained from all participants or their representatives.

## Serum Sample Preparation

Blood samples of patients with AIS were collected within 24 h after admission; when collecting, it was ensured that the patients have fasted for at least 8 h. The serum was centrifuged and extracted within 1 h and refrigerated at −80°C. To separate metabolites with different polarities, the same sample underwent two different treatment methods. After melting in ice at 4°C for 30–60 min, 40 μl of serum was taken into a 1.5-ml centrifuge tube for a reversed-phase ultra-performance liquid chromatographic analysis, adding 300 μl methanol and 1 ml methyl tert-butyl ether to precipitate the protein for 15 s. The sample was then placed in a centrifuge at 12,000 rpm at a constant temperature of 4°C for 10 min, the upper solution (400 μl) was then evaporated, and the sample was finally redissolved in 100 μl methanol. For the serum analyzed by Hydrop interaction liquid chromatography (HILIC), 50 μl plus 150 μl acetonitrile was added to the centrifuge tube to precipitate the protein, and 100 μl of the upper solution was centrifuged under the above-mentioned conditions to be determined.

## Chromatographic Condition

For the C18 separation, mobile phase A consisted of acetonitrile/water (60/40), and mobile phase B was isopropanol/acetonitrile (90/10); both A and B contained 0.1% formic acid and 10 mmol/L ammonium acetate. The column was an HSS T3 column (2.1 × 100 mm, 1.8 μm, Waters) operated at 45°C. The flow rate was 300 μl/min, and the injection volume was 1 μl. For the HILIC separation, mobile phase A was acetonitrile, and mobile phase B was water; both A and B contained 0.1% formic acid and 10 mmol/L ammonium acetate. The column was a BEH amide column (2.1 × 100 mm, 1.7 μm, Waters) operated at 40°C. The flow rate was 300 μl/min, and the injection volume was 1 μ l.

## LC–MS Detection

A metabolomics analysis was performed using a Thermo Scientific Q Exactive hybrid quadrupole Orbitrap mass spectrometer equipped with a HESI-II probe. The positive and negative HESI-II spray voltages were 3.7 and 3.5 kV, respectively, the heated capillary temperature was 320°C, the sheath gas pressure was 30 psi, the auxiliary gas setting was 10 psi, and the heated vaporizer temperature was 300°C. Both the sheath gas and the auxiliary gas consisted of nitrogen. The collision gas was also nitrogen at a pressure of 1.5 mTorr. The parameters of the full mass scan were as follows: resolution of 70,000, auto gain control target under 1 × 106, maximum isolation time of 50 ms, and m/z range of 50–1,500. The LC–MS system was controlled using Xcalibur 2.2 SP1.48 software (Thermo Fisher Scientific), and data were collected and processed using the same software.

## Untargeted Metabolome Data Processing

All data obtained from the four assays in the two systems in both positive and negative ion modes were processed using Progenesis QI data analysis software (Non-linear Dynamics, Newcastle, United Kingdom) to impute raw data, peak alignment, picking, and normalization to produce peak intensities for retention time ($t_R$) and m/z data pairs. The ranges of automatic peak picking for the C18 were between 1 and 16 min and between 1 and 12 min, respectively. Next, the adduct ions of each "feature" (m/z, $t_R$) were deconvoluted, and these features were identified in the human metabolome database (HMDB) and Lipidmaps.

To monitor a system's stability and performance and the reproducibility of the sample, quality control (QC) samples were prepared by pooling equal volumes of each serum sample. The pretreatment of serum QC samples was performed in accordance with real samples. For repeatable metabolic analyses, three features of the analytical system must be stable: (1) retention time, (2) signal intensity, and (3) mass accuracy. In this study, three QCs were continuously injected at the beginning of the run. QC samples are then injected at regular intervals of six or eight samples throughout the analytical run to provide data from which repeatability can be assessed.

The features were selected based on their coefficients of variation (CVs) with QC samples; features with CVs over 15% were eliminated.

## Statistical Analysis

The classified variables and continuous variables in the baseline information on participants were compared by $\chi^2$ test and *t*-test in SPSS, respectively. Data are presented as mean ± SD or the percentage, as appropriate. A multivariate statistical analysis was performed using principal component analysis (PCA) and orthogonal projections to latent structures—discriminant analysis (OPLS-DA) multivariate statistical methods in SIMCA (14.1) software. In this study, the variable importance in the projection (VIP) value of the OPLS-DA model (threshold > 1) and the *P*-value of *t*-test ($P < 0.05$) were used to identify the different metabolites. The qualitative method of different metabolites consists of searching in HMDB (to compare the m/z or molecular mass, error limit 0.01 Da). The OPLS-DA model was then validated by permutation tests. A pathway analysis was performed using MetaboAnalyst 4.0.

# RESULTS

## Baseline Characteristics

In this study, there were 99 people in the AIS group and 50 people in the control group. The baseline characteristics are shown in **Table 1**. There were 73 males and 26 females with an average age of 58.06 years in the AIS group and 36 men and 14 women with an average age of 57.60 years in the control group. We found no significant difference in age, sex, hyperlipidemia, and diabetes mellitus between the AIS and control groups ($P > 0.05$).

## PCA and OPLS-DA

The PCA of the unsupervised model was used to analyze the differences and intra-group variation among the LAA, SAO, and HC groups, in which $R^2X$ was used to judge the quality of the model, and $Q^2$ represented the predictable variables of the model. As shown in **Figures 1A–C**, there was a slight separation among the three groups on the score plots in both the C18 column and the HILIC column (C18-positive: $R^2X = 0.789$, $Q^2 = 0.547$; C18-negative: $R^2X = 0.817$, $Q^2 = 0.616$; HILIC: $R^2X = 0.732$, $Q^2 = 0.453$). To obtain the metabolite information that leads to this difference, supervised models and OPLS-DA were performed. The serum samples in the AIS group and the HC group were separated in the C18 column and HILIC column (C18-positive: $R^2Y = 0.883$, $Q^2 = 0.726$; C18-negative: $R^2Y = 0.964$, $Q^2 = 0.857$; HILIC: $R^2Y = 0.985$, $Q^2 = 0.914$) as were the LAA group and the SAO group (C18-positive: $R^2Y = 0.916$, $Q^2 = 0.778$; C18-negative: $R^2Y = 0.909$, $Q^2 = 0.800$; HILIC: $R^2Y = 0.953$, $Q^2 = 0.726$), highlighting the excellence of the models (**Figures 1D–I**).

## Different Metabolites

The analysis of OPLS-DA in the supervised model is summarized in **Table 2**. A total of 18 significantly changed metabolites (SCMs)

**TABLE 1 |** Comparison of baseline characteristics between the patients and the healthy controls.

| | Patients (*n* = 99) | | | Controls (*n* = 50) | *P*-value |
|---|---|---|---|---|---|
| | LAA (*n* = 49) | SAO (*n* = 50) | Total | | |
| Age (years), mean ± SD | 58.08 ± 12.827 | 58.04 ± 10.234 | 58.06 ± 11.531 | 57.60 ± 2.718 | 0.781 |
| Male, *n* (%) | 36 (73.5) | 37 (74.0) | 73 (73.7) | 36 (72.0) | 0.821 |
| **Medical history, *n* (%)** | | | | | |
| Hypertension | 29 (52.9) | 24 (48.0) | 53 (53.5) | 13 (26.0) | 0.001 |
| Diabetes | 13 (26.5) | 7 (14.0) | 20 (20.2) | 8 (16.0) | 0.535 |
| Hyperlipidemia | 5 (10.2) | 4 (8.0) | 9 (9.1) | 8 (16.0) | 0.210 |
| Coronary heart disease | 3 (6.1) | 2 (4.0) | 5 (5.1) | 1 (2.0) | 0.664 |
| Smoking | 21 (42.9) | 20 (40.0) | 41 (41.4) | 9 (18.0) | 0.004 |
| Drinking | 18 (36.7) | 19 (38.0) | 37 (37.4) | 16 (32.0) | 0.518 |

*LAA, large artery atherosclerosis; SAO, small artery occlusion.*

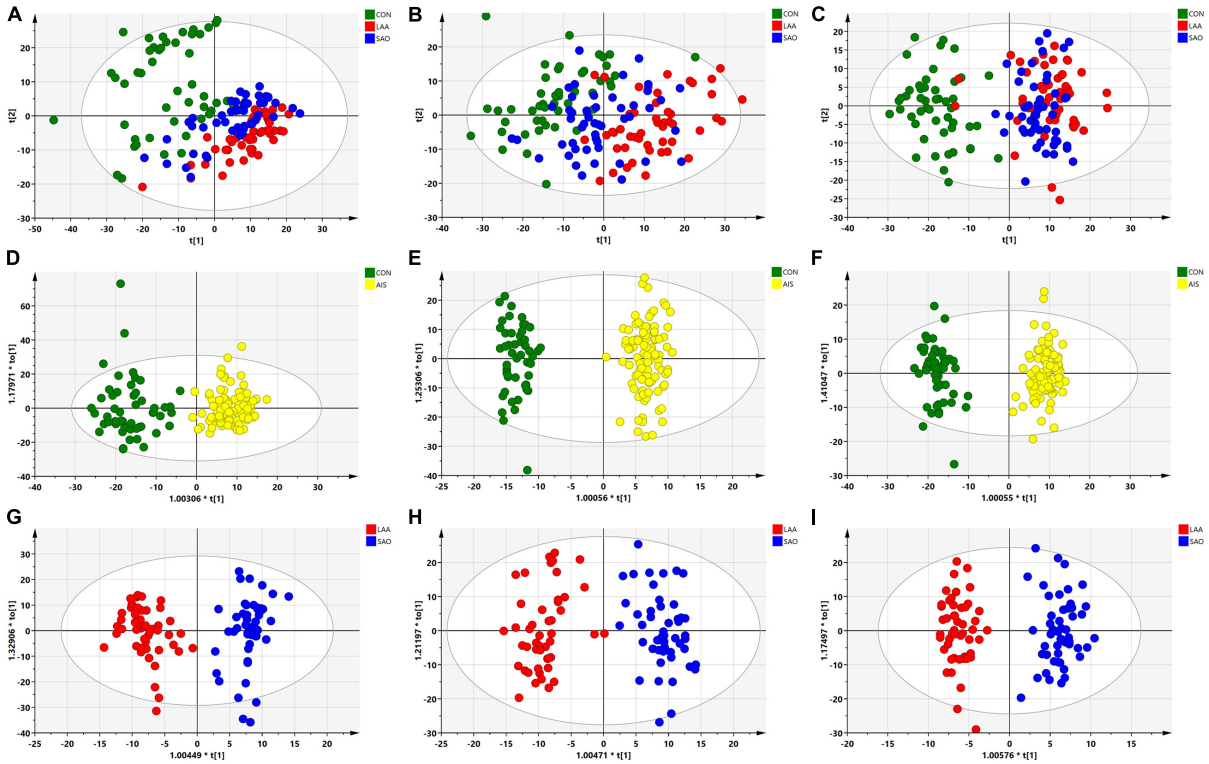



**FIGURE 1 |** Multivariate statistical analysis of serum metabolic profiling between acute ischemic stroke (AIS) and healthy control (HC) groups. **(A)** Principal component analysis (PCA) score plots in the C18-positive column ($R^2X$ = 0.789, $Q^2$ = 0.547). **(B)** PCA score plots in the C18-negative column ($R^2X$ = 0.817, $Q^2$ = 0.616). **(C)** PCA score plots in the Hydrop interaction liquid chromatography (HILIC) column ($R^2X$ = 0.732, $Q^2$ = 0.453). **(D)** Orthogonal projections to latent structures—discriminant analysis (OPLS-DA) score plots of patients with AIS and HCs in the C18-positive column ($R^2Y$ = 0.883, $Q^2$ = 0.726). **(E)** OPLS-DA score plots of patients with AIS and HCs in the C18-negative column ($R^2Y$ = 0.964, $Q^2$ = 0.857). **(F)** OPLS-DA score plots of the patients with AIS and HCs in the HILIC column ($R^2Y$ = 0.985, $Q^2$ = 0.914). **(G)** OPLS-DA score plots of the large artery atherosclerosis (LAA) and the small artery occlusion (SAO) groups in the C18-positive column ($R^2Y$ = 0.916, $Q^2$ = 0.778). **(H)** OPLS-DA score plots of the LAA and the SAO groups in the C18-negative column ($R^2Y$ = 0.909, $Q^2$ = 0.800). **(I)** OPLS-DA score plots of the LAA and the SAO groups in the HILIC column ($R^2Y$ = 0.953, $Q^2$ = 0.726).

(VIP > 1, $P$ < 0.05) were screened between the AIS group and the HC group through different chromatographic columns, and a total of eight SCMs were screened between the LAA and SAO groups from the HMDB.

Compared with the HCs, the AIS patients exhibited higher levels of oleic acid, linoleic acid, arachidonic acid (AA), docosahexaenoic acid (DHA), L-palmitoylcarnitine, tetradecanoylcarnitine, dodecanoylcarnitine, and decanoylcarnitine and lower levels of Cer (14:0), Cer (16:0), nonadecanoic acid, 4-hydroxyproline, phosphatidylethanolamine (PE) (18:1), PE (18:0), propionylcarnitine, L-glutamine, L-arginine, and L-proline as shown in the heat map in **Figure 2**. In comparison to the SAO group, the LAA group was characterized by decreased levels of L-pipecolic acid, 1-methylhistidine, PE (18:2), LysoPE (18:2), LysoPC (18:3), LysoPC (20:0), and LysoPC (18:2) and by increased levels of PE (16:0). Detailed information is shown in **Table 2**.

## Metabolic Pathways

MetaboAnalyst 4.0 was used to analyze the different metabolic pathways of the groups. The potential different metabolic pathways between the AIS patients and the HCs include linoleic acid metabolism, AA metabolism, arginine and proline metabolism, and alanine, aspartate, and glutamate metabolism. The metabolic pathways of the LAA group and the SAO group probably differ in glycerophospholipid metabolism, glycosylphosphatidylinositol (GPI)-anchor biosynthesis, histidine metabolism, and lysine degradation (**Figure 3**).

# DISCUSSION

We obtained the serum metabolic profiling of stroke patients by non-targeted metabolomics and discovered that the AIS patients had metabolic disorders. Furthermore, the metabolic profiles of the LAA and SAO subtypes of IS were different. Among the abnormal metabolic indicators, the metabolic disorders of lipids and amino acids are the most obvious. Changes in metabolite patterns lay the foundation for us to further clarify the pathophysiological mechanisms of stroke and find ways to implement innovative clinical diagnoses of stroke classifications.

The brain consumes approximately 20% of total human energy consumed. Approximately 20% of the energy consumed

**TABLE 2 |** Characteristics of the different metabolites.

| Metabolites | Retention time (min) | Mass-to-charge ratio | VIP value | Fold change | *P*-value |
|---|---|---|---|---|---|
| **Between the AIS and HCs groups** | | | | | |
| Oleic acid | 5.116 | 281.249 | 2.107 | 2.216 | <0.00001 |
| Linoleic acid | 4.465 | 279.233 | 1.844 | 1.970 | 0.00001 |
| Cer (d18:0/14:0) | 8.993 | 512.503 | 1.735 | 0.212 | <0.00001 |
| Cer (d18:0/16:0) | 9.546 | 540.534 | 1.732 | 0.213 | <0.00001 |
| Arachidonic acid | 4.421 | 303.233 | 1.556 | 1.564 | 0.00030 |
| Non-adecanoic acid | 6.306 | 297.280 | 1.371 | 0.652 | <0.00001 |
| Docosahexaenoic acid | 4.207 | 327.233 | 1.306 | 1.342 | 0.00964 |
| 4-Hydroxyproline | 5.939 | 132.066 | 1.270 | 0.520 | <0.00001 |
| PE[18:2(9Z,12Z)/18:1(9Z)] | 8.534 | 742.538 | 1.099 | 0.492 | <0.00001 |
| PE[18:2(9Z,12Z)/18:0] | 8.989 | 744.553 | 1.099 | 0.514 | <0.00001 |
| L-Palmitoylcarnitine | 1.433 | 400.342 | 1.098 | 1.338 | <0.00001 |
| Propionylcarnitine | 2.604 | 218.139 | 1.097 | 0.716 | <0.00001 |
| Tetradecanoylcarnitine | 1.473 | 372.311 | 1.095 | 1.382 | 0.00015 |
| L-Glutamine | 6.353 | 147.077 | 1.094 | 0.810 | <0.00001 |
| L-Arginine | 7.316 | 175.119 | 1.093 | 0.742 | <0.00001 |
| Dodecanoylcarnitine | 1.522 | 344.280 | 1.094 | 1.349 | 0.00695 |
| L-Proline | 5.346 | 116.071 | 1.092 | 0.758 | <0.00001 |
| Decanoylcarnitine | 1.593 | 316.248 | 1.090 | 1.330 | 0.02110 |
| **Between the LAA and SAO group** | | | | | |
| L-Pipecolic acid | 4.735 | 130.087 | 1.578 | 1.790 | 0.00065 |
| 1-Methylhistidine | 7.267 | 170.093 | 1.317 | 1.625 | 0.04886 |
| PE [22:6(4Z,7Z,10Z,13Z,16Z,19Z)/16:0] | 8.209 | 764.521 | 1.270 | 0.780 | 0.02704 |
| PE [P-18:0/18:2(9Z,12Z)] | 9.314 | 728.557 | 1.192 | 1.276 | 0.01645 |
| LysoPE [18:2(9Z,12Z)/0:0] | 2.428 | 478.292 | 1.113 | 1.446 | 0.00011 |
| LysoPC [18:3(9Z,12Z,15Z)] | 1.991 | 518.323 | 1.045 | 1.387 | 0.00144 |
| LysoPC (20:0) | 5.009 | 552.402 | 1.033 | 1.254 | 0.00669 |
| LysoPC [18:2(9Z,12Z)] | 2.455 | 520.339 | 1.000 | 1.336 | 0.00013 |

*LAA, large artery atherosclerosis; SAO, small artery occlusion; AIC, acute ischemic stroke; HC, healthy control; PE, phosphatidylethanolamine.*

by the brain is provided by the oxidative reaction of fatty acids (Ebert et al., 2003). Neurons are very sensitive to conditions such as ischemia and hypoxia. In order to regulate the lack of energy caused by the AIS, the brain can initiate energy production responses such as fatty acid degradation through negative feedback to maintain homeostasis (Schwartz et al., 2000; Belgardt and Brüning, 2010). Oleic acid and linoleic acid are long-chain fatty acids that can cross the blood–brain barrier to provide energy to the brain (Panov et al., 2014), and L-palmitoylcarnitine is also involved in fatty acid degradation. In this study, compared with the HC group, the increase in oleic acid, linoleic acid, and L-palmitoylcarnitine in the AIS group may be associated with increased fatty acid catabolism in the acute phase of the IS to maintain energy homeostasis. In addition, previous studies have reported that changes in lipid metabolism are associated with mitochondrial dysfunction caused by oxidative stress (Tobe, 2013), which is one of the three main pathophysiological reactions (neurotoxicity, oxidative stress, and inflammation) in IS (Fukuyama et al., 1998; Chamorro et al., 2012; Lai et al., 2014). This result is consistent with the transcriptomic profiling results of IS (Li et al., 2015; Cai et al., 2019). Cai et al. (2019) assessed the patterns of transcriptomic changes at different stages of IS using a mouse model. The results showed that mmu-miR-199a-5p and mmu-miR-199b-3p inhibit the inflammatory response during the recovery phase of IS and exert neuroprotective effects by regulating the Taok1 gene (Cai et al., 2019). This may imply that fatty acid metabolism is related to the regulation of these genes. However, this requires further verification in animal experiments.

Both AA and DHA are polyunsaturated fatty acids which are released from the metabolic pathway of glycerol phospholipid degradation and are the main components of the phospholipid membrane. AA and DHA participate in membrane fluidity, signal transduction, and gene transcription during the whole life process (Rapoport et al., 2001). They are also involved in many pathological processes, including stroke (Rapoport, 2008). AA is stored in the phospholipid membranes of cells, which produce free AA by deacylation mediated by phospholipase A2 (PLA2). In pathological environments such as stroke, free AA increases the production of free radicals through the "arachidonic acid cascade reaction" (Rink and Khanna, 2011). This reaction can occur as early as 1 h after a stroke (Shohami et al., 1982), a major factor in the oxidative damage of tissues after a stroke. Consistent with the results of our study, previous studies reported a significant increase in the types of reactive oxygen species and AA metabolism after reperfusion in IS (Gürsoy-Ozdemir et al., 2004).

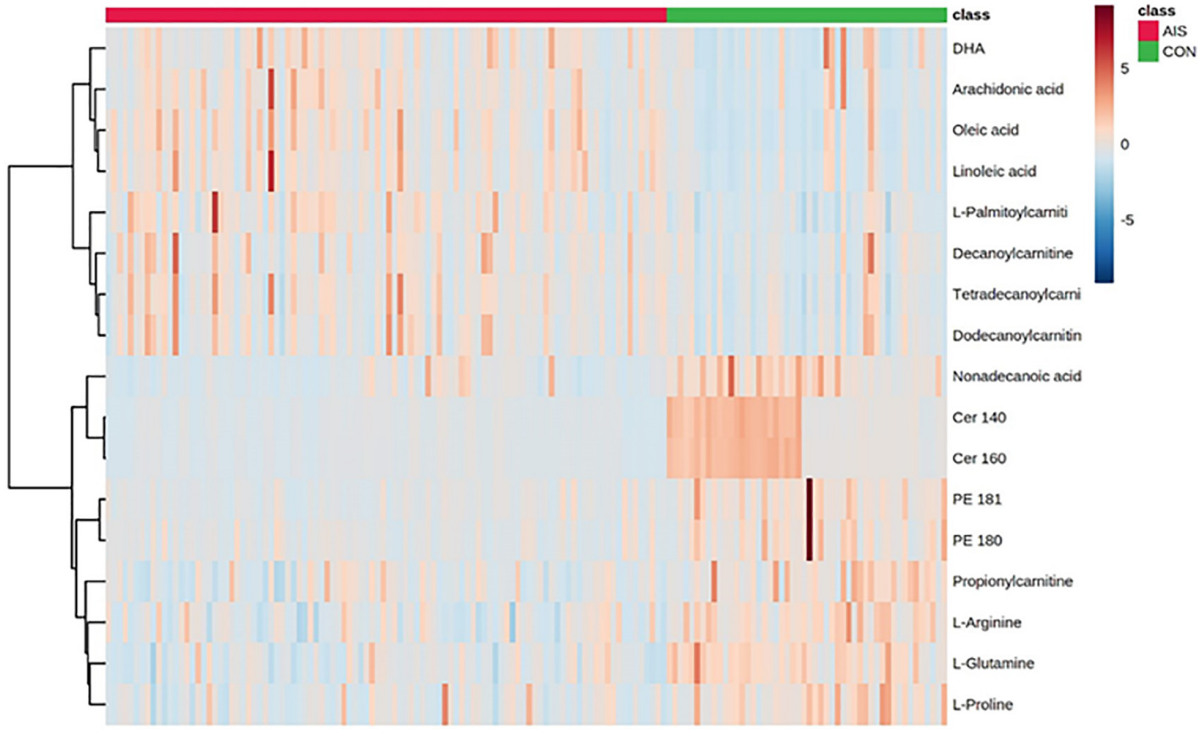



**FIGURE 2 |** Heat map of different serum metabolites between patients with acute ischemic stroke and healthy control groups.

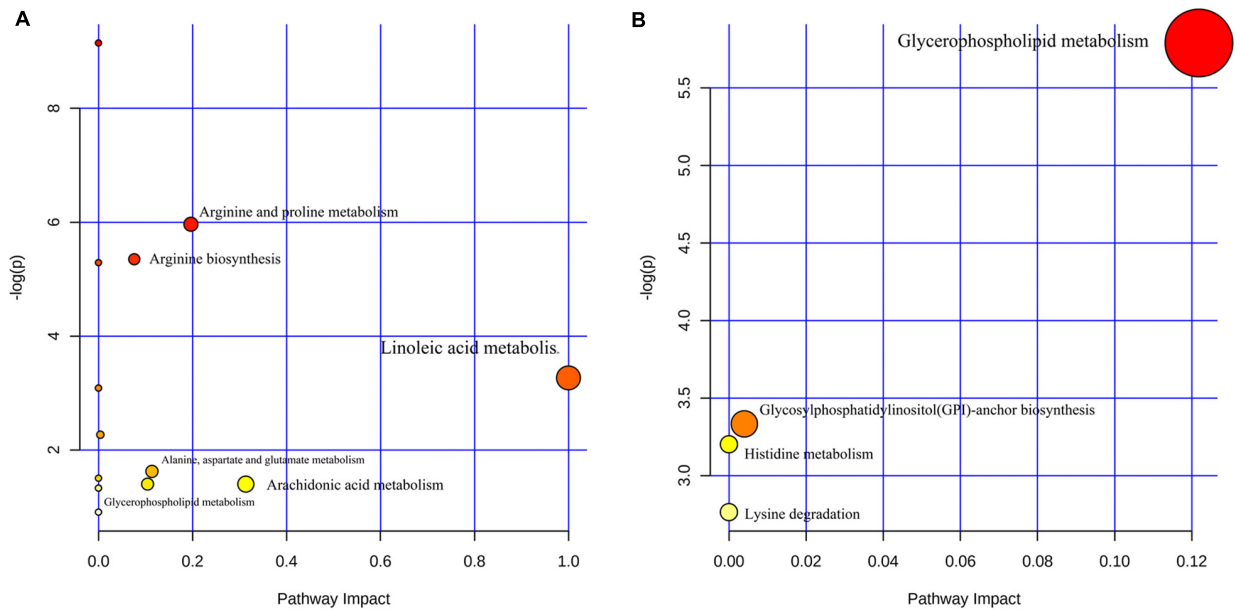



**FIGURE 3 |** Different metabolic pathways **(A)** between patients with acute ischemic stroke and healthy control groups and **(B)** between the large artery atherosclerosis and the small artery occlusion groups.

Ceramide is a type of waxy lipid composed of sphingosine and fatty acids, which plays a role in plaque formation (Holland and Summers, 2008). In addition, ceramide levels in the high-risk groups with IS were higher than those in the low-risk groups (Wang et al., 2017). However, previous studies reported that the level of ceramide in patients with AIS is lower than that in the HC group, which may be related to ceramide-mediated apoptosis (Taha et al., 2006). The specific mechanism warrants further study. Glutamine (Gln) and glutamate (Glu) can be converted into each other in the human body. The level of glutamine and

the ratio of Gln to Glu were negatively correlated with the risk factors of cardiovascular disease (including body mass index, waist circumference, fasting blood glucose, insulin, triglyceride, *etc*.) (Cheng et al., 2012). Zheng et al. (2016) demonstrated that the ratio of Gln to Glu was associated with a reduced risk of cardiovascular disease. Similarly, glutamine levels in the AIS group were lower than those in the HC group of our study.

In the different metabolites of the LAA group and the SAO group, L-pipecolic acid mainly affects the lysine degradation pathway because lysine is produced in the process of L-pipecolic acid degradation, and lysine has previously been shown to decrease in patients with IS (Kimberly et al., 2013; Lee et al., 2017). 1-Methylhistidine is involved in histidine metabolism, which is a metabolic byproduct of the antioxidant molecule carnosine and its analogs in the brain (Bellia et al., 2011). Hu et al. (2019) showed that the level of L-pipecolic acid in patients with post-stroke depression is lower than that in HCs but higher than that in patients with stroke. PE, LysoPC, and LysoPE participate in glycerophospholipid metabolism and GPI-anchor biosynthesis. LysoPE is a product of PE hydrolyzed by PLA2, which plays a role in cell-mediated cell signaling and the activation of other enzymes (Park et al., 2007). PE and LysoPC are intermediate products of glycerophospholipid metabolism, while AA and DHA are glycerophospholipid degradation products. It should be noted that the metabolic changes of glycerophospholipids can not only help diagnose AIS but also help distinguish different subtypes of IS. This might shed insight on our exploration of the pathological mechanisms of different subtypes. Studies have confirmed the correlation between lipid metabolites and AIS using lipidomic and metabolomics techniques (Yang et al., 2017; Au, 2018). Our results are consistent with these (Liu et al., 2017). To date, there has been no research exploring the mechanisms of different metabolite isomers showing different behaviors. This may require further lipidomic assessments in a larger sample size and further verification in animal models.

In this study, non-targeted metabolomics based on LC–MS was used to identify the different metabolites between patients with AIS and the HCs and between the LAA and SAO groups, providing new insights and encouraging further study of the pathophysiological mechanisms among different subtypes of IS. However, this study still has many limitations. First, this was a single-center study with a relatively small sample size. Multicenter studies with larger sample sizes will be needed to validate our findings. In addition, because the metabolites in the human body change dynamically, we collected the serum after the onset of the disease, which may have affected the estimation of the correlation between the metabolic differences and the disease. A longitudinal comparison of multiple blood samples, after the onset, from the same patient will enable a clearer assessment of the changes in serum metabolites in AIS patients. Finally, targeted metabolomics technology in another set of samples is needed to further verify the different metabolites. In the future, to validate the results and investigate the potential of metabolites as biomarkers, we will include patients and follow them up prospectively to obtain their modified Rankin Scale scores and further explore metabolites associated with prognosis.

## CONCLUSION

In summary, this study identified the different metabolites and metabolic pathways in patients with AIS and HCs and between the LAA and SAO subtypes of IS by non-targeted metabolomics. We demonstrated that metabolomics might be used to diagnose AIS and distinguish its subtypes. Further research is needed to explore the pathophysiological mechanisms that affect the changes in metabolites and lead to new clinical diagnoses and potential interventions.

## DATA AVAILABILITY STATEMENT

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

## ETHICS STATEMENT

The studies involving human participants were reviewed and approved by The First Affiliated Hospital of Zhengzhou University. The patients/participants provided their written informed consent to participate in this study.

## AUTHOR CONTRIBUTIONS

XW, BS, and J-SK conceived and designed the research. XW, MT, and LZ conducted the experiments. XW, MT, LZ, L-lP, WS, and JL performed the data collection. XW, XL, RZ, JW, and SS analyzed the data. XW wrote the manuscript. All authors have read and approved the manuscript.

## FUNDING

This study was funded by the National Key Research and Development Program, Major Chronic Non-communicable Disease Prevention and Control Research Key Special Project (2017YFC1308202), and Henan Provincial Medical Science and Technology Research Plan (SBGJ2018031).

## REFERENCES

Adams, H. P., Bendixen, B. H., Kappelle, L. J., Biller, J., Love, B. B., Gordon, D. L., et al. (1993). Classification of subtype of acute ischemic stroke. Definitions for use in a multicenter clinical trial. TOAST. Trial of Org 10172 in acute stroke treatment. *Stroke* 24, 35–41. doi: 10.1161/01.str.24.1.35

Au, A. (2018). Metabolomics and lipidomics of ischemic stroke. *Adv. Clin. Chem.* 85, 31–69. doi: 10.1016/bs.acc.2018.02.002

Belgardt, B. F., and Brüning, J. C. (2010). CNS leptin and insulin action in the control of energy homeostasis. *Ann. N. Y. Acad. Sci.* 1212, 97–113.

Bellia, F., Vecchio, G., Cuzzocrea, S., Calabrese, V., and Rizzarelli, E. (2011). Neuroprotective features of carnosine in oxidative driven diseases. *Mol. Aspects Med.* 32, 258–266. doi: 10.1016/j.mam.2011.10.009

Benjamin, E. J., Blaha, M. J., Chiuve, S. E., Cushman, M., Das, S. R., Deo, R., et al. (2017). Heart disease and stroke statistics-2017 update: a report from the american heart association. *Circulation* 135, e146–e603. doi: 10.1161/CIR.0000000000000485

Cai, Y., Zhang, Y., Ke, X., Guo, Y., Yao, C., Tang, N., et al. (2019). Transcriptome sequencing unravels potential biomarkers at different stages of cerebral ischemic stroke. *Front. Genet.* 10:814. doi: 10.3389/fgene.2019.00814

Chamorro, Á, Meisel, A., Planas, A. M., Urra, X., van de Beek, D., and Veltkamp, R. (2012). The immunology of acute stroke. *Nat. Rev. Neurol.* 8, 401–410. doi: 10.1038/nrneurol.2012.98

Chen, P.-H., Gao, S., Wang, Y.-J., Xu, A.-D., Li, Y.-S., and Wang, D. (2012). Classifying ischemic stroke, from TOAST to CISS. *CNS Neurosci. Therap.* 18, 452–456. doi: 10.1111/j.1755-5949.2011.00292.x

Cheng, S., Rhee, E. P., Larson, M. G., Lewis, G. D., McCabe, E. L., Shen, D., et al. (2012). Metabolite profiling identifies pathways associated with metabolic risk in humans. *Circulation* 125, 2222–2231. doi: 10.1161/CIRCULATIONAHA.111.067827

Ebert, D., Haller, R. G., and Walton, M. E. (2003). Energy contribution of octanoate to intact rat brain metabolism measured by 13C nuclear magnetic resonance spectroscopy. *J. Neurosci.* 23, 5928–5935. doi: 10.1523/jneurosci.23-13-05928.2003

Fukuyama, N., Takizawa, S., Ishida, H., Hoshiai, K., Shinohara, Y., and Nakazawa, H. (1998). Peroxynitrite formation in focal cerebral ischemia-reperfusion in rats occurs predominantly in the peri-infarct region. *J. Cereb. Blood Flow Metab.* 18, 123–129. doi: 10.1097/00004647-199802000-00001

Gürsoy-Ozdemir, Y., Can, A., and Dalkara, T. (2004). Reperfusion-induced oxidative/nitrative injury to neurovascular unit after focal cerebral ischemia. *Stroke* 35, 1449–1453. doi: 10.1161/01.str.0000126044.83777.f4

Holland, W. L., and Summers, S. A. (2008). Sphingolipids, insulin resistance, and metabolic disease: new insights from in vivo manipulation of sphingolipid metabolism. *Endocr. Rev.* 29, 381–402. doi: 10.1210/er.2007-2025

Hu, Z., Fan, S., Liu, M., Zhong, J., Cao, D., Zheng, P., et al. (2019). Objective diagnosis of post-stroke depression using NMR-based plasma metabonomics. *Neuropsychiatr. Dis. Treat.* 15, 867–881. doi: 10.2147/NDT.S192307

Jickling, G. C., and Sharp, F. R. (2015). Biomarker panels in ischemic stroke. *Stroke* 46, 915–920. doi: 10.1161/STROKEAHA.114.005604

Jung, J. Y., Lee, H.-S., Kang, D.-G., Kim, N. S., Cha, M. H., Bang, O.-S., et al. (2011). 1H-NMR-based metabolomics study of cerebral infarction. *Stroke* 42, 1282–1288. doi: 10.1161/STROKEAHA.110.598789

Kelly, P. J., Albers, G. W., Chatzikonstantinou, A., De Marchis, G. M., Ferrari, J., George, P., et al. (2016). Validation and comparison of imaging-based scores for prediction of early stroke risk after transient Ischaemic attack: a pooled analysis of individual-patient data from cohort studies. *Lancet Neurol.* 15, 1238–1247. doi: 10.1016/S1474-4422(16)30236-30238

Kimberly, W. T., Wang, Y., Pham, L., Furie, K. L., and Gerszten, R. E. (2013). Metabolite profiling identifies a branched chain amino acid signature in acute cardioembolic stroke. *Stroke* 44, 1389–1395. doi: 10.1161/STROKEAHA.111.000397

Lai, T. W., Zhang, S., and Wang, Y. T. (2014). Excitotoxicity and stroke: identifying novel targets for neuroprotection. *Prog. Neurobiol.* 115, 157–188. doi: 10.1016/j.pneurobio.2013.11.006

Latchaw, R. E., Alberts, M. J., Lev, M. H., Connors, J. J., Harbaugh, R. E., Higashida, R. T., et al. (2009). Recommendations for imaging of acute ischemic stroke: a scientific statement from the American heart association. *Stroke* 40, 3646–3678. doi: 10.1161/STROKEAHA.108.192616

Lee, Y., Khan, A., Hong, S., Jee, S. H., and Park, Y. H. (2017). A metabolomic study on high-risk stroke patients determines low levels of serum lysine metabolites: a retrospective cohort study. *Mol. Biosyst.* 13, 1109–1120. doi: 10.1039/c6mb00732e

Li, Y., Mao, L., Gao, Y., Baral, S., Zhou, Y., and Hu, B. (2015). MicroRNA-107 contributes to post-stroke angiogenesis by targeting Dicer-1. *Sci. Rep.* 5:13316. doi: 10.1038/srep13316

Liu, P., Li, R., Antonov, A. A., Wang, L., Li, W., Hua, Y., et al. (2017). Discovery of metabolite biomarkers for acute ischemic stroke progression. *J. Proteome Res.* 16, 773–779. doi: 10.1021/acs.jproteome.6b00779

Montaner, J., Perea-Gainza, M., Delgado, P., Ribó, M., Chacón, P., Rosell, A., et al. (2008). Etiologic diagnosis of ischemic stroke subtypes with plasma biomarkers. *Stroke* 39, 2280–2287. doi: 10.1161/STROKEAHA.107.505354

Panov, A., Orynbayeva, Z., Vavilin, V., and Lyakhovich, V. (2014). Fatty acids in energy metabolism of the central nervous system. *Biomed. Res. Intern.* 2014:472459. doi: 10.1155/2014/472459

Park, K. S., Lee, H. Y., Lee, S. Y., Kim, M.-K., Kim, S. D., Kim, J. M., et al. (2007). Lysophosphatidylethanolamine stimulates chemotactic migration and cellular invasion in SK-OV3 human ovarian cancer cells: involvement of pertussis toxin-sensitive G-protein coupled receptor. *FEBS Lett.* 581, 4411–4416. doi: 10.1016/j.febslet.2007.08.014

Rapoport, S. I. (2008). Arachidonic acid and the brain. *J. Nutr.* 138, 2515–2520.

Rapoport, S. I., Chang, M. C., and Spector, A. A. (2001). Delivery and turnover of plasma-derived essential PUFAs in mammalian brain. *J. Lipid Res.* 42, 678–685.

Rink, C., and Khanna, S. (2011). Significance of brain tissue oxygenation and the arachidonic acid cascade in stroke. *Antioxid. Redox Signal.* 14, 1889–1903. doi: 10.1089/ars.2010.3474

Schwartz, M. W., Woods, S. C., Porte, D., Seeley, R. J., and Baskin, D. G. (2000). Central nervous system control of food intake. *Nature* 404, 661–671. doi: 10.1038/35007534

Shohami, E., Rosenthal, J., and Lavy, S. (1982). The effect of incomplete cerebral ischemia on prostaglandin levels in rat brain. *Stroke* 13, 494–499. doi: 10.1161/01.str.13.4.494

Song, B., Fang, H., Zhao, L., Gao, Y., Tan, S., Lu, J., et al. (2013). Validation of the ABCD3-I score to predict stroke risk after transient ischemic attack. *Stroke* 44, 1244–1248. doi: 10.1161/STROKEAHA.113.000969

Taha, T. A., Mullen, T. D., and Obeid, L. M. (2006). A house divided: ceramide, sphingosine, and sphingosine-1-phosphate in programmed cell death. *Biochim. Biophys. Acta* 1758, 2027–2036. doi: 10.1016/j.bbamem.2006.10.018

Tobe, E. H. (2013). Mitochondrial dysfunction, oxidative stress, and major depressive disorder. *Neuropsychiatr. Dis. Treat.* 9, 567–573. doi: 10.2147/NDT.S44282

Wang, D. D., Toledo, E., Hruby, A., Rosner, B. A., Willett, W. C., Sun, Q., et al. (2017). Plasma ceramides, mediterranean diet, and incident cardiovascular disease in the PREDIMED trial (Prevención con Dieta Mediterránea). *Circulation* 135, 2028–2040. doi: 10.1161/CIRCULATIONAHA.116.024261

Wang, H., Liddell, C. A., Coates, M. M., Mooney, M. D., Levitz, C. E., Schumacher, A. E., et al. (2014). Global, regional, and national levels of neonatal, infant, and under-5 mortality during 1990-2013: a systematic analysis for the Global Burden of Disease Study 2013. *Lancet* 384, 957–979. doi: 10.1016/S0140-6736(14)60497-60499

Wang, X., Tian, X., Pei, L.-L., Niu, P.-P., Guo, Y., Hu, R., et al. (2020). The association between serum Apelin-13 and the prognosis of acute ischemic stroke. *Transl. Stroke Res.* 11, 700–707. doi: 10.1007/s12975-019-00769-w

World Health Organization [WHO] (1989). Stroke–1989. Recommendations on stroke prevention, diagnosis, and therapy. Report of the WHO task force on stroke and other cerebrovascular disorders. *Stroke* 20, 1407–1431. doi: 10.1161/01.str.20.10.1407

Yang, L., Lv, P., Ai, W., Li, L., Shen, S., Nie, H., et al. (2017). Lipidomic analysis of plasma in patients with lacunar infarction using normal-phase/reversed-phase two-dimensional liquid chromatography-quadrupole time-of-flight mass spectrometry. *Anal. Bioanal. Chem.* 409, 3211–3222. doi: 10.1007/s00216-017-0261-266

Zheng, Y., Hu, F. B., Ruiz-Canela, M., Clish, C. B., Dennis, C., Salas-Salvado, J., et al. (2016). Metabolites of glutamate metabolism are associated with incident cardiovascular events in the PREDIMED PREvención con DIeta MEDiterránea (PREDIMED) Trial. *J. Am. Heart Assoc.* 5:e003755. doi: 10.1161/JAHA.116.003755

**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 Neuroscience

MINI REVIEW
published: 22 April 2021
doi: 10.3389/fnins.2021.591122

# Single-Cell Transcriptomics: Current Methods and Challenges in Data Acquisition and Analysis

*Asif Adil[1†], Vijay Kumar[2†], Arif Tasleem Jan[3*] and Mohammed Asger[1*]*

[1] Department of Computer Sciences, Baba Ghulam Shah Badshah University, Rajouri, India, [2] Department of Biotechnology, Yeungnam University, Gyeongsan, South Korea, [3] School of Biosciences and Biotechnology, Baba Ghulam Shah Badshah University, Rajouri, India

Rapid cost drops and advancements in next-generation sequencing have made profiling of cells at individual level a conventional practice in scientific laboratories worldwide. Single-cell transcriptomics [single-cell RNA sequencing (SC-RNA-seq)] has an immense potential of uncovering the novel basis of human life. The well-known heterogeneity of cells at the individual level can be better studied by single-cell transcriptomics. Proper downstream analysis of this data will provide new insights into the scientific communities. However, due to low starting materials, the SC-RNA-seq data face various computational challenges: normalization, differential gene expression analysis, dimensionality reduction, etc. Additionally, new methods like 10× Chromium can profile millions of cells in parallel, which creates a considerable amount of data. Thus, single-cell data handling is another big challenge. This paper reviews the single-cell sequencing methods, library preparation, and data generation. We highlight some of the main computational challenges that require to be addressed by introducing new bioinformatics algorithms and tools for analysis. We also show single-cell transcriptomics data as a big data problem.

**Keywords: single-cell transcriptomics, Sc-RNA-seq, big data, single-cell big data, normalization, single-cell analysis, downstream analysis**

OPEN ACCESS

***Edited by:***
*Kumardeep Chaudhary, Icahn School of Medicine at Mount Sinai, United States*

***Reviewed by:***
*Ankush Sharma, University of Oslo, Norway*
*Xun Zhu, University of Hawaii Cancer Center, United States*

***Correspondence:**
*Arif Tasleem Jan*
*atasleem@bgsbu.ac.in*
*Mohammed Asger*
*masgerghazi@bgsbu.ac.in*

*† These authors have contributed equally to this work*

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

***Received:*** *03 August 2020*
***Accepted:*** *19 March 2021*
***Published:*** *22 April 2021*

***Citation:***
*Adil A, Kumar V, Jan AT and Asger M (2021) Single-Cell Transcriptomics: Current Methods and Challenges in Data Acquisition and Analysis. Front. Neurosci. 15:591122. doi: 10.3389/fnins.2021.591122*

## INTRODUCTION

The human body exhibits a diverse range of cells that undergo transit from one state to another in life (development, disease, and regeneration). Though derived from the same zygote, the cell, with its types and states, is greatly influenced by the internal processes and external factors (Song et al., 2019). In its progression through proliferation and the differentiation states to generate multiple cell types for organ formation, complex heterogeneities in the cellular architecture are observed. The cellular heterogeneity in terms of morphology, function, and gene expression profiles lie between various tissues, but has also been observed among the same cell types that allow them to perform different roles. Dysregulation in any particular cell type (irrespective of tissues, organs, and organ-system) influences the entire system that progresses to disorders and even severe diseases like cancer (Macaulay et al., 2017).

Recent technological advancements have enabled biologists to profile *cells at individual levels* on a variety of omics layers (genomes, transcriptomes, epigenomes, and proteomes) (Hu et al., 2016); among these, single cell (SC) transcriptomics is widely studied. The cells of a human body, being

heterogeneous, often show a drastic variation at the individual level (Wang and Bodovitz, 2010; Xin et al., 2016). The SC experiments were found much conclusive compared with bulk cell sequencing that involves sequencing in bulk (assuming cells of a particular type are identical) and estimating an average of expressions. The SC transcriptomics was awarded as method of the year by *Nature* in 2013 (Xue et al., 2015). With the advent of next-generation sequencing, it becomes possible to develop sequencing methods to probe the dynamics of the genome and variations thereof. Of them, RNA sequencing (RNA-seq)-mediated transcriptomic profiling revealed information of novel RNA species that deepened our understanding of the transcriptome dynamics (Tang et al., 2009; Wang et al., 2009; Ozsolak and Milos, 2011). Lately, these sequencing approaches have been extended to study intra-population heterogeneity of SCs (Wills et al., 2013), whereby it enabled the study of cell fates, their transition to different subtypes, and the dynamics of gene expression masked in bulk population studies (Altschuler and Wu, 2010; Trapnell et al., 2014). Compared with bulk sequencing, where libraries are prepared from thousands of cells, libraries for single-cell RNA sequencing (SC-RNA-seq) are cell-specific towards investigating cellular functionalities of DNA and RNA in different cellular subsets (Gross et al., 2015; Xue et al., 2015). Though SC-RNA-seq has revealed novel findings in different cellular backgrounds, it poses specific challenges: Pre-processing of the SC-RNA-seq data is majorly different from bulk RNA-seq, stricter protocols for library preparation and low starting material. Another challenge is the lack of analytical approaches required to accommodate large datasets generated during SC-RNA-seq experiments. Keeping this in view, we investigated the methods adopted in SC experiments, sequencing approaches, and challenges thereof, as part of realizing the goal of precision medicine.

## SINGLE-CELL RNA SEQUENCE PROFILING TECHNIQUES

With the first report in 2009, a surge in the SC transcriptomics methods capable of sequencing millions of cells with great accuracy and viability in a short span of time was observed (Tang et al., 2009). These methods are generally different from each other in terms of cell isolation methods, cell lysis procedure, amplification process, cDNA generation, transcript coverage, and Unique Molecular Identifier (UMI) tagging (at either 3′ end or 5′ end). The most critical distinction in the SC-RNA profiling techniques is that some provide full-length transcript coverage and some only partially sequence from either 3′ end or 5′ end of the transcript (Chen et al., 2019). **Table 1** highlights widely used SC-RNA profiling methods in terms of different properties.

**TABLE 1 |** Current SC-RNA-seq profiling techniques, based on transcript coverage and UMI insertion possibility.

| Method | Length of transcript | UMI insertion possibility | References |
|---|---|---|---|
| ScNaUmi-seq | Full length | Yes | Lebrigand et al., 2020 |
| MATQ-seq | Full length | Yes | Sheng and Zong, 2019 |
| 10× Chromium | 3′ end | Yes | Zheng et al., 2017 |
| CEL-seq2 | 3′ end | Yes | Hashimshony et al., 2016 |
| Drop-seq | 3′ end | Yes | Macosko et al., 2015 |
| InDrop | 3′ end | Yes | Klein et al., 2015 |
| Smart-seq2 | Full length | No | Picelli et al., 2014 |
| STRT-seq | 5′ end | Yes | Islam et al., 2014 |
| MARS-seq | 3′ end | Yes | Jaitin et al., 2014 |
| Smart-seq | Full length | No | Ramskold et al., 2013 |

*SC-RNA-seq, single-cell RNA sequencing; UMI, Unique Molecular Identifier.*

## OPTIMAL METHODOLOGY OF SINGLE-CELL TRANSCRIPTOMICS

Of the various sequencing platforms, Drop-seq, InDrop, and 10× Chromium are well-known platforms for sequencing hundreds and thousands of cells in an unbiased manner (Kulkarni et al., 2019). In SC transcriptomics, each cell needs to be isolated from its originating tissue. The Droplet-based techniques, which at the core use microfluidics to attach cells with beads containing a unique barcode, are widely incorporated to separate cells. The performance criteria for isolation methods are based on three parameters: throughput, purity, and recovery (Tomlinson et al., 2013; Gross et al., 2015). *Throughput* indicates the number of cells that can be isolated per unit time, *purity* refers to the number of cells collected after separation from tissue, and *recovery* is the final amount of the target cells, in hand, after separation. The morphological complexity of cells like those of the central nervous system (CNS) makes the separation process a little challenging. The segregation process exposes them to specific environmental, chemical, and harsh dissociation steps that often bias data analysis (Kulkarni et al., 2019). The dissociation of intact cells from a frozen postmortem tissue is also challenging, as cell membranes are prone to damage from mechanical and physical stresses as part of the freeze–thaw process (McGann et al., 1988). Though each cell separation methods currently in use shows an advantage different for the above three parameters, it becomes imperative to select a well-suited method for the isolation of a cell. The current methodology of cell separation is broadly categorized into two groups based on (1) cellular properties like cell density, cell shape, cell size, etc., and (2) biological characteristics of a cell that comprises affinity methods (Tomlinson et al., 2013). **Tables 2**, **3** show some of the widely used methods concerning the operational mode, throughput, advantages, and disadvantages.

Though high-throughput SC-RNA approaches such as 10× Chromium allows analysis of cells in an unbiased manner, it lacks in providing an in-depth information on sequence diversity, splicing, and chimeric transcripts generated in the process (Lebrigand et al., 2020). The problem is overcome by performing Nanopore long-read sequencing [using a cell barcode (cellBC) assignment to long reads] to obtain a full-length sequence corresponding to the 10× Chromium system's data. As SC library preparation requires robust amplification, chimeric cDNA generation and amplification bias issues are

**TABLE 2 |** Commonly used methods for cell isolation based on biological characteristics.

| Technique | Mode of operation | Throughput | Advantages | Disadvantages | References |
|---|---|---|---|---|---|
| Fluorescence-activated cell sorting | Automatic | High | High rate of rare cell sorting, high purity | Cost-intensive, high skills required | Herzenberg et al., 2002; Gross et al., 2015 |
| Magnetic-activated cell separation | Automatic | High | High purity, cost-efficient | Cell capture is non-specific | Schmitz et al., 1994; Welzel et al., 2015 |

**TABLE 3 |** Commonly used methods for cell isolation on the bases of physical characteristics.

| Technique | Mode of operation | Throughput | Advantages | Disadvantages | References |
|---|---|---|---|---|---|
| Microfluidic cell separation | Automatic | High | Works with low starting materials, amplification integration | High skills required, dissociated cells | Wyatt Shields et al., 2015 |
| Micromanipulation manual cell picking | Manual | Low | More control over cell, live and intact cell separation | Laborious, high skills needed | Citri et al., 2012 |
| Laser-capture microdissection | Manual | Low | Undamaged live cell capture, highly advanced | Too complex to operate, threat of contamination by neighboring cells | Espina et al., 2006 |
| Density gradient centrifugation | Manual | Low | Cost-efficient | Too slow and laborious, low yield | Beakke, 1951 |

currently addressed by employing a 3′ or 5′ end tag-based approach (Trombetta et al., 2015; Natarajan et al., 2019). The sequence length method determines the quality of alignment across the total length of a gene, while tag-based methods integrate UMIs at either 3′ end or 5′ end of the transcript (Kivioja et al., 2012; Smith et al., 2017; Sena et al., 2018). The UMI addition makes it easier to identify and quantify the individual transcripts by eliminating PCR artifacts and minimizes false annotation of PCR-generated chimeric cDNAs as novel transcripts. The full length-based methodology provides an all-inclusive coverage of the reads, yet they contribute a bias for long genes, as the genes with shorter length are often missed (Phipson et al., 2017). Additionally, the higher sequencing error rate of long-read sequencers and UMI problems account for a serious issue pertaining to these platforms (Gupta et al., 2018; Lebrigand et al., 2020; Volden and Vollmers, 2020). Despite this, the Tag-based methods have shown a fair dominance in SC-RNA library preparation for quantifying the transcripts in SC analysis when cell number is large (**Figure 1**).

## QUANTIFICATION OF EXPRESSION AND QUALITY CONTROL

Like bulk RNA-seq, the transcripts in SC-RNA are sequenced into reads that generate the raw fastq data. The quality of the sequence reads generated in a sequencing method is considered an important quality indicator of SC-RNA-seq data. As the alignment of the transcript reads for SC-RNA-seq is same as bulk RNA-seq, the methods and tools used for the gene or transcript quantification for bulk RNA-seq can also be used for quantifying transcripts generated by SC-RNA-seq (Li and Homer, 2010; Fonseca et al., 2012). HISAT2 (Kim et al., 2019), TopHat2 (Kim et al., 2013), and STAR (Dobin et al., 2013) are currently the most popular alignment tools, which can map billions of reads to a reference transcriptome with greater accuracy and high speed. Transcriptome reconstruction can be either *de novo* (for samples lacking reference genome) or reference based, also called genome-guided assembly (Chen et al., 2011). However, the former technique sometimes lacks accuracy in comparison with the reference-based assembly approach (Garber et al., 2011). For SC-RNA-seq methods that generate data on a whole-transcriptome basis, Smart-seq2 (Picelli et al., 2014) and MATQ-seq (Sheng and Zong, 2019) use Cufflinks, RSEM, Stringtie, etc., for the quantification of transcripts, while methods that incorporate the 3′ end UMI tagging [like Drop-seq (Macosko et al., 2015), InDrop (Klein et al., 2015), MARS-seq (Jaitin et al., 2014), etc.] require specific algorithms to generate the expression count for the transcript. Another efficient tool for the UMI-based methods was developed by Huang and Sanguinetti (2017) for calculating the expression count of SCs accurately. **Table 4** provides information about the current tools for read alignment and expression quantification. The SC-RNA-seq exhibits certain limitations, which results in higher technical noise (Kolodziejczyk et al., 2015). In SC-RNA-seq data, many transcripts appear to be lost during reverse transcription due to the small number and low capture efficiency of RNA molecules in SCs (Saliba et al., 2014). Consequently, in one cell, some transcripts are highly expressed but are missing in another cell. This pattern is described as a "dropout" event. It has been reported that even the most sensitive protocol for SC-RNA-seq fails to detect some of the transcripts as part of Dropout events (Haque et al., 2017). When the cells are dissociated or isolated, a certain number of cells become dead or get destroyed. The SC-RNA-seq methods generate low-quality data from these cells (Ilicic et al., 2016). After alignment and quantification of the transcripts, the quality control check of cells is necessary to remove low-quality cells for an accurate downstream analysis.

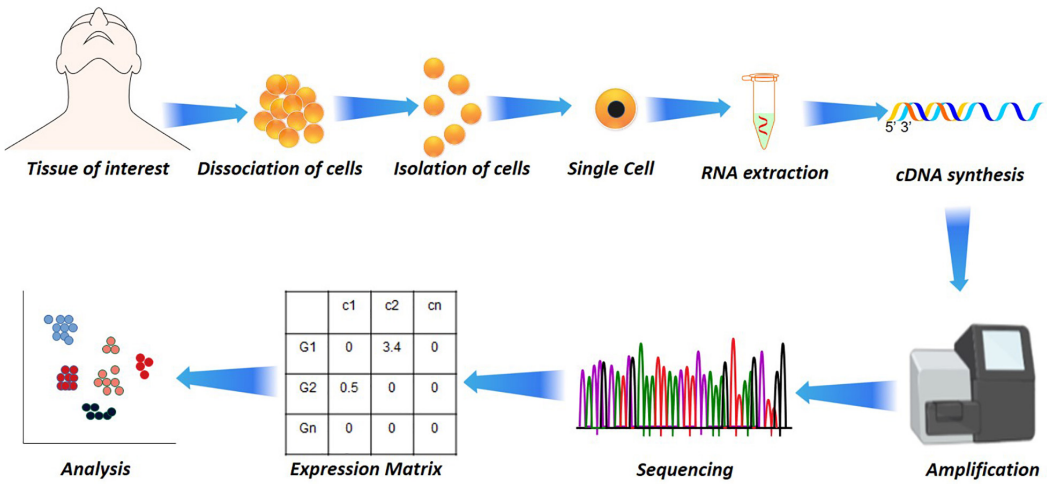



**FIGURE 1 |** Single-cell analysis in disease and health. Starting from the dissociation of target cells from the target tissue/organ, their isolation based on fluorescence-activated cell sorting (FACS) or other microfluidic techniques to RNA extraction. The RNA extraction is followed by cDNA synthesis by reverse transcriptase, followed by amplification and sequencing. From the sequencing, the reads are aligned and subjected to quantification that results in a quantification matrix or Gene Expression Matrix.

**TABLE 4 |** Widely used tools for read alignment and expression quantification.

| Tool | Function | Feature | URL | References |
|---|---|---|---|---|
| Salmon | Expression quantification | k-mer-based read quantification | https://combine-lab.github.io/salmon/ | Patro et al., 2017 |
| Kallisto | Expression quantification | Pseudoalignment-based rapid read determination | https://pachterlab.github.io/kallisto/ | Bray et al., 2016 |
| StringTIe | Expression quantification | Alignment based, splice aware | https://ccb.jhu.edu/software/stringtie/ | Pertea et al., 2015 |
| HISAT2 | Read alignment | Alignment based, splice aware | https://daehwankimlab.github.io/hisat2/ | Sirén et al., 2014 |
| Sailfish | Expression quantification | k-mer-based read quantification | http://www.cs.cmu.edu/~{}ckingsf/software/sailfish/ | Patro et al., 2014 |
| RNA-Skim | Expression quantification | *Sig-mer* (a type of k-mer)-based read quantification of transcripts | http://www.csbio.unc.edu/rs/ | Zhang and Wang, 2014 |
| TopHat2 | Read alignment | Alignment based, splice aware | https://ccb.jhu.edu/software/tophat/index.shtml | Kim et al., 2013 |
| STAR | Read alignment | Alignment based, splice aware | https://github.com/alexdobin/STAR | Dobin et al., 2013 |
| Bowtie | Read alignment | Maintains quality threshold, hence less no. of mismatches | http://bowtie-bio.sourceforge.net/index.shtml | Langmead et al., 2009 |
| Cufflinks | Expression quantification | Alignment based, splice aware | https://github.com/cole-trapnell-lab/cufflinks | Trapnell et al., 2010 |

## CHALLENGES IMPEDING SINGLE-CELL RNA SEQUENCE DATA ANALYSIS

Though SC-RNA-seq has deepened our understanding of the cellular heterogeneity and molecular basis of life, it is impeded by several technical and computational challenges. The foremost among them is that its datasets exhibit a considerable amount of noise attributed to meager starting materials that often causes faulty downstream analysis and erroneous results (Brennecke et al., 2013). The SC-RNA-seq data analysis is performed as subtle execution in computational steps; read alignment, expression count generation, cell quality control, normalizing the data, and then further downstream analysis including SC clustering, differential gene expression (DGE), pseudo-temporal analysis, etc. In addition to low starting materials, the technical noise in the datasets is contributed by various factors, like batch effects (Haghverdi et al., 2018) and the low capture efficiency of protocols (Hwang et al., 2018). A few of the analytical steps, including read alignment and generation of count matrix, can be resolved using already available computational methods

designed for bulk RNA-seq. However, data processing tasks like normalization, DGE analysis, cell imputation, and dimensionality reduction, etc., call for the development of novel computational techniques, algorithms, and tools for smooth execution of SC-RNA-seq data analysis. The nature of the challenges that SC-RNA-seq data possess, including big data problem (Costa, 2012; Yu and Lin, 2016; Angerer et al., 2017; He et al., 2017), is highlighted in the following subsections:

## Normalization

In SC-RNA-seq, coverage of sequences between the libraries exhibit systematic differences from experimental procedures, dropout events, depth of the sequencing, and other technical effects (Stegle et al., 2015). These differences must be corrected by normalizing the data such that there is no interference in the comparison of the gene expression between cells. Being crucial, normalization of the SC-RNA-seq datasets eventually leads to lucid downstream analysis, including identifying different cell subsets and revealing differential expression of genes. In bulk RNA-seq, expression counts from various libraries are usually normalized by computing the fragments per kilobase of transcript counts of per million mapped fragments (FPKM) (Mortazavi et al., 2008), transcripts per million (TPM) (Li and Dewey, 2011), reads per kilobase of transcripts per million mapped reads (RPKM), upper quartile (UQ) (Bullard et al., 2010), DESeq (Love et al., 2014), removed unwanted variation (RUV) (Risso et al., 2014), and Gamma regression model (Ding et al., 2015). Generally, there are two types of normalization: (1) normalization of data within the sample, and (2) normalization of the data between the sample (Vallejos et al., 2015, 2017). In the former, FPKM/RPKM or TPM are used to exclude gene-specific biases (Vallejos et al., 2017) such as guanine–cytosine (GC) content and gene length, while in the latter, the normalization method tunes the sample-specific differences such as sequencing depth and capture efficiency. While ignoring the underlying stochasticity, normalization generates a relative expression estimate (Stegle et al., 2015), assuming the overall processed RNA per sample is equal (AlJanahi et al., 2018; Olsen and Baryawno, 2018). The bulk-based strategies for normalization have been reported unsuitable for SC-RNA-seq datasets because the datasets are highly zero-inflated and have higher technical noise. Multiple methods have been developed for normalizing the SC-RNA-seq data (Vallejos et al., 2015; Lun et al., 2016; Sengupta et al., 2016; Bacher et al., 2017; Yip et al., 2017). However, $O(nlogn)$ is considered more efficient than others in performing normalization of SC-RNA-seq data (Yip et al., 2017).

## Dimensionality Reduction

High dimensionality is yet another challenge that SC-RNA-seq data present. Owing to the data coming from cells showing high dimensions, i.e., a large number of genes, it is necessary to reduce (while optimally preserving the critical properties) the set of random variables and work with the principle variables which describe the data profoundly (Andrews and Hemberg, 2019). The two most frequently used methods for dimensionality reduction are principal component analysis (PCA) (Van Der Maaten et al., 2009) and T-distribution stochastic neighbor embedding (t-SNE) (Van Der Maaten and Hinton, 2008; Kobak and Berens, 2019). PCA uses a linear process to transform a set of variables (possibly correlated) into an uncorrelated variable known as a principal component, while t-SNE is a non-linear probability distribution-based approach. Both PCA and t-SNE methods of dimensionality reduction have certain limitations (Chen et al., 2019); based on the assumption that approximately all the data are distributed normally, PCA does not effectively amount to the underlying complexities in the structure of SC-RNA-seq data, and t-SNE has a larger time complexity reaching $O(n^2)$ (Pezzotti et al., 2017). The most recent algorithm employed for dimensionality reduction "UMAP" (Uniform Manifold Approximation and Projection) (McInnes et al., 2018; Becht et al., 2019) outperforms PCA and t-SNE for SC-RNA-seq in terms of high reproducibility and meaningful organization of cells (Becht et al., 2018). UMAP is a non-linear graph-based algorithm that tends to identify the closest neighbors of a data point and assigns them a larger weight, thereby preserving the topological structure of the data. The idea is to project a low-dimensional representation of the data while preserving the nearest neighbours of an individual data point (i.e., cells). This helps to group more closely related neighbours and partly conserves the relation of points in the "long-range" using the intermediate data points. Although the interpretation of the distances in a reduced space becomes difficult, UMAP has been largely able to uncover the elusive features of the data. UMAP is computationally faster than t-SNE, preserves the global structure, and maintains the continuity of cell subsets (Becht et al., 2018). At the core, UMAP assumes the subsistence of a "manifold structure" in the data. This assumption makes it find the manifolds in the noise of data. Since SC-RNA-seq suffers from a significant amount of noise, it is necessary to consider it before applying UMAP (McInnes et al., 2018).

Another method to perform dimensionality reduction is the linear discriminant analysis (LDA). LDA is a supervised dimensionality reduction method that tends to maximize the separability between the predetermined classes, using the covariance of "between-class" and "within-class." It first calculates the mean of the distances between the classes and then the mean of distances within the classes. The goal is to find a projection to maximize the ratio of between-class variability to the lower within-class variability (Tharwat et al., 2017; Qiao and Meister, 2020).

The SC-RNA-seq exhibits potential challenges similar to text mining, such as polysemy and synonymy, noise, and sparsity. Recently, a popular text mining technique, latent semantic analysis (LSA), has been used in SC-RNA-seq dimensionality reduction (Cheng et al., 2019). LSA at core uses a linear algebra-based method, called singular value decomposition (SVD), to cluster the semantically similar terms. SVD approximates a low-rank matrix to the given cell-gene matrix, such that the dimensions of the new matrix are much less than the original. This approximation is made by taking a combined product of the matrices of left-singular vector, right-singular vector, and the diagonal singular values.

## Differential Gene Expression Analysis

The expression of genes is stochastic in a cell; expression values thus observed are quite heterogeneous at the individual level among seemingly similar cells. The DGE analysis helps to understand the innate cellular processes and stochasticity of gene expressions (McDavid et al., 2013). The problem faced in DGE analysis is identifying genes that are largely expressed in a group of cells without any or no preliminary information of primary cell subtypes (Stegle et al., 2015). Additionally, gene expressions in individual cells show multimodality (Kippner et al., 2014). As expression variability of genes between cells of the same type indicates transcriptional heterogeneity (Johnson et al., 2015; Angermueller et al., 2016), it needs robust computational approaches to detect the true heterogeneity. In addition to multimodality, the sparsity due to—but not limited to—dropout events brings irregularities in the data, consequent of which the differential genes are difficult to detect. Various parametric as well as non-parametric approaches like Single-cell Differential Expression, Model-based Analysis of Single-cell Transcriptome (MAST), D3E, scDD, SigEMD, and DEsingle (Kharchenko et al., 2014; Finak et al., 2015; Delmans and Hemberg, 2016; Korthauer et al., 2016; Miao et al., 2018; Wang and Nabavi, 2018) have been developed/proposed for the DGE analysis in the SC-RNA-seq data. However, these tools try to manage either the gene dropouts or multimodality (Wang et al., 2019). For the subtle DGE analysis, these two crucial challenges need to be taken care of together.

## Cluster Analysis

Cluster analysis of SC-RNA-seq data is required to identify both known and unknown rare cell types (Menon, 2018). Along with the technical dropout events, the cells show a huge variation in gene expression levels even from the same set. As mentioned above, SC-RNA-seq suffers from massive inflation of zeros. There are three reasons for the observation of zeros in data: (1) the transcript was absent explicitly, hence a "true zero"; (2) the depth of sequencing was very low, and the transcript was present but not accounted for; and (3) at the time of library preparation, the transcript could not be captured or failed to amplify. The measurements from the latter two are considered to be the "false zeros." The concentration of too many zeros in the data brings in irregularities. These technical and biological factors lead to significant noise, due to which cluster analysis becomes challenging. For this, methods like Seurat, DropClust, and SCANPY (Satija et al., 2015; Ntranos et al., 2016; Yip et al., 2017; Sinha et al., 2018) have been proposed for clustering of SCs. There are certain limitations associated with these as well. Seurat and SCANPY work well with large datasets but underperforms when the dataset is smaller (Kiselev et al., 2019). The anticipated complexity in data and the rate of generation of SC data will be a challenge for all these tools. UMAP is yet another method for cluster identification of SC-RNA-seq data; however, as UMAP tends to preserve the local-topological structure, it is rather difficult to establish a relationship between clusters when the underlying cell subtypes are unknown.

In addition to the sparsity in data, SC-RNA-seq data suffer from a huge level of noise from faulty experimental designs usually referred to as "batch-effects." The noise in the data may contribute to the overfitting of the data. The overfitting can be avoided using regularization. Regularization is a process of restricting or reducing the features at the time of modeling.

So far, the clustering methods cluster the cells as per the transcription similarity, but the biological annotation of cell clusters remains a challenge. A possible solution could come from the generation of the data itself, as the more data are accumulated, the more can unknown clusters be matched with the previously known clusters. Another popular approach for cluster annotation is to use Gene Ontology (GO) analysis of the marker genes (Ashburner et al., 2000).

## Single-Cell Spatial Transcriptomics and RNA Velocity

Spatial transcriptomics (ST) gives measurement of gene expression changes with reference to geographical coordinates of the cells in tissues. It allows measurements of the transcripts with an advantage of conserving the spatial information, providing an additional analytical edge (Burgess, 2019). ST conform to *in situ* methods like seqFISH (Shah et al., 2016), seqFISH+ (Eng et al., 2019), FISSEQ (Fluorescence *in situ* Sequence) (Lee et al., 2015), MERFISH (Chen et al., 2015), and SC-RNA-seq-based methods like slide-seq (Rodriques et al., 2019) and Niche-seq (Medaglia et al., 2017). *In situ* labeling of the transcripts in tissues is advantageous for visualizing the location; however, a chance of molecular overcrowding results in fluorescence signal overlap. This overcrowding can be overcome by using SC spatial RNA-seq; however, the dissociation of cells prior to sequencing makes it difficult to link the transcriptomes back to their original locations (Burgess, 2019). These complementary strengths and limitations make it necessary to integrate the datasets generated by each technology.

In ST, a pair of images are generated, one containing whole tissue with fairly visible spots and the other having clearly visible fluorescence array spots (Wong et al., 2018). To leverage the ST, the image data from ST need to be integrated with the SC-RNA-seq data. As the principle challenges in both ST and SC-RNA-seq are the sparsity of the data and noise from technical and biological sources, an accurate data normalization and transformation is necessary before any downstream analysis (Wagner et al., 2016). Few tools have been developed to determine the cell types with respect to their spatial identities (Edsgärd et al., 2018; Svensson et al., 2018; Dries et al., 2019; Queen et al., 2019). These tools lack interactive processing of images and fails in providing a comprehensive three-dimensional view of the tissue. Recently, STUtility (Bergenstråhle et al., 2020b)—an R package using non-negative matrix factorization (NMF) for reducing the dimensions, spatial correlation (based on Pearson correlation), and K-means clustering—was found capable of providing a holistic view of the expression in tissues. SpatialCPie (Bergenstråhle et al., 2020a) is another easy-to-use R package that uses clustering at various resolutions to interactively uncover the gene expression patterns. Elosua-Bayes et al. (2021)

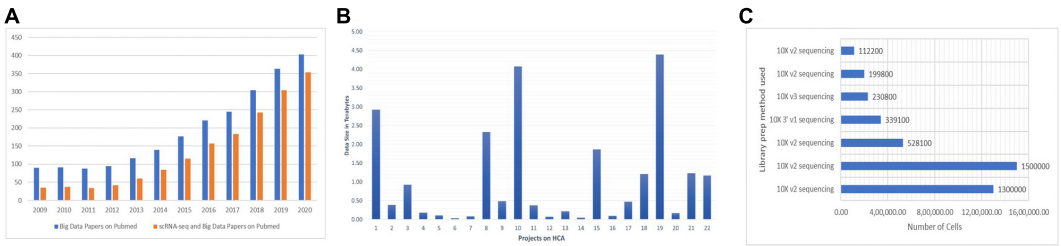



**FIGURE 2 | (A)** There is a steep rise every year for the publications of studies addressing the big data and SC-RNA-seq. For big data papers on PubMed, we used the query "*[big data (All Fields) AND MapReduce (All Fields) AND Hadoop (All fields)]*." For SC-RNA-seq and big data papers on PubMed, we used "*[(scRNA-seq OR Big Data) OR (Single-cell AND big data)]*." **(B,C)** Numbers were collected from the Human Cell Atlas Data portal of some exemplary projects.

developed SPOTlight, which uses NMF along with non-negative least squares (NNLS). NMF helps in dimensional reduction, followed by selection of marker genes using seurat package and then using NNLS to deconvolute each captured location (Elosua-Bayes et al., 2021).

The SC-RNA measurements have advanced our understanding of the intrinsic cellular functionalities; however, the destruction of cells in the process ceases the possibility of further resampling for an additional transcriptional state analysis. A new methodology, RNA velocity, is capable of deducing the future transcriptional state of a cell (La Manno et al., 2018). The idea behind the study is that the transcriptional upregulation of gene at a particular stage leads to the short-spanned abundance of unspliced transcripts. Similarly, the downregulation of the gene at a point of time results in a decrease of spliced transcripts. The ratio of this variation between unspliced and spliced transcripts is used to estimate the future state of a cell.

## Single-Cell Multi-omics and Data Integration

Biological activities in cells are perplexing, and the measurements of these processes show contrasting variation at temporal and histological levels. To comprehensively understand the intricate biological process of cells and organisms, it is necessary to investigate them at a multi-omics scale. Contingent upon the research question, SC experiments have flexed its reach to variety of layers, the majority of which include the following: (1) SCI-seq for Single-cell Genome Sequencing (Vitak et al., 2017), (2) scBS-seq for Single-cell DNA methylation (Smallwood et al., 2014), (3) scATAC-seq for Single-cell chromatin accessibility (Buenrostro et al., 2015), (4) CITE-seq for cell Surface Proteins (Stoeckius et al., 2017), (5) scCHIP-seq for Histone Modifications (Gomez et al., 2013), and (6) scGESTALT (Frieda et al., 2017) and MEMOIR (Raj et al., 2018) for chromosomal conformation. A universal challenge for all the SC technologies is that the measurements from a very low starting material led to generation of highly sparse and extremely noisy data. Hence, the integration of this data requires a statistically sound and robust computational framework. A primary challenge thereof remains to find an empirical strategy to normalize, batch-effect correction and linking the data from different sources so that the biological meaning and inference remain uncompromised.

For the integration and analysis of the SC multi-omics data, several methods developed for the variety of SC-mono-omics data have been fused or extended further to fulfill the requirement. However, each tool follows a different strategy for the analysis, which can be categorized as follows: (1) correlation and unsupervised cluster analysis; (2) data integration of different samples from a single measurement type and a single experiment type, e.g., SC-RNA-seq; (3) analysis and integration of data from different experiments and a single measurement type across different samples, e.g., sc-Spatial Transcriptomics; (4) integration of data from SC population, with more than one measurement type, different samples, and a single experiment; and (5) integration of data across multiple cells, multiple experiments, and multiple measurement types, e.g., combination of the SC-RNA-seq, scATAC, scCHIP-seq, CITE-seq, etc., of different cells collected at different time points (Stuart et al., 2019; Lähnemann et al., 2020; Lee et al., 2020).

Computational methods and tools for integration of biological data are evolving gradually. A number of techniques have been developed that have been discussed in section "Cluster Analysis." Seurat (Butler et al., 2018) is currently at the top of integrative analysis of SC multi-omics data, integrating the datasets based on the second principle. Along with Seurat, mutual nearest neighbor (MNN)-based method (Haghverdi et al., 2018) has been exploited to analyze the data combined on the basis of the second category. For the fourth category, analytical methods developed for bulk cellular analysis like MOFA (Argelaguet et al., 2018), MINT (Rohart et al., 2017a), mixOmics (Rohart et al., 2017b), and DIABLO (Singh et al., 2019) are being utilized. Cardelino (McCarthy et al., 2018), MATCHER (Welch et al., 2017), and cloealign (Campbell et al., 2019) are currently the tools used for integrative analysis under the fourth category. To our knowledge, there are no tools available for the last category.

## Big Data Pertaining to Single-Cell RNA Sequencing

The data-intensive scientific discoveries rely on three paradigms—theory, experimentation, and simulation modeling (Tolle et al., 2011). As big data is described with three characteristics (volume, velocity, and variety) (Stephens et al., 2015; Adil et al., 2016), data generated by SC-RNA-seq are tantamount to these three quantitative characteristics

(Ivanov et al., 2013). With the introduction of new methods in microfluidics (Zare and Kim, 2010), combinatorial indexing procedures (Fan et al., 2015), and rapid drop in the sequencing cost, SC assay profiling has widely become a routine practice among biologists for analyzing millions of cells in hours, paving the way for the accumulation of a large amount of data. The most popular next-generation sequencing platform, Illumina HiSeq, results in the accumulation of around 100 gigabytes of raw RNA-seq data per study. It usually takes hours to align these raw data to their reference genome. SC experiments generating petabytes of data on a variety of layers contribute to the big data paradigm. A human genome has 20,000–25,000 genes composed of 3 million base pairs, totaling to 100 gigabytes of data, equivalent to 102,400 photos[1]; it is expected that more or less "25 petabytes" of genomic data will be generated annually around the globe by the year 2030 (Khoury et al., 2020). It is anticipated that human genomic data can potentially overtake the data produced by online social networks (Check Hayden, 2015). The Human Cell Atlas (HCA)—a project to prepare a reference map of each cell in the human body at various stages, will accumulate a massive amount of data by the end of its completion (Regev et al., 2017). There is a need for comprehensive integration of big data and SC-RNA-seq technologies. A large number of publications on SC-RNA and big data have emerged lately (**Figure 2A**). The datasets of 4.5 million cells are already published in Data[2], the largest of which contains more than 1.5 million $CD34^+$ hematopoietic cells of human bone marrow (Setty et al., 2019) and 1.3 million transcriptomes of mouse brain cells (**Figures 2B,C**).

Consequently, the data acquired from these experiments constitute a data revolution in the field of SC biology (Lähnemann et al., 2019). As SC-RNA-seq data have a greater potential of uncovering the hidden patterns at the molecular level, the data pertaining to it thus require an extremely parallel, scalable, and statistically sound computational framework as its handling tools. Big data technologies like Apache's Hadoop (Taylor, 2010; O'Driscoll et al., 2013) and Spark (Zaharia et al., 2016; Guo et al., 2018) embody the required computational parallelism and data distribution mechanisms. Hadoop uses MapReduce technology for parallel and scalable processing (Dean and Ghemawat, 2008) to disintegrate the larger problems into smaller subproblems on a distributed file system called Hadoop Distributed File System (HDFS). Incorporating big data technologies in the analysis of rapidly increasing SC genomics data will help in transforming and processing it with limitless scalability and fault tolerance at a very low cost.

[1] https://www.experfy.com/blog/intersection-genomics-big-data

[2] https://data.humancellatlas.org/

## CONCLUSION AND FUTURE PERSPECTIVE

As a consequence of meager RNA capture rate, low starting materials, and challenging experimental protocols, the SC-RNA-seq faces computational and analytical challenges. The noise and sparsity due to the technical (dropout events) and biological factors make the downstream analysis of SC-RNA-seq data a complicated task. Additionally, the rapidity in the development of new and exciting experimental methods for SC-RNA-seq is paving the way for a large accumulation of data. This large agglomeration of data is nothing but the genomic face of "big data." These two challenges together give rise to a new paradigm of Big Single-Cell Data Science. Although a plethora of algorithms and computational tools have already been developed, it is essential to address these challenges collectively and produce a robust, accurate, parallel, and scalable framework.

## AUTHOR CONTRIBUTIONS

MA and ATJ conceived the idea, edited the manuscript, and contributed to the compilation of data for designing of figures. AA, VK, and ATJ contributed to the writing of the manuscript. All authors contributed to the article and approved the submitted version.

## FUNDING

ATJ is grateful to DST-SERB for financial support (CRG/2019/004106) that helped in to establishing the infrastructural facilities.

## ACKNOWLEDGMENTS

The authors would like to thank their colleagues for the help in improving the contents of the manuscript.

## REFERENCES

Adil, A., Kar, H. A., Jangir, R., and Sofi, S. A. (2016). "Analysis of multi-diseases using big data for improvement in healthcare," in *Proceedings of the 2015 IEEE UP Section Conference on Electrical Computer and Electronics, UPCON 2015*, Allahabad. doi: 10.1109/UPCON.2015.7456696

AlJanahi, A. A., Danielsen, M., and Dunbar, C. E. (2018). An introduction to the analysis of single-cell RNA-sequencing data. *Mol. Ther. Methods Clin. Dev.* 10, 189–196. doi: 10.1016/j.omtm.2018.07.003

Altschuler, S. J., and Wu, L. F. (2010). Cellular heterogeneity: do differences make a difference? *Cell* 141, 559–563. doi: 10.1016/j.cell.2010.04.033

Andrews, T. S., and Hemberg, M. (2019). M3Drop: dropout-based feature selection for scRNASeq. *Bioinformatics (Oxford, England)* 35, 2865–2867. doi: 10.1093/bioinformatics/bty1044

Angerer, P., Simon, L., Tritschler, S., Wolf, F. A., Fischer, D., and Theis, F. J. (2017). Single cells make big data: new challenges and opportunities in transcriptomics. *Curr. Opin. Syst. Biol.* 4, 85–91. doi: 10.1016/j.coisb.2017.07.004

Angermueller, C., Clark, S. J., Lee, H. J., Macaulay, I. C., Teng, M. J., Hu, T. X., et al. (2016). Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity. *Nat. Methods* 13, 229–232. doi: 10.1038/nmeth.3728

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.* 14:8124. doi: 10.15252/msb.20178124

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., et al. (2000). Gene ontology: tool for the unification of biology. *Nat. Genet.* 25, 25–29. doi: 10.1038/75556

Bacher, R., Chu, L. F., Leng, N., Gasch, A. P., Thomson, J. A., Stewart, R. M., et al. (2017). SCnorm: robust normalization of single-cell RNA-seq data. *Nat. Methods* 14, 584–586. doi: 10.1038/nmeth.4263

Beakke, M. K. (1951). Density gradient centrifugation: a new separation technique. *J. Am. Chem. Soc.* 73, 1847–1848. doi: 10.1021/ja01148a508

Becht, E., Dutertre, C.-A., Kwok, I., Ng, L. G., Ginhoux, F., and Newell, E. (2018). Evaluation of UMAP as an alternative to t-SNE for single-cell data. *bioRxiv* [Preprint]. doi: 10.1101/298430

Becht, E., McInnes, L., Healy, J., Dutertre, C. A., Kwok, I. W. H., Ng, L. G., et al. (2019). Dimensionality reduction for visualizing single-cell data using UMAP. *Nat. Biotechnol.* 37, 38–44. doi: 10.1038/nbt.4314

Bergenstråhle, J., Bergenstråhle, L., and Lundeberg, J. (2020a). SpatialCPie: an R/Bioconductor package for spatial transcriptomics cluster evaluation. *BMC Bioinform.* 21:161. doi: 10.1186/s12859-020-3489-7

Bergenstråhle, J., Larsson, L., and Lundeberg, J. (2020b). Seamless integration of image and molecular analysis for spatial transcriptomics workflows. *BMC Genomics* 21:482. doi: 10.1186/s12864-020-06832-3

Bray, N. L., Pimentel, H., Melsted, P., and Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* 34, 525–527. doi: 10.1038/nbt.3519

Brennecke, P., Anders, S., Kim, J. K., Kołodziejczyk, A. A., Zhang, X., Proserpio, V., et al. (2013). Accounting for technical noise in single-cell RNA-seq experiments. *Nat. Methods* 10, 1093–1098. doi: 10.1038/nmeth.2645

Buenrostro, J. D., Wu, B., Litzenburger, U. M., Ruff, D., Gonzales, M. L., Snyder, M. P., et al. (2015). Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* 523, 486–490. doi: 10.1038/nature14590

Bullard, J. H., Purdom, E., Hansen, K. D., and Dudoit, S. (2010). Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics* 11:94. doi: 10.1186/1471-2105-11-94

Burgess, D. J. (2019). Spatial transcriptomics coming of age. *Nat. Rev. Genet.* 20:317. doi: 10.1038/s41576-019-0129-z

Butler, A., Hoffman, P., Smibert, P., Papalexi, E., and Satija, R. (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* 36, 411–420. doi: 10.1038/nbt.4096

Campbell, K. R., Steif, A., Laks, E., Zahn, H., Lai, D., McPherson, A., et al. (2019). Clonealign: statistical integration of independent single-cell RNA and DNA sequencing data from human cancers. *Genome Biol.* 20:54. doi: 10.1186/s13059-019-1645-z

Check Hayden, E. (2015). Genome researchers raise alarm over big data. *Nature* 312–314. doi: 10.1038/nature.2015.17912

Chen, G., Ning, B., and Shi, T. (2019). Single-cell RNA-seq technologies and related computational data analysis. *Front. Genet.* 10:317. doi: 10.3389/fgene.2019.00317

Chen, G., Wang, C., and Shi, T. L. (2011). Overview of available methods for diverse RNA-Seq data analyses. *Sci. China Life Sci.* 54, 1121–1128. doi: 10.1007/s11427-011-4255-x

Chen, K. H., Boettiger, A. N., Moffitt, J. R., Wang, S., and Zhuang, X. (2015). Spatially resolved, highly multiplexed RNA profiling in single cells. *Science* 348:6090. doi: 10.1126/science.aaa6090

Cheng, C., Easton, J., Rosencrance, C., Li, Y., Ju, B., Williams, J., et al. (2019). Latent cellular analysis robustly reveals subtle diversity in large-scale single-cell RNA-seq data. *Nucleic Acids Res.* 47:e143. doi: 10.1093/nar/gkz826

Citri, A., Pang, Z. P., Südhof, T. C., Wernig, M., and Malenka, R. C. (2012). Comprehensive qPCR profiling of gene expression in single neuronal cells. *Nat. Protoc.* 7, 118–127. doi: 10.1038/nprot.2011.430

Costa, F. F. (2012). Big data in genomics: challenges and solutions. *G.I.T. Lab. J.* 1–4.

Dean, J., and Ghemawat, S. (2008). MapReduce: simplified data processing on large clusters. *Commun. ACM* 51, 107–113. doi: 10.1145/1327452.1327492

Delmans, M., and Hemberg, M. (2016). Discrete distributional differential expression (D3E) - a tool for gene expression analysis of single-cell RNA-seq data. *BMC Bioinform.* 17:110. doi: 10.1186/s12859-016-0944-6

Ding, B., Zheng, L., Zhu, Y., Li, N., Jia, H., Ai, R., et al. (2015). Normalization and noise reduction for single cell RNA-seq experiments. *Bioinformatics* 31, 2225–2227. doi: 10.1093/bioinformatics/btv122

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., et al. (2013). STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* 29, 15–21. doi: 10.1093/bioinformatics/bts635

Dries, R., Zhu, Q., Eng, C. H. L., Sarkar, A., Bao, F., George, R. E., et al. (2019). Giotto, a pipeline for integrative analysis and visualization of single-cell spatial transcriptomic data. *bioRxiv* [Preprint]. doi: 10.1101/701680

Edsgärd, D., Johnsson, P., and Sandberg, R. (2018). Identification of spatial expression trends in single-cell gene expression data. *Nat. Methods* 15, 339–342. doi: 10.1038/nmeth.4634

Elosua-Bayes, M., Nieto, P., Mereu, E., Gut, I., and Heyn, H. (2021). SPOTlight: seeded NMF regression to deconvolute spatial transcriptomics spots with single-cell transcriptomes. *Nucleic Acids Res.* gkab043. doi: 10.1093/nar/gkab043

Eng, C. H. L., Lawson, M., Zhu, Q., Dries, R., Koulena, N., Takei, Y., et al. (2019). Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH+. *Nature* 568:235. doi: 10.1038/s41586-019-1049-y

Espina, V., Wulfkuhle, J. D., Calvert, V. S., VanMeter, A., Zhou, W., Coukos, G., et al. (2006). Laser-capture microdissection. *Nat. Protoc.* 1, 586–603. doi: 10.1038/nprot.2006.85

Fan, H. C., Fu, G. K., and Fodor, S. P. A. (2015). Combinatorial labeling of single cells for gene expression cytometry. *Science* 347:1258367. doi: 10.1126/science.1258367

Finak, G., McDavid, A., Yajima, M., Deng, J., Gersuk, V., Shalek, A. K., et al. (2015). MAST: A flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol.* 16:278. doi: 10.1186/s13059-015-0844-5

Fonseca, N. A., Rung, J., Brazma, A., and Marioni, J. C. (2012). Tools for mapping high-throughput sequencing data. *Bioinformatics* 28, 3169–3177. doi: 10.1093/bioinformatics/bts605

Frieda, K. L., Linton, J. M., Hormoz, S., Choi, J., Chow, K. H. K., Singer, Z. S., et al. (2017). Synthetic recording and in situ readout of lineage information in single cells. *Nature* 541, 59–64. doi: 10.1038/nature20777

Garber, M., Grabherr, M. G., Guttman, M., and Trapnell, C. (2011). Computational methods for transcriptome annotation and quantification using RNA-seq. *Nat. Methods* 8, 469–477. doi: 10.1038/nmeth.1613

Gomez, D., Shankman, L. S., Nguyen, A. T., and Owens, G. K. (2013). Detection of histone modifications at specific gene loci in single cells in histological sections. *Nat. Methods* 10, 171–177. doi: 10.1038/nmeth.2332

Gross, A., Schoendube, J., Zimmermann, S., Steeb, M., Zengerle, R., and Koltay, P. (2015). Technologies for single-cell isolation. *Int. J. Mol. Sci.* 16, 16897–16919. doi: 10.3390/ijms160816897

Guo, R., Zhao, Y., Zou, Q., Fang, X., and Peng, S. (2018). Bioinformatics applications on apache spark. *GigaScience* 7:giy098. doi: 10.1093/gigascience/giy098

Gupta, I., Collier, P. G., Haase, B., Mahfouz, A., Joglekar, A., Floyd, T., et al. (2018). Single-cell isoform RNA sequencing characterizes isoforms in thousands of cerebellar cells. *Nat. Biotechnol.* 36, 1197–1202. doi: 10.1038/nbt.4259

Haghverdi, L., Lun, A. T. L., Morgan, M. D., and Marioni, J. C. (2018). Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat. Biotechnol.* 36, 421–427. doi: 10.1038/nbt.4091

Haque, A., Engel, J., Teichmann, S. A., and Lönnberg, T. (2017). A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. *Genome Med.* 9, 1–12. doi: 10.1186/s13073-017-0467-4

Hashimshony, T., Senderovich, N., Avital, G., Klochendler, A., de Leeuw, Y., Anavy, L., et al. (2016). CEL-Seq2: Sensitive highly-multiplexed single-cell RNA-Seq. *Genome Biol.* 17:77. doi: 10.1186/s13059-016-0938-8

He, K. Y., Ge, D., and He, M. M. (2017). Big data analytics for genomic medicine. *Int. J. Mol. Sci.* 18, 1–18. doi: 10.3390/ijms18020412

Herzenberg, L. A., Parks, D., Sahaf, B., Perez, O., Roederer, M., and Herzenberg, L. A. (2002). The history and future of the fluorescence activated cell sorter and flow cytometry: a view from Stanford. *Clin. Chem.* 48, 1819–1827.

Hu, P., Zhang, W., Xin, H., and Deng, G. (2016). Single cell isolation and analysis. *Front. Cell Dev. Biol.* 4:116. doi: 10.3389/fcell.2016.00116

Huang, Y., and Sanguinetti, G. (2017). BRIE: transcriptome-wide splicing quantification in single cells. *Genome Biol.* 18:123. doi: 10.1186/s13059-017-1248-5

Hwang, B., Lee, J. H., and Bang, D. (2018). Single-cell RNA sequencing technologies and bioinformatics pipelines. *Exp. Mol. Med.* 50, 1–14. doi: 10.1038/s12276-018-0071-8

Ilicic, T., Kim, J. K., Kolodziejczyk, A. A., Bagger, F. O., McCarthy, D. J., Marioni, J. C., et al. (2016). Classification of low quality cells from single-cell RNA-seq data. *Genome Biol.* 17:29. doi: 10.1186/s13059-016-0888-1

Islam, S., Zeisel, A., Joost, S., La Manno, G., Zajac, P., Kasper, M., et al. (2014). Quantitative single-cell RNA-seq with unique molecular identifiers. *Nat. Methods* 11, 163–166. doi: 10.1038/nmeth.2772

Ivanov, T., Korfiatis, N., and Zicari, R. V. (2013). *On the Inequality of the 3V's of Big Data Architectural Paradigms: A Case For Heterogeneity*. Available online at: https://arxiv.org/abs/1311.0805

Jaitin, D. A., Kenigsberg, E., Keren-Shaul, H., Elefant, N., Paul, F., Zaretsky, I., et al. (2014). Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. *Science* 343, 776–779. doi: 10.1126/science.1247651

Johnson, M. B., Wang, P. P., Atabay, K. D., Murphy, E. A., Doan, R. N., Hecht, J. L., et al. (2015). Single-cell analysis reveals transcriptional heterogeneity of neural progenitors in human cortex. *Nat. Neurosci.* 18, 637–646. doi: 10.1038/nn.3980

Kharchenko, P. V., Silberstein, L., and Scadden, D. T. (2014). Bayesian approach to single-cell differential expression analysis. *Nat. Methods* 11, 740–742. doi: 10.1038/nmeth.2967

Khoury, M. J., Armstrong, G. L., Bunnell, R. E., Cyril, J., and Iademarco, M. F. (2020). The intersection of genomics and big data with public health: opportunities for precision public health. *PLoS Med.* 17:e1003373. doi: 10.1371/journal.pmed.1003373

Kim, D., Paggi, J. M., Park, C., Bennett, C., and Salzberg, S. L. (2019). Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* 37, 907–915. doi: 10.1038/s41587-019-0201-4

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S. L. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* 14:R36. doi: 10.1186/gb-2013-14-4-r36

Kippner, L. E., Kim, J., Gibson, G., and Kemp, M. L. (2014). Ingle cell transcriptional analysis reveals novel innate immune cell types. *PeerJ* 2:e452. doi: 10.7717/peerj.452

Kiselev, V. Y., Andrews, T. S., and Hemberg, M. (2019). Challenges in unsupervised clustering of single-cell RNA-seq data. *Nat. Rev. Genet.* 20, 273–282. doi: 10.1038/s41576-018-0088-9

Kivioja, T., Vähärautio, A., Karlsson, K., Bonke, M., Enge, M., Linnarsson, S., et al. (2012). Counting absolute numbers of molecules using unique molecular identifiers. *Nat. Methods* 9, 72–74. doi: 10.1038/nmeth.1778

Klein, A. M., Mazutis, L., Akartuna, I., Tallapragada, N., Veres, A., Li, V., et al. (2015). Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* 161, 1187–1201. doi: 10.1016/j.cell.2015.04.044

Kobak, D., and Berens, P. (2019). The art of using t-SNE for single-cell transcriptomics. *Nat. Commun.* 10:5416. doi: 10.1038/s41467-019-13056-x

Kolodziejczyk, A. A., Kim, J. K., Svensson, V., Marioni, J. C., and Teichmann, S. A. (2015). The technology and biology of single-cell RNA sequencing. *Mol. Cell* 58, 610–620. doi: 10.1016/j.molcel.2015.04.005

Korthauer, K. D., Chu, L. F., Newton, M. A., Li, Y., Thomson, J., Stewart, R., et al. (2016). A statistical approach for identifying differential distributions in single-cell RNA-seq experiments. *Genome Biol.* 17:222. doi: 10.1186/s13059-016-1077-y

Kulkarni, A., Anderson, A. G., Merullo, D. P., and Konopka, G. (2019). Beyond bulk: a review of single cell transcriptomics methodologies and applications. *Curr. Opin. Biotechnol.* 58, 129–136. doi: 10.1016/j.copbio.2019.03.001

La Manno, G., Soldatov, R., Zeisel, A., Braun, E., Hochgerner, H., Petukhov, V., et al. (2018). RNA velocity of single cells. *Nature* 560, 494–498. doi: 10.1038/s41586-018-0414-6

Lähnemann, D., Köster, J., Szczurek, E., Mccarthy, D. J., Hicks, S. C., Mark, D., et al. (2019). 12 grand challenges in single-cell data science. *PeerJ* 7:e27885v3. doi: 10.7287/peerj.preprints.27885v2

Lähnemann, D., Köster, J., Szczurek, E., McCarthy, D. J., Hicks, S. C., Robinson, M. D., et al. (2020). Eleven grand challenges in single-cell data science. *Genome Biol.* 21:31. doi: 10.1186/s13059-020-1926-6

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* 10:R25. doi: 10.1186/gb-2009-10-3-r25

Lebrigand, K., Magnone, V., Barbry, P., and Waldmann, R. (2020). High throughput error corrected Nanopore single cell transcriptome sequencing. *Nat. Commun.* 11, 1–8. doi: 10.1038/s41467-020-17800-6

Lee, J., Hyeon, D. Y., and Hwang, D. (2020). Single-cell multiomics: technologies and data analysis methods. *Exp. Mol. Med.* 52, 1428–1442. doi: 10.1038/s12276-020-0420-2

Lee, J. H., Daugharthy, E. R., Scheiman, J., Kalhor, R., Ferrante, T. C., Terry, R., et al. (2015). Fluorescent in situ sequencing (FISSEQ) of RNA for gene expression profiling in intact cells and tissues. *Nat. Protoc.* 10, 442–458. doi: 10.1038/nprot.2014.191

Li, B., and Dewey, C. N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinform.* 12:323. doi: 10.1186/1471-2105-12-323

Li, H., and Homer, N. (2010). A survey of sequence alignment algorithms for next-generation sequencing. *Brief. Bioinform.* 11, 473–483. doi: 10.1093/bib/bbq015

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 15:550. doi: 10.1186/s13059-014-0550-8

Lun, A. T. L., Bach, K., and Marioni, J. C. (2016). Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome Biol.* 17:75. doi: 10.1186/s13059-016-0947-7

Macaulay, I. C., Ponting, C. P., and Voet, T. (2017). Single-cell multiomics: multiple measurements from single cells. *Trends Genet.* 33, 155–168. doi: 10.1016/j.tig.2016.12.003

Macosko, E. Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., et al. (2015). Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* 161, 1202–1214. doi: 10.1016/j.cell.2015.05.002

McCarthy, D. J., Rostom, R., Huang, Y., Kunz, D. J., Danecek, P., Bonder, M. J., et al. (2018). Cardelino: integrating whole exomes and single-cell transcriptomes to reveal phenotypic impact of somatic variants. *bioRxiv* [Preprint]. doi: 10.1101/413047

McDavid, A., Finak, G., Chattopadyay, P. K., Dominguez, M., Lamoreaux, L., Ma, S. S., et al. (2013). Data exploration, quality control and testing in single-cell qPCR-based gene expression experiments. *Bioinformatics* 29, 461–467. doi: 10.1093/bioinformatics/bts714

McGann, L. E., Yang, H. Y., and Walterson, M. (1988). Manifestations of cell damage after freezing and thawing. *Cryobiology* 25, 178–185. doi: 10.1016/0011-2240(88)90024-7

McInnes, L., Healy, J., Saul, N., and Großberger, L. (2018). UMAP: uniform manifold approximation and projection. *J. Open Source Softw.* 3:861. doi: 10.21105/joss.00861

Medaglia, C., Giladi, A., Stoler-Barak, L., De Giovanni, M., Salame, T. M., Biram, A., et al. (2017). Spatial reconstruction of immune niches by combining photoactivatable reporters and scRNA-seq. *Science* 358, 1622–1626. doi: 10.1126/science.aao4277

Menon, V. (2018). Clustering single cells: a review of approaches on high-and low-depth single-cell RNA-seq data. *Brief. Funct. Genomics* 18:434. doi: 10.1093/bfgp/ely001

Miao, Z., Deng, K., Wang, X., and Zhang, X. (2018). DEsingle for detecting three types of differential expression in single-cell RNA-seq data. *Bioinformatics (Oxford, England)* 34, 3223–3224. doi: 10.1093/bioinformatics/bty332

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* 5, 621–628. doi: 10.1038/nmeth.1226

Natarajan, K. N., Miao, Z., Jiang, M., Huang, X., Zhou, H., Xie, J., et al. (2019). Comparative analysis of sequencing technologies for single-cell transcriptomics. *Genome Biol.* 20:70. doi: 10.1186/s13059-019-1676-5

Ntranos, V., Kamath, G. M., Zhang, J. M., Pachter, L., and Tse, D. N. (2016). Fast and accurate single-cell RNA-seq analysis by clustering of transcript-compatibility counts. *Genome Biol.* 17, 1–14. doi: 10.1186/s13059-016-0970-8

O'Driscoll, A., Daugelaite, J., and Sleator, R. D. (2013). Big data", Hadoop and cloud computing in genomics. *J. Biomed. Inform.* 46, 774–781. doi: 10.1016/j.jbi.2013.07.001

Olsen, T. K., and Baryawno, N. (2018). Introduction to single-cell RNA sequencing. *Curr. Protoc. Mol. Biol.* 122:57. doi: 10.1002/cpmb.57

Ozsolak, F., and Milos, P. M. (2011). RNA sequencing: Advances, challenges and opportunities. *Nat. Rev. Genet.* 12, 87–98. doi: 10.1038/nrg2934

Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., and Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* 14, 417–419. doi: 10.1038/nmeth.4197

Patro, R., Mount, S. M., and Kingsford, C. (2014). Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. *Nat. Biotechnol.* 32, 462–464. doi: 10.1038/nbt.2862

Pertea, M., Pertea, G. M., Antonescu, C. M., Chang, T. C., Mendell, J. T., and Salzberg, S. L. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* 33, 290–295. doi: 10.1038/nbt.3122

Pezzotti, N., Lelieveldt, B. P. F., Van Der Maaten, L., Höllt, T., Eisemann, E., and Vilanova, A. (2017). Approximated and user steerable tSNE for progressive visual analytics. *IEEE Trans. Visualization Comp. Graphics* 23, 1739–1752. doi: 10.1109/TVCG.2016.2570755

Phipson, B., Zappia, L., and Oshlack, A. (2017). Gene length and detection bias in single cell RNA sequencing protocols. *F1000Research* 6:595. doi: 10.12688/f1000research.11290.1

Picelli, S., Faridani, O. R., Björklund, ÅK., Winberg, G., Sagasser, S., and Sandberg, R. (2014). Full-length RNA-seq from single cells using Smart-seq2. *Nat. Protoc.* 9, 171–181. doi: 10.1038/nprot.2014.006

Qiao, M., and Meister, M. (2020). *Factorized Linear Discriminant Analysis for Phenotype-Guided Representation Learning of Neuronal Gene Expression Data*. Available online at: https://arxiv.org/abs/2010.02171v4

Queen, R., Cheung, K., Lisgo, S., Coxhead, J., and Cockell, S. (2019). Spaniel: analysis and interactive sharing of spatial transcriptomics data. *bioRxiv* [Preprint]. doi: 10.1101/619197

Raj, B., Wagner, D. E., McKenna, A., Pandey, S., Klein, A. M., Shendure, J., et al. (2018). Simultaneous single-cell profiling of lineages and cell types in the vertebrate brain. *Nat. Biotechnol.* 36, 442–450. doi: 10.1038/nbt.4103

Ramskold, D., Luo, S., Wang, Y., Li, R., Deng, Q., Omid, R., et al. (2013). Full-Length mRNA-Seq from single Cell levels of RNA and individual circulating tumor cells. *Nat. Biotechnol.* 30, 777–782. doi: 10.1038/nbt.2282.Full-Length

Regev, A., Teichmann, S., Lander, E., Amit, I., Benoist, C., Birney, E., et al. (2017). Science forum: the human cell atlas. *eLife* 6:e27041.

Risso, D., Ngai, J., Speed, T. P., and Dudoit, S. (2014). Normalization of RNA-seq data using factor analysis of control genes or samples. *Nat. Biotechnol.* 32, 896–902. doi: 10.1038/nbt.2931

Rodriques, S. G., Stickels, R. R., Goeva, A., Martin, C. A., Murray, E., Vanderburg, C. R., et al. (2019). Slide-seq: a scalable technology for measuring genome-wide expression at high spatial resolution. *Science* 363, 1463–1467. doi: 10.1126/science.aaw1219

Rohart, F., Eslami, A., Matigian, N., Bougeard, S., and Lê Cao, K. A. (2017a). MINT: a multivariate integrative method to identify reproducible molecular signatures across independent experiments and platforms. *BMC Bioinform.* 18:128. doi: 10.1186/s12859-017-1553-8

Rohart, F., Gautier, B., Singh, A., and Lê Cao, K. A. (2017b). mixOmics: an R package for 'omics feature selection and multiple data integration. *PLoS Comput. Biol.* 13:1005752. doi: 10.1371/journal.pcbi.1005752

Saliba, A. E., Westermann, A. J., Gorski, S. A., and Vogel, J. (2014). Single-cell RNA-seq: Advances and future challenges. *Nucleic Acids Res.* 42, 8845–8860. doi: 10.1093/nar/gku555

Satija, R., Farrell, J. A., Gennert, D., Schier, A. F., and Regev, A. (2015). Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.* 33, 495–502. doi: 10.1038/nbt.3192

Schmitz, B., Radbruch, A., Kümmel, T., Wickenhauser, C., Korb, H., Hansmann, M. L., et al. (1994). Magnetic activated cell sorting (MACS) - a new imrnunomagnetic method for megakarvocvtic cell isolation. *Eur. J. Heamatol.* 52, 267–275.

Sena, J. A., Galotto, G., Devitt, N. P., Connick, M. C., Jacobi, J. L., Umale, P. E., et al. (2018). Unique Molecular Identifiers reveal a novel sequencing artefact with implications for RNA-Seq based gene expression analysis. *Sci. Rep.* 8:13121. doi: 10.1038/s41598-018-31064-7

Sengupta, D., Rayan, N. A., Lim, M., Lim, B., and Prabhakar, S. (2016). Fast, scalable and accurate differential expression analysis for single cells. *bioRxiv* [Preprint]. doi: 10.1101/049734

Setty, M., Kiseliovas, V., Levine, J., Gayoso, A., Mazutis, L., and Pe'er, D. (2019). Characterization of cell fate probabilities in single-cell data with Palantir. *Nat. Biotechnol.* 37, 451–460. doi: 10.1038/s41587-019-0068-4

Shah, S., Lubeck, E., Zhou, W., and Cai, L. (2016). In situ transcription profiling of single cells reveals spatial organization of cells in the mouse hippocampus. *Neuron* 92, 342–357. doi: 10.1016/j.neuron.2016.10.001

Sheng, K., and Zong, C. (2019). Single-cell RNA-Seq by multiple annealing and tailing-based quantitative single-cell RNA-Seq (MATQ-Seq). *Methods Mol. Biol.* 1979, 57–71. doi: 10.1007/978-1-4939-9240-9_5

Singh, A., Shannon, C. P., Gautier, B., Rohart, F., Vacher, M., Tebbutt, S. J., et al. (2019). DIABLO: an integrative approach for identifying key molecular drivers from multi-omics assays. *Bioinformatics* 35, 3055–3062. doi: 10.1093/bioinformatics/bty1054

Sinha, D., Kumar, A., Kumar, H., Bandyopadhyay, S., and Sengupta, D. (2018). Dropclust: Efficient clustering of ultra-large scRNA-seq data. *Nucleic Acids Res.* 46:e36. doi: 10.1093/nar/gky007

Sirén, J., Välimäki, N., and Mäkinen, V. (2014). HISAT2 - fast and sensitive alignment against general human population. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 11, 375–388. doi: 10.1109/TCBB.2013.2297101

Smallwood, S. A., Lee, H. J., Angermueller, C., Krueger, F., Saadeh, H., Peat, J., et al. (2014). Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. *Nat. Methods* 11, 817–820. doi: 10.1038/nmeth.3035

Smith, T., Heger, A., and Sudbery, I. (2017). UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. *Genome Res.* 27, 491–499. doi: 10.1101/gr.209601.116

Song, Y., Xu, X., Wang, W., Tian, T., Zhu, Z., and Yang, C. (2019). Single cell transcriptomics: Moving towards multi-omics. *Analyst* 144, 3172–3189. doi: 10.1039/c8an01852a

Stegle, O., Teichmann, S. A., and Marioni, J. C. (2015). Computational and analytical challenges in single-cell transcriptomics. *Nat. Rev. Genet.* 16, 133–145. doi: 10.1038/nrg3833

Stephens, Z. D., Lee, S. Y., Faghri, F., Campbell, R. H., Zhai, C., Efron, M. J., et al. (2015). Big data: astronomical or genomical? *PLoS Biol.* 13:e1002195. doi: 10.1371/journal.pbio.1002195

Stoeckius, M., Hafemeister, C., Stephenson, W., Houck-Loomis, B., Chattopadhyay, P. K., Swerdlow, H., et al. (2017). Simultaneous epitope and transcriptome measurement in single cells. *Nat. Methods* 9:2579.

Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W. M., et al. (2019). Comprehensive integration of single-cell data. *Cell* 177, 1888–1902.e21. doi: 10.1016/j.cell.2019.05.031

Svensson, V., Teichmann, S. A., and Stegle, O. (2018). SpatialDE: Identification of spatially variable genes. *Nat. Methods* 15, 343–346. doi: 10.1038/nmeth.4636

Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., et al. (2009). mRNA-Seq whole-transcriptome analysis of a single cell. *Nat. Methods* 6, 377–382. doi: 10.1038/nmeth.1315

Taylor, R. C. (2010). An overview of the Hadoop/MapReduce/HBase framework and its current applications in bioinformatics. *BMC Bioinform.* 11:S1. doi: 10.1186/1471-2105-11-S12-S1

Tharwat, A., Gaber, T., Ibrahim, A., and Hassanien, A. E. (2017). Linear discriminant analysis: a detailed tutorial. *AI Commun.* 30, 169–190. doi: 10.3233/AIC-170729

Tolle, K. M., Tansley, D. S. W., and Hey, A. J. G. (2011). The fourth Paradigm: Data-intensive scientific discovery. *Proc. IEEE* 99, 1334–1337. doi: 10.1109/JPROC.2011.2155130

Tomlinson, M. J., Tomlinson, S., Yang, X. B., and Kirkham, J. (2013). Cell separation: Terminology and practical considerations. *J. Tissue Eng.* 4, 1–14. doi: 10.1177/2041731412472690

Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., et al. (2014). The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.* 32, 381–386. doi: 10.1038/nbt.2859

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

Trombetta, J., Gennert, D., Lu, D., and Sattija, R. (2015). Preparation of single-cell RNA-seq libraries for NGS. *Curr. Protoc. Mol. Biol.* 19, 161–169. doi: 10.3851/IMP2701.Changes

Vallejos, C. A., Marioni, J. C., and Richardson, S. (2015). BASiCS: Bayesian analysis of single-cell sequencing data. *PLoS Comput. Biol.* 11:e1004333. doi: 10.1371/journal.pcbi.1004333

Vallejos, C. A., Risso, D., Scialdone, A., Dudoit, S., and Marioni, J. C. (2017). Normalizing single-cell RNA sequencing data: challenges and opportunities. *Nat. Methods* 14, 565–571. doi: 10.1038/nmeth.4292.Normalizing

Van Der Maaten, L. J. P., and Hinton, G. E. (2008). Visualizing high-dimensional data using t-sne. *J. Machine Learn. Res.* 9, 2579–2605.

Van Der Maaten, L. J. P., Postma, E. O., and Van Den Herik, H. J. (2009). "Dimensionality reduction: a comparative review," in *Technical Report TiCC-TR 2009-005* (Tilburg: Tillburg University).

Vitak, S. A., Torkenczy, K. A., Rosenkrantz, J. L., Fields, A. J., Christiansen, L., Wong, M. H., et al. (2017). Sequencing thousands of single-cell genomes with combinatorial indexing. *Nat. Methods* 472, 90–94. doi: 10.1038/nmeth.4154

Volden, R., and Vollmers, C. (2020). Highly multiplexed single-cell full-length cDNA Sequencing of human immune cells with 10X genomics and R2C2. *bioRxiv* [Preprint]. doi: 10.1101/2020.01.10.902361

Wagner, A., Regev, A., and Yosef, N. (2016). Revealing the vectors of cellular identity with single-cell genomics. *Nat. Biotechnol.* 34, 1145–1160. doi: 10.1038/nbt.3711

Wang, D., and Bodovitz, S. (2010). Single cell analysis: the new frontier in "omics.". *Trends Biotechnol.* 28, 281–290. doi: 10.1016/j.tibtech.2010.03.002

Wang, T., Li, B., Nelson, C. E., and Nabavi, S. (2019). Comparative analysis of differential gene expression analysis tools for single-cell RNA sequencing data. *BMC Bioinform.* 20:40. doi: 10.1186/s12859-019-2599-6

Wang, T., and Nabavi, S. (2018). SigEMD: a powerful method for differential gene expression analysis in single-cell RNA sequencing data. *Methods* 145, 25–32. doi: 10.1016/j.ymeth.2018.04.017

Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: A revolutionary tool for transcriptomics. *Nat. Rev. Genet.* 10, 57–63. doi: 10.1038/nrg2484

Welch, J. D., Hartemink, A. J., and Prins, J. F. (2017). MATCHER: manifold alignment reveals correspondence between single cell transcriptome and epigenome dynamics. *Genome Biol.* 18:138. doi: 10.1186/s13059-017-1269-0

Welzel, G., Seitz, D., and Schuster, S. (2015). Magnetic-activated cell sorting (MACS) can be used as a large-scale method for establishing zebrafish neuronal cell cultures. *Sci. Rep.* 5:7959. doi: 10.1038/srep07959

Wills, Q. F., Livak, K. J., Tipping, A. J., Enver, T., Goldson, A. J., Sexton, D. W., et al. (2013). Single-cell gene expression analysis reveals genetic associations masked in whole-tissue experiments. *Nat. Biotechnol.* 31, 748–752. doi: 10.1038/nbt.2642

Wong, K., Navarro, J. F., Bergenstråhle, L., Ståhl, P. L., and Lundeberg, J. (2018). ST Spot Detector: a web-based application for automatic spot and tissue detection for spatial transcriptomics image datasets. *Bioinformatics* 34, 1966–1968. doi: 10.1093/bioinformatics/bty030

Wyatt Shields, C. IV, Reyes, C. D., and López, G. P. (2015). Microfluidic cell sorting: a review of the advances in the separation of cells from debulking to rare cell isolation. *Lab Chip* 5, 1230–1249. doi: 10.1039/c4lc01246a

Xin, Y., Kim, J., Ni, M., Wei, Y., Okamoto, H., Lee, J., et al. (2016). Use of the Fluidigm C1 platform for RNA sequencing of single mouse pancreatic islet cells. *Proc. Natl. Acad. Sci. U.S.A.* 113, 3293–3298. doi: 10.1073/pnas.1602306113

Xue, R., Li, R., and Bai, F. (2015). Single cell sequencing: technique, application, and future development. *Sci. Bull.* 60, 33–42. doi: 10.1007/s11434-014-0634-6

Yip, S. H., Wang, P., Kocher, J. P. A., Sham, P. C., and Wang, J. (2017). Linnorm: improved statistical analysis for single cell RNA-seq expression data. *Nucleic Acids Res.* 45:e179. doi: 10.1093/nar/gkx828

Yu, P., and Lin, W. (2016). Single-cell transcriptome study as big data. *Genomics Proteomics Bioinform.* 14, 21–30. doi: 10.1016/j.gpb.2016.01.005

Zaharia, M., Franklin, M. J., Ghodsi, A., Gonzalez, J., Shenker, S., Stoica, I., et al. (2016). Apache spark. *Commun. ACM* 59, 56–65. doi: 10.1145/2934664

Zare, R. N., and Kim, S. (2010). Microfluidic platforms for single-cell analysis. *Annu. Rev. Biomed. Eng.* 12, 187–201. doi: 10.1146/annurev-bioeng-070909-105238

Zhang, Z., and Wang, W. (2014). RNA-skim: a rapid method for RNA-Seq quantification at transcript level. *Bioinformatics* 30, i283–i292. doi: 10.1093/bioinformatics/btu288

Zheng, G. X. Y., Terry, J. M., Belgrader, P., Ryvkin, P., Bent, Z. W., Wilson, R., et al. (2017). Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* 8:14049. doi: 10.1038/ncomms14049

**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: 03 May 2021
doi: 10.3389/fgene.2021.648524

# Omics Approaches for Understanding Biogenesis, Composition and Functions of Fungal Extracellular Vesicles

*Daniel Zamith-Miranda[1,2*†], Roberta Peres da Silva[3†], Sneha P. Couvillion[4], Erin L. Bredeweg[5], Meagan C. Burnet[4], Carolina Coelho[3], Emma Camacho[6], Leonardo Nimrichter[7], Rosana Puccia[8], Igor C. Almeida[9], Arturo Casadevall[6], Marcio L. Rodrigues[10,11], Lysangela R. Alves[10], Joshua D. Nosanchuk[1,2] and Ernesto S. Nakayasu[4*]*

[1] Department of Microbiology and Immunology, Albert Einstein College of Medicine, Bronx, NY, United States, [2] Division of Infectious Diseases, Department of Medicine, Albert Einstein College of Medicine, Bronx, NY, United States, [3] MRC Centre for Medical Mycology, University of Exeter, Exeter, United Kingdom, [4] Biological Sciences Division, Pacific Northwest National Laboratory, Richland, WA, United States, [5] Environmental and Molecular Sciences Laboratory, Pacific Northwest National Laboratory, Richland, WA, United States, [6] Department of Molecular Microbiology and Immunology, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, United States, [7] Laboratório de Glicobiologia de Eucariotos, Instituto de Microbiologia Paulo de Góes, Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brazil, [8] Departamento de Microbiologia, Imunologia e Parasitologia, Escola Paulista de Medicina-Universidade Federal de São Paulo, São Paulo, Brazil, [9] Department of Biological Sciences, Border Biomedical Research Center, University of Texas at El Paso, El Paso, TX, United States, [10] Laboratório de Regulação da Expressão Gênica, Instituto Carlos Chagas-FIOCRUZ PR, Curitiba, Brazil, [11] Instituto de Microbiologia Paulo de Góes, Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brazil

Extracellular vesicles (EVs) are lipid bilayer structures released by organisms from all kingdoms of life. The diverse biogenesis pathways of EVs result in a wide variety of physical properties and functions across different organisms. Fungal EVs were first described in 2007 and different omics approaches have been fundamental to understand their composition, biogenesis, and function. In this review, we discuss the role of omics in elucidating fungal EVs biology. Transcriptomics, proteomics, metabolomics, and lipidomics have each enabled the molecular characterization of fungal EVs, providing evidence that these structures serve a wide array of functions, ranging from key carriers of cell wall biosynthetic machinery to virulence factors. Omics in combination with genetic approaches have been instrumental in determining both biogenesis and cargo loading into EVs. We also discuss how omics technologies are being employed to elucidate the role of EVs in antifungal resistance, disease biomarkers, and their potential use as vaccines. Finally, we review recent advances in analytical technology and multi-omic integration tools, which will help to address key knowledge gaps in EVs biology and translate basic research information into urgently needed clinical applications such as diagnostics, and immuno- and chemotherapies to fungal infections.

**Keywords: extracellular vesicles, fungi, virulence, systems biology, proteomics, metabolomics, lipidomics, transcriptomics**

OPEN ACCESS

***Edited by:***
*Bhabatosh Das, Translational Health Science and Technology Institute (THSTI), India*

***Reviewed by:***
*Yifan Ge, Massachusetts General Hospital and Harvard Medical School, United States*
*Ali Salehzadeh-Yazdi, University of Rostock, Germany*
*Krishnamohan Atmakuri, Translational Health Science and Technology Institute (THSTI), India*

***Correspondence:**
*Daniel Zamith-Miranda*
*daniel.zamithmiranda@einsteinmed.org*
*Ernesto S. Nakayasu*
*ernesto.nakayasu@pnnl.gov*

*† These authors have contributed equally to this work*

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

***Received:*** *31 December 2020*
***Accepted:*** *06 April 2021*
***Published:*** *03 May 2021*

***Citation:***
*Zamith-Miranda D, Peres da Silva R, Couvillion SP, Bredeweg EL, Burnet MC, Coelho C, Camacho E, Nimrichter L, Puccia R, Almeida IC, Casadevall A, Rodrigues ML, Alves LR, Nosanchuk JD and Nakayasu ES (2021) Omics Approaches for Understanding Biogenesis, Composition and Functions of Fungal Extracellular Vesicles. Front. Genet. 12:648524. doi: 10.3389/fgene.2021.648524*

## INTRODUCTION

Cells secrete a variety of molecules to the extracellular milieu, from the smallest metabolites to large proteins and glycoconjugates. These secreted molecules range from toxic catabolites to cell communication molecules, virulence factors, and enzymes involved in nutrient acquisition. One particularly intriguing mechanism of secretion is the release of extracellular vesicles (EVs). Organisms from all kingdoms have been described to release EVs (Toyofuku et al., 2019; Woith et al., 2019; Rayamajhi and Aryal, 2020; Rizzo et al., 2020b). In fungi, EVs were first described and partially characterized in *Cryptococcus neoformans* (Rodrigues et al., 2007). Since the original description, different omics approaches have been instrumental for the characterization of EVs, from their putative mechanisms of biogenesis to their function in the fungal biology. In this article, we review the contribution of different omics approaches to study fungal EVs.

## INITIAL CHARACTERIZATION OF FUNGAL EXTRACELLULAR VESICLES

*Cryptococcus neoformans* EVs comprise a heterogeneous population of lipid bilayer vesicles, including pigment-containing, electron-lucid, electron-dense, and membrane-associated electron-dense vesicles. The first proteomic analysis of cryptococcal EVs led to the identification of 76 proteins involved in a variety of functions, including virulence, oxidative stress, unfolded-protein response, cellular metabolism, protein translation, signal transduction, cytoskeleton organization, and also proteins found in the plasma membrane (Rodrigues et al., 2008). Subsequent studies in *Histoplasma capsulatum* identified 206 EV-associated proteins with a similar diversity of functions to *C. neoformans* EVs, besides proteins involved in cell synthesis and remodeling (Albuquerque et al., 2008). Lipidomic analysis of *H. capsulatum* EVs led to the identification of 18 phospholipids, including phosphatidylethanolamine, phosphatidylcholine, and phosphatidylserine species (Albuquerque et al., 2008). These initial characterization of fungal EVs opened new questions about their biogenesis and roles in infection, along with their potential use for clinical and biotechnological applications, as discussed in the subsequent sections.

## CELLULAR SITES AND MECHANISMS OF EXTRACELLULAR VESICLE FORMATION IN FUNGI

The precise cellular sites and mechanisms of fungal EVs formation are still not fully defined. At first, there was some skepticism in the field that EVs may be products of dying cells whereby released lipids self-assembled into vesicles. Skepticism about EVs was also fueled by concerns of how such large structures could cross the cell wall, which was viewed as a rigid structure that would preclude vesicular transport. However, we have shown that heat-killed *C. neoformans* failed to secrete EVs (Rodrigues et al., 2007). With regard to the cell wall transit, recent studies have shown that this structure is easily penetrated by vesicles (Walker et al., 2018). To further investigate this issue, we compared lipidomic analysis data of *H. capsulatum* EVs (Cleare et al., 2020) and whole cells (Burnet et al., 2020). We found a significant depletion of the energy storage lipid, triacylglycerol, and of the mitochondrial lipid, cardiolipin, in EVs when compared with whole cells (**Figure 1**). The fact that EVs and whole cells have distinct lipid composition suggests they are formed by specific EVs biogenesis process(es), rather than being a product of cell death, and/or breakdown. Lipidomic analysis of EVs from two *Paracoccidioides brasiliensis* isolates of different phylogenetic groups showed differences in sterol and fatty acid composition of EVs when compared with whole cells, also suggesting the involvement of specific organelles in EVs biogenesis (Vallejo et al., 2012a). In addition, the deletion of the sterol biosynthesis gene Erg6 induced changes in the lipid and protein content of *C. neoformans* EVs, suggesting a role for sterols in EVs formation (Oliveira et al., 2020). Studies have shown that fungal EVs can originate from intracellular organelles, such as endosomes (Oliveira et al., 2010; Zarnowski et al., 2018; Zhao et al., 2019; Park et al., 2020), or at the plasma membrane (Rodrigues et al., 2000, 2013; Rizzo et al., 2020a). Morphological studies of *C. neoformans* showed structures resembling multivesicular bodies (MBVs) that can fuse to the plasma membrane, resulting in the release of intraluminal MBV vesicles into the fungal periplasm (Takeo et al., 1973). Those images suggested that populations of fungal EVs might correspond to exosomes, which are mammalian EVs released to the extracellular milieu by the fusion of MVBs to the plasma membrane (Raposo and Stoorvogel, 2013). This biogenesis pathway was supported by recent studies in *C. neoformans* (Park et al., 2020) and *Candida albicans* (Zarnowski et al., 2018), in which deletion of genes affecting MVB formation resulted in aberrant vesicles and/or decreased EVs production. In *Saccharomyces cerevisiae*, deletion of several regulators of either conventional or unconventional secretion resulted in alterations of EVs composition, as measured by proteomic analysis (Oliveira et al., 2010). In *C. neoformans*, the deletion of SEC6, a gene participating in the post-Golgi secretory pathway, also resulted in reduced EVs formation (Panepinto et al., 2009). In the filamentous fungus *Neurospora crassa*, GFP-localization of SEC-6, -5, -8, and -15 subunits of the exocyst complex each form a crescent just beyond the cluster of vesicles of the Spitzenkörper at an extending hyphal tip (Riquelme et al., 2014). The exocyst allows a physical linkage of the vesicle cluster to the apical membrane. The cellular events of these EVs is also a matter of debate, including fusion or vesicle budding and secretion (Rizzoli and Jahn, 2007; Miura and Ueda, 2018). These studies show that intracellular regulators of secretory pathways, such as the post-Golgi pathway, participate in fungal EVs formation, similar to what occurs in mammalian EVs.

As observed in protozoan parasites (Marcilla et al., 2014; Szempruch et al., 2016) and mammals (Raposo and Stoorvogel, 2013; Stahl and Raposo, 2019), fungal EVs can also be formed at the plasma membrane. Immunofluorescence of *C. neoformans*

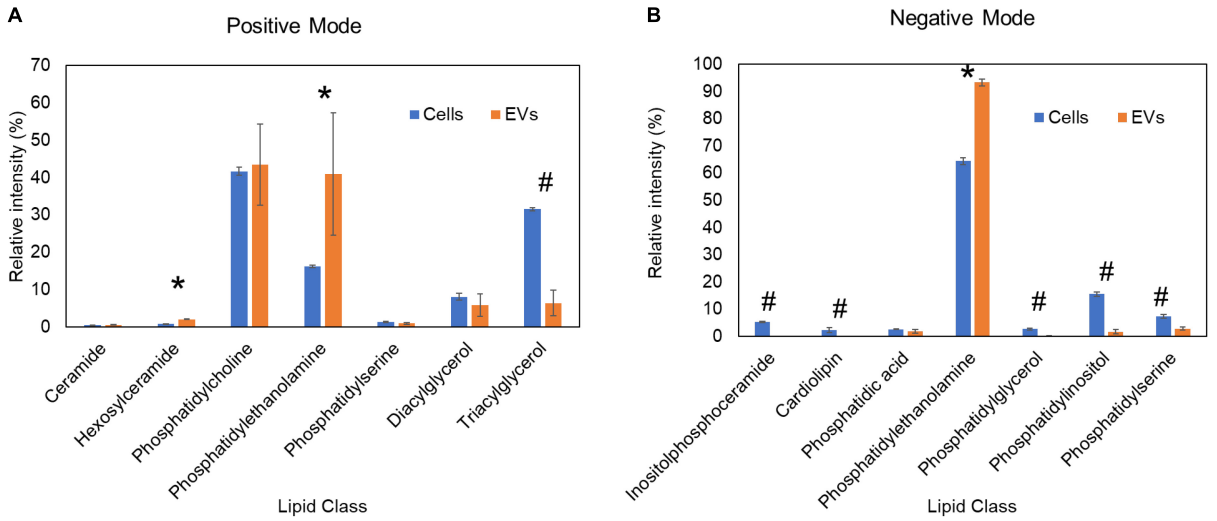



**FIGURE 1 |** Comparative lipidomic analysis of *H. capsula*tum EVs and whole cells. Lipidomics data from previous publications (Burnet et al., 2020; Cleare et al., 2020) of yeast cells grown in F12 medium. Each lipid class was normalized by the total ion intensity of the identified lipid species. Lipid classes significantly ($p \leq 0.05$ by Student's *t*-test) enriched in extracellular vesicles and whole cells are shown by asterisks and hash signs, respectively. **(A)** Lipids quantified by mass spectrometry in the positive ionization mode. **(B)** Lipids quantified by mass spectrometry in the negative ionization mode.

surface lipids revealed plasma membrane projections, suggesting that the plasma membrane could bud and release EVs (Rodrigues et al., 2000). The participation of the plasma membrane in fungal EVs formation was also shown using "wall-less" *Aspergillus fumigatus* cells (Rizzo et al., 2020a). Ultra-resolution microscopy analyses of these fungal protoplasts demonstrated the occurrence of EVs budding from the plasma membrane. Shedding of these plasma membrane-derived EVs was increased during cell wall synthesis, suggesting their participation in this process. Accordingly, protoplast EVs contain cell-wall polysaccharides and polysaccharide synthases, which are plasma membrane-associated enzymes (Rizzo et al., 2020a). These morphological and compositional studies support the presence of plasma membrane-derived EVs in fungi similar to the mammalian microvesicles (or ectosomes; Raposo and Stoorvogel, 2013; Stahl and Raposo, 2019). However, additional mechanisms of plasma membrane-derived EVs formation that differ from those described for mammalian microvesicles have been described in fungi. In *S. cerevisiae*, electron tomography studies revealed deep invaginations of the plasma membrane organized as two parallel membranes extending a few hundred nanometers toward the cell center. These structures can curve back to the cell surface, resulting in fusion with the plasma membrane and EVs formation (Rodrigues et al., 2013). Based on these observations, it has been proposed that fungal EVs might also originate from cytoplasmic content loading into a vesicle derived from the reshaping of the plasma membrane. The mechanisms behind this process are currently unknown, but they could represent a new pathway of EVs formation.

Overall, the studies described above show that fungi release exosome-like and microvesicle-like EVs, suggesting a conserved mechanism of EVs formation in lower and higher eukaryotes.

## CELL WALL REMODELING BY EXTRACELLULAR VESICLE CARGO

The cell wall is responsible for shaping fungal cells and for their resistance to diverse types of stress (Nimrichter et al., 2016). Nutrient availability, ambient pH, temperature, osmotic stressors, and other extracellular stimuli can lead to cell wall remodeling, which includes structural changes in their major components such as chitin, glucan, and glycoproteins (Nimrichter et al., 2016). The interplay between rigidity and plasticity of the cell wall is a key factor for fungal adaptation, survival, growth, and virulence (Nimrichter et al., 2016; Beauvais and Latgé, 2018). Although the cell wall synthesis and shaping are classically attributed to plasma membrane proteins, EVs might also play a role in this process (Vallejo et al., 2012b; Nimrichter et al., 2016; Ikeda et al., 2018; Zhao et al., 2019; Dawson et al., 2020). In *P. brasiliensis*, 60% of the non-covalently bound cell wall proteins, detected by proteomic analysis of two distinct isolates, have been described in fungal EVs (Longo et al., 2014). The EVs content can vary deeply depending on the growth conditions and fungal species (Vallejo et al., 2012a,b; Longo et al., 2014; Peres et al., 2015; Alves et al., 2019; Peres da Silva et al., 2019; Cleare et al., 2020). In addition, enrichment of cell wall remodeling enzymes is a conserved feature across fungal EVs (Vallejo et al., 2012b; Longo et al., 2014; Ikeda et al., 2018; Zhao et al., 2019; Dawson et al., 2020). Cell wall synthases or hydrolases have been found in EVs from *Candida auris, C. albicans*, *Cryptococcus deneoformans, Cryptococcus deuterogattii, C. neoformans*, *Fusarium oxysporum, P. brasiliensis*, *Sporothrix brasiliensis*, *S. cerevisiae*, *H. capsulatum*, and *Trichoderma reesei* (Vallejo et al., 2012b; Wolf et al., 2014; Ikeda et al., 2018; Bleackley et al., 2019; de Paula et al., 2019; Konečná et al., 2019; Zhao et al., 2019; Cleare et al., 2020; Dawson et al., 2020; Zamith-Miranda et al., 2020;

Rizzo et al., 2020c). When we compared the proteomic data of *H. capsulatum* EVs (Cleare et al., 2020) with whole cells (Burnet et al., 2020), the results showed that EVs were highly enriched in cell wall synthases and hydrolases, such as β-glucanase, β-1,3-glucanosyltransferase, chitin synthase, and chitinase (**Figure 2**). In *S. cerevisiae,* supplementation with chitin synthase Chs3-enriched EVs rescued the growth of cells treated with the cell wall-targeting antifungal caspofungin, suggesting that the EVs cargo *per se* is sufficient to supply components for cell remodeling (Zhao et al., 2019). Enzymes that have already been described in fungal EVs are involved in evasion of the immune system by modifying cell wall epitopes. The presence of lactate or exposure to hypoxia induced β-1,3-glucan masking in *C. albicans*. This effect was mediated by the secreted exo-β-1,3-glucanase, Xog1, which has been described as an EVs cargo in proteomic studies (Nimrichter et al., 2016; Konečná et al., 2019; Childers et al., 2020; Dawson et al., 2020). However, the direct participation of EVs in this process still needs to be confirmed. The ability to actively secrete cell wall synthases and hydrolases through EVs in response to extracellular environmental signals could represent a new mechanism of cell wall remodeling, which could affect the exposure of epitopes during infection, consequently resulting in modulation of the immune response.

## EXTRACELLULAR VESICLES AND ANTIFUNGALS

A seminal work by Andes laboratory showed that EVs were involved in biofilm formation in *C. albicans* (Zarnowski et al., 2018). Defective biofilm formation leads to an increased susceptibility to fluconazole. The authors performed gain-of-function experiments and showed that addition of wild-type EVs to biofilm-deficient strains restored biofilm formation and re-established fluconazole resistance. This work is akin to the previously mentioned work of EV-associated Chs3 restoring and rescuing cell wall defects and improving tolerance to antifungals (Zhao et al., 2019). Proteomic analyses performed in both studies showed that EVs carry a myriad of functional enzymes. Interestingly, it is possible that the delivery of a combination of enzymes in EVs allows for higher efficiency in cell wall remodeling.

## RNA CONTENT IN FUNGAL EXTRACELLULAR VESICLES

Omic approaches caused a major impact in deciphering the RNA content carried by EVs, being most of the data available characterized using RNA-seq. In fungi, RNA export via EVs was originally described in *S. cerevisiae*, *C. albicans*, *C. neoformans*, and *P. brasiliensis* (Peres et al., 2015). Similar to what has been published for mammalian EVs, the fungal EVs transcripts were composed mainly of small RNA (sRNA) sequences of up to 250 nt. The most abundant classes were non-coding (nc)RNA sequences of the small nucleolar RNA (snoRNAs), small nuclear RNA (snRNAs), and tRNA types (Peres et al., 2015). Subsequent transcriptomics analysis of EVs from *H. capsulatum* (Alves et al., 2019) and *Paracoccidioides* (Peres da Silva et al., 2019) isolates revealed that the small ncRNAs were mostly represented by short 25-nt fragments that aligned to a specific region of a particular mRNA. The presence of anti-sense RNA fragments in EVs might have a role in gene silencing, maybe similarly to that of micro

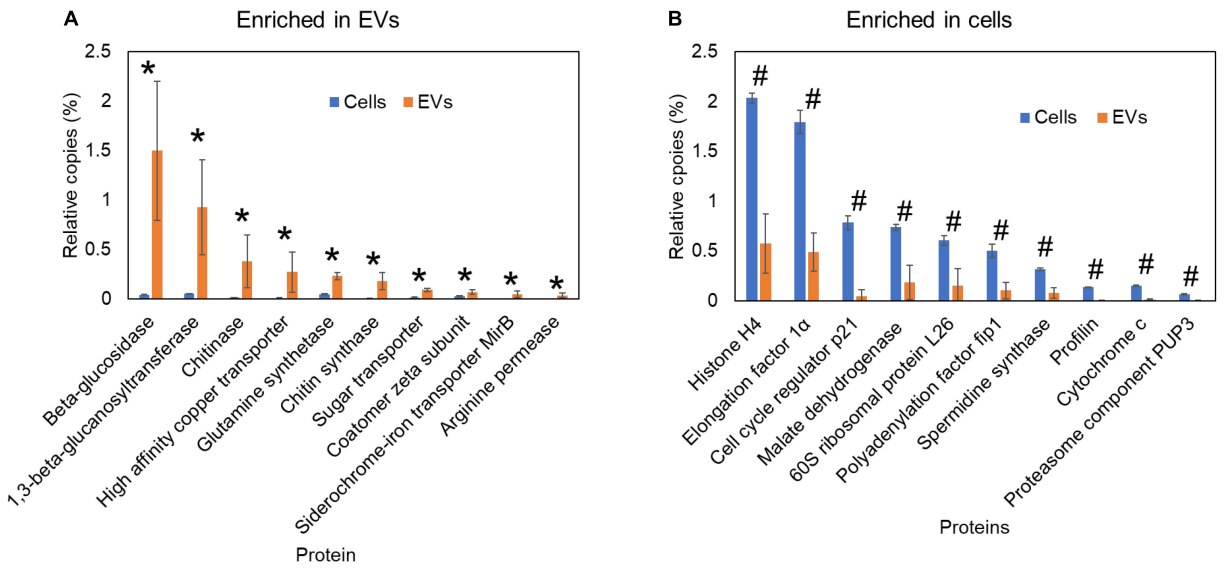



**FIGURE 2 |** Comparative proteomic analysis of *Histoplasma capsulatum* extracellular vesicles and whole cells. Proteomics data from previous publications (Burnet et al., 2020; Cleare et al., 2020) of yeast cells grown in F12 medium. The relative copy number of proteins in cells or extracellular vesicles was calculated using intensity based absolute quantification (iBAQ) by normalizing to the total iBAQ of each sample. Proteins significantly ($p \leq 0.05$ by Student's *t*-test) enriched in extracellular vesicles **(A)** and whole cells **(B)** are indicated as asterisk and hash signs, respectively.

RNAs (miRNAs) and fungal exonic short interfering RNAs (Nicolás and Ruiz-Vázquez, 2013; Son et al., 2017). The fungus *Malassezia sympodialis*, a member of the human skin microbiota, also exports EVs containing 16 to 22 nucleotides long RNAs. However, *M. sympodialis* lacks an RNAi machinery, suggesting that this fungal species might bear an alternative miRNA production pathway (Rayner et al., 2017). In the dimorphic fungus *Pichia fermentans,* the length of EVs RNAs ranges from 25 to 130 nt. These transcripts are involved in the transition of this fungus from yeast to pseudohyphal morphology, which occurs in response to specific environmental conditions (Leone et al., 2018).

Although most of the fungal EVs transcripts were small ncRNAs, full-length mRNAs have also been found. In general, they corresponded to genes of metabolic pathways, transcription regulation, cell cycle, vesicle-mediated transport, cellular responses to stress, and translation, depending on the species and the species isolate studied (Peres et al., 2015; Alves et al., 2019; Peres da Silva et al., 2019). An *in vitro* translation experiment has shown that mRNAs carried by *P. brasiliensis* and *P. lutzii* EVs are functional (Peres da Silva et al., 2019). Based on these results, it is reasonable to speculate that EVs mRNAs can be transferred and translated into the host cell, possibly modulating gene expression that could benefit the pathogen infection and survival. In *Cryptococcus gattii*, the EVs derived from a virulent strain induced, inside macrophages, survival and proliferation of a less virulent strain, which would normally be cleared by the host cell. This phenotype was decreased by EVs pre-treatment with RNase, supporting a role for EV-associated RNAs in the transfer of virulence traits (Bielska et al., 2018). However, the nature of these RNAs and their mechanism of action still needs to be further investigated.

Regarding RNA loading into EVs, in mammals, the autophagy protein LC3 has been reported as a recruiter of RNA-binding proteins to these compartments (Leidal et al., 2020). In fungi, the composition of RNA in the EVs can be affected by alterations in the intracellular vesicle and secretion pathway. The knockout of Golgi reassembly and stacking protein in *C. neoformans* deeply affected the EVs RNA composition, suggesting a role of the Golgi in the EVs RNA loading (Peres et al., 2018). To further evaluate the existence of a specific mechanism of RNA loading into fungal EVs, we compared the published transcriptomics data of *H. capsulatum* EVs with that of the whole cell (Alves et al., 2019). We observed a striking enrichment of specific RNA sequences in EVs, while the most expressed RNAs in the cells were present only in trace amounts in EVs (**Figure 3**). These results support the hypothesis that RNA sorting to EVs is finely regulated. In addition, robust RNA-seq data comparing the transcriptomics of EVs with that of their corresponding *C. albicans* cells cultivated both under control and mild stress conditions. We observed that the EVs and the cell transcriptomics was distinct in all growth conditions and that the RNA content of both EVs and cells was modulated under the stress conditions analyzed (Leitão, 2017).

## EXTRACELLULAR VESICLE COMPONENTS AS DISEASE BIOMARKERS AND CELL BIOLOGY MARKERS

While the terms biomarkers and markers are often and inappropriately used interchangeably, they have distinct definitions. Biomarkers, by definition, are molecular signatures

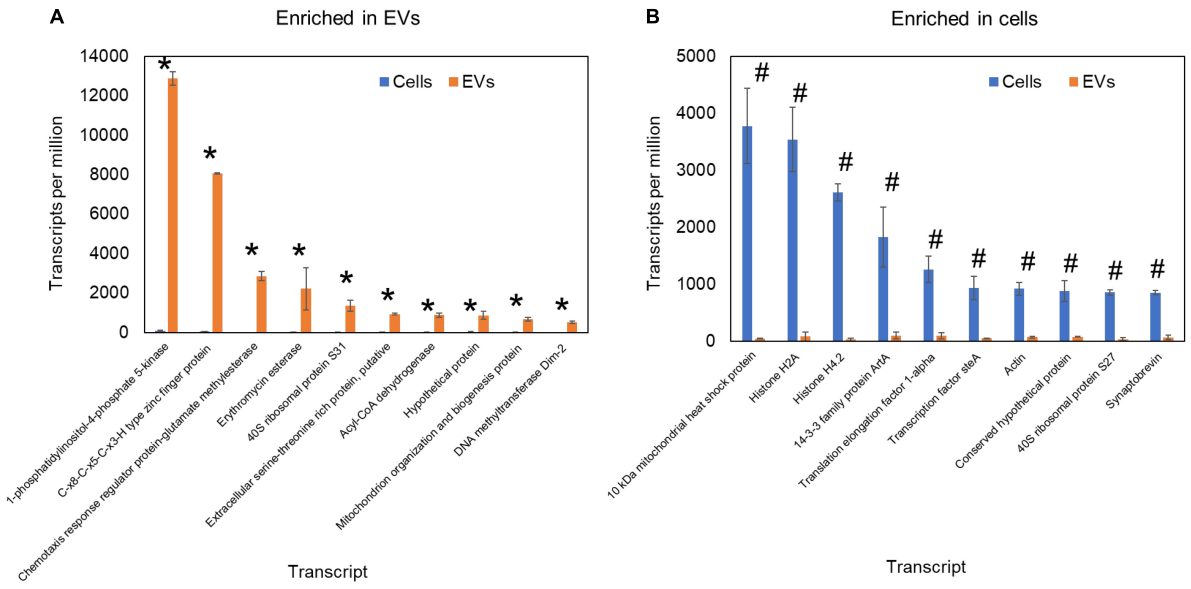



**FIGURE 3 |** Comparative transcriptomic analysis of RNA content *H. capsulatum* extracellular vesicles and whole cells. The expression level is represented as transcripts per million. Transcripts significantly ($p \leq 0.05$ by *t*-test) enriched in extracellular vesicles **(A)** and whole cells **(B)** are shown by asterisk and hash signs, respectively.

that can be used in the clinic to diagnose or predict the appearance or outcome of a disease (Strimbu and Tavel, 2010), whereas cell biology markers are molecules that can differentiate cell populations, cellular processes or cellular compartments (Zhang et al., 2019). The distinct composition of EVs from fungi and human cells make them good candidates for clinical diagnostic biomarkers and/or disease follow-up, i.e., for early assessment of chemotherapeutic outcomes and/or disease progression. Unfortunately, this subject has been understudied in fungi. In parasites, proteins from the murine malarial parasite *Plasmodium yoelii* have been detected in proteomics analysis of EVs from infected reticulocytes (Martin-Jaular et al., 2011). Similar findings have been recently observed in EVs isolated from plasma of a patient with chronic Chagas disease (Cortes-Serra et al., 2020). The use of EVs as biomarkers has been better explored in cancer biology. For instance, in prostate cancer, urinary EVs has been shown to carry RNAs that are signature of the disease outcome and are considered promising biomarker candidates (Pang et al., 2020). More recently, the proteomic analysis of EVs and particles from plasma and tissue showed that they can distinguish between normal and cancer cells with >90% sensitivity and specificity (Hoshino et al., 2020).

The discovery of specific markers of fungal EVs would have a major impact in cell biology research toward understanding their biogenesis, traffic, and function. In mammalian cells, good markers are the tetraspanins CD9, CD63, and CD81 (Andreu and Yáñez-Mó, 2014), which are currently unavailable for fungal EVs. In 2012, Vallejo et al. showed that 63% of the *P. brasiliensis* EVs proteins had orthologs described in EVs of *H. capsulatum, C. neoformans*, and *S. cerevisiae* (Vallejo et al., 2012b). Dawson et al. (2020) analyzed proteins that were enriched in EVs from three different strains of *C. albicans* when compared to the proteome of whole cells from the same strains. They found 47 commonly enriched proteins including Sur7, Evp1, and a variety of cell-wall synthesis and remodeling proteins. It should be noted, however, that the whole cell fraction that the authors analyzed did not include plasma membrane and, therefore, the results should be carefully considered (Dawson et al., 2020). Due to presence of cell-wall synthases and hydrolases in EVs from a variety of species (as discussed above) and since cell wall is conserved across the kingdom Fungi, it is reasonable to speculate that cell-wall synthesis and remodeling proteins could be a common marker for fungal EVs. The validation of these marker candidates and subsequent development of reagents may open new avenues to study the cell biology of fungal EVs, while biomarkers could be used for translational research as new diagnostic and prognostic tools.

## EXTRACELLULAR VESICLES AS TRANSPORTERS OF VIRULENCE FACTORS

Early characterization of EVs from *C. neoformans* showed the presence of important previously described virulence factors, specifically, glucoronoxylomannan (GXM), melanin, monohexosylceramide, laccase, urease, and phosphatase (Rodrigues et al., 2008; Eisenman et al., 2009). So far, however, only a few studies have investigated the participation of fungal EVs during infection in vertebrates. Using a murine model of cryptococcosis, Huang and colleagues demonstrated that the co-injection of *C. neoformans* with EVs facilitated the yeast transversal of the blood-brain barrier and enhanced the disease development (Huang et al., 2012). In addition, EVs are associated with a higher fungal burden and an increased lesion diameter in early stages of sporotrichosis caused by *S. brasiliensis* (Ikeda et al., 2018). Over the last decade, a number of EVs proteomic analysis carried out in diverse pathogenic fungal species described the finding of proteins that are associated with fungal virulence, but the association between effect and EVs cargo is still speculative, due to the lack of appropriate molecular tools, especially genetically deficient strains for most fungal species and pharmacological inhibitors. These EV-associated proteins include hydrolytic enzymes involved in protein and lipid degradation, proteins that protect against host oxidative responses and other types of stress (**Table 1**).

A combination of virulence factors is loaded in EVs from *Candida* species, including aspartyl proteases (SAPs), adhesion molecules, and lipases (Gil-Bona et al., 2015; Vargas et al., 2015; Karkowska-Kuleta et al., 2020; Martínez-López et al., 2020; Zamith-Miranda et al., 2020). In *P. brasiliensis*, six previously characterized virulence factors were detected in EVs (Vallejo et al., 2012b): gp43 (Torres et al., 2013), 14-3-3 (Marcos et al., 2016), catalase (Tamayo et al., 2017), cytochrome C peroxidase (Parente-Rocha et al., 2015), superoxide dismutase (Tamayo et al., 2016), and PbCDC42 (Almeida et al., 2009). In *C. neoformans*, proteomic studies revealed the presence of antioxidant enzymes such as catalase, superoxide dismutase, thioredoxin, thioredoxin reductase, and thiol-specific antioxidant protein (Rodrigues et al., 2008). Antioxidant proteins were found in *H. capsulatum* EVs, including catalase B, superoxide dismutase, and a thiol-specific antioxidant protein (Albuquerque et al., 2008), and *A. fumigatus*, of which Asp F3 and a putative thioredoxin reductase were found (Souza et al., 2019). Rodrigues and colleagues confirmed the urease and laccase activities in EVs released by *C. neoformans* (Rodrigues et al., 2008). Urease improves the survival of *C. neoformans* inside macrophages by modulating the phagosomal pH (Fu et al., 2018). EVs urease seems to be relevant during brain invasion (Huang et al., 2012). Laccase promotes pathogenesis of cryptococcal infections via multiple pathways: (1) synthesizing prostaglandins that may suppress local inflammatory responses (Erb-Downward et al., 2008); (2) inducing extrapulmonary dissemination to the brain (Noverr et al., 2004); (3) inhibiting the Th17-type cytokine response and neutrophils recruitment (Hansakon et al., 2020); (4) enhancing fungal survival in macrophages by mediating its escape through non-lytic exocytosis (De Oliveira Frazão et al., 2020); and (5) catalyzing the synthesis of melanin. However, the role in pathogenesis of these virulence factors present in EVs still need to be investigated.

Lipidomic, glycomic, and metabolomic studies of fungal EVs have also led to the identification of potential virulence factors. Lipidomic analysis comparing EVs from two *P. brasiliensis* isolates with different degrees of virulence showed distinct

**TABLE 1 |** Effects and virulence factors in fungal EVs.

| Fungus | *In vivo* effect of EVs | *In vitro* effect of EVs | Virulence factors carried by EVs | |
|---|---|---|---|---|
| *C. neoformans* | Pathogenesis | Stimulates cytokine production and antifungal activity in macrophages (Oliveira et al., 2010) | GXM | |
| | (Promotes brain infection) | Enhance adhesion and trans endothelial passage through endothelial cells activating lipid rafts (Huang et al., 2012) | Catalase and superoxide dismutase | |
| | Protection in *G. mellonella* and mice models of cryptococcosis (Rizzo et al., 2020c) | Melanin synthesis (Rodrigues et al., 2008) | Urease | |
| | | | Melanin/laccase | |
| *C. gattii* | nd* | Associated with virulence transference (Bielska et al., 2018) | Protein and RNA | |
| *C. albicans* | **Yeast EVs:** | **Yeast EVs:** | **Yeast EVs** | **Hyphae EVs** |
| | Protection in *G. mellonella* and mice models of candidiasis (Vargas et al., 2015, 2020) | Stimulates macrophages to produce NO and cytokines. Stimulates dendritic cells to produce cytokine and up-regulates MHCII and CD86 (Vargas et al., 2015) | **SAPs | ***SAPs |
| | | | Als3 and 4 | Als3 |
| | | Biofilm EVs: | PLB | PLB5 and PLC2 |
| | | Matrix production and biofilm drug resistance (Zarnowski et al., 2018) | | ****Ece1p |
| | | Hyphae EVs: | | |
| | | Induced TNFα release in THP-1 cells (Martínez-López et al., 2020) | | |
| *C. auris* | Induces adhesion to epithelium and activation of bone marrow-derived dendritic cells (Zamith-Miranda et al., 2020) | Adhesion to epithelial cells | Phosphatase | |
| | | | Peroxisomal catalase | |
| | | Dendritic cell activation | Superoxide dismutase | |
| | | | SAP10 | |
| | | | Phospholipases B and D | |
| | | | Thioredoxin Reductase | |
| *C. glabrata* | nd | nd | Phospholipase B | |
| *C. parapsilosis* | nd | nd | Lipase (Lip2) | |
| *C. tropicalis* | nd | nd | SAP | |
| | | | Hwp1-like protein | |
| | | | Lysophospholipase | |
| *H. capsulatum* | nd | Inhibits phagocytosis and killing by macrophages and impacts ROS generation (Matos Baltazar et al., 2016; Baltazar et al., 2018) | Catalase B, Superoxide Dismutase and a Thiol-specific antioxidant protein | |
| *P. brasiliensis* | nd | Induces production of proinflammatory mediators and the M1 polarization of macrophages. | gp43, 14-3-3, PbCdC42, catalase, superoxide dismutase | |
| | | Enhance the fungicidal activity of macrophages (da Silva et al., 2016) | | |
| *A. fumigatus* | | Induces the production of TNF-alpha and CCL-2 by macrophages | Asp F3 and a putative thioredoxin reductase | |
| | | Enhances the antifungal activity of macrophages and neutrophils (Souza et al., 2019) | | |
| *A. flavus* | Protection in *G. mellonella* model of aspergillosis | Induces the production of inflammatory mediators (NO and cytokines) and the M1 polarization of macrophages. Enhance the fungicidal activity of macrophages (Brauer et al., 2020) | nd | |
| *S. brasiliensis* | Increase in fungal burden and lesion diameter in a mice model of sporotrichosis (Ikeda et al., 2018) | Enhancement of yeast phagocytosis and fungal burden in dendritic cells. | 70 KDa-glycoprotein | |
| | | Increase in cytokine production (IL-12p40 and TNF-alpha; Ikeda et al., 2018) | | |

**nd – not dertermined.*
***So far SAP 4 was the only member never reported in yeast C. albicans EVs.*
****In hyphae EVs SAP2, 4, 5, 6, 7, 8, 9, and 10 were identified.*
*****Key proteins were found in hyphae EVs, but the toxin peptide candidalysin was not detected.*

phospholipid and sterol contents, which might be associated with differential virulence. The more virulent Pb18 isolate had higher ergosterol to brassicasterol ratio than the Pb3 strain (Vallejo et al., 2012a), and considering the function of ergosterol in triggering macrophage pyroptosis it can represent a virulence mechanism (Koselny et al., 2018). EVs from *C. albicans* and *C. auris* carry a variety of lysophospholipids (Zamith-Miranda et al., 2020), which might be correlated with expression of phospholipases in these organisms. In fact, lysophosphatidylcholine well-characterized regulators of the host immune response (Soehnlein et al., 2009; Carneiro et al., 2013; Gazos-Lopes et al., 2014) and might have a role in candidiasis virulence. Monohexosylceramides have been found in EVs released by *H. capsulatum*, *P. brasiliensis*, *C. neoformans*, *C. albicans*, and *C. auris* (Rodrigues et al., 2007; Vallejo et al., 2012a; Cleare et al., 2020; Zamith-Miranda et al., 2020). Monohexosylceramide has been associated with *C. neoformans* ability to grow in neutral and basic pH (Rittershaus et al., 2006), and to promote *C. albicans* infection (Rittershaus et al., 2006). *N*-acetylsphingosine (also known as C2-ceramide), a regulator of the T-cell function (Menné et al., 2000; Detre et al., 2006), has been reported in EVs from *C. auris* (Zamith-Miranda et al., 2020) but its function in virulence still need to be investigated.

Peres da Silva et al. (2015) showed that *P. brasiliensis* and *P. lutzii* have a polysaccharide (or hydrolysis fragments) with glycogen structure and a galactofuranosylmannan oligomer as the main glycans in EVs. Small amounts of 1,3- and 1,6-cell wall glucans were also found. Indeed, β-1,3-glucan is an important cell wall inflammatory pathogen-associated molecular pattern. The study also included glycan and plant/mammalian lectin microarray profiling of EVs surface, revealing the presence of ligands of DC-SIGN receptors, exposed mannose and *N*-acetylglucosamine residues, and *N*-acetylglucosamine-binding lectin(s) that can potentially mediate interaction with the host. *P. brasiliensis* EVs carbohydrate content could indeed be implicated in the transcriptome modulation of murine monocyte-derived dendritic cells (Peres da Silva et al., 2015). A mechanism of fungal resistance to the host defenses by EVs has been shown in *C. neoformans* by shutting off the host inflammasome. Metabolomic analysis identified that the aromatic metabolite DL-Indole-3-lactic acid is secreted inside EVs, which in turn could impair the inflammasome activation by the host cells (Bürgel et al., 2020).

Overall, omics analyses have found a variety of molecules associated with virulence and regulation of the host immune response. However, how EVs promote virulence with their molecules still needs additional investigations.

## HOST RESPONSE TO EXTRACELLULAR VESICLES

The presence of virulence factors and antigens suggests that fungal EVs could modulate the host response to infection. *A. flavus* and *P. brasiliensis* EVs enhance the phagocytosis of their respective yeast cells by macrophages (da Silva et al., 2016; Brauer et al., 2020). The EVs also induce the polarization of macrophages toward the proinflammatory M1 phenotype, which has been associated with high antifungal activity (da Silva et al., 2016; Brauer et al., 2020). Similar induction of pro-inflammatory cytokines has also been reported for EVs from *C. neoformans*, *C. albicans*, and *A. fumigatus* (Oliveira et al., 2010; Vargas et al., 2015; Souza et al., 2019). Conversely, EVs can also impair specific host responses. EVs released by *M. sympodialis* drive the production of the cytokine IL-4 by human peripheral blood mononuclear cells (Gehrmann et al., 2011). It is believed *M. sympodialis* EVs have a function in allergic responses. Proteomic analysis identified that 10 of 13 previously characterized allergens produced by the fungus are present in EVs. Two of these proteins were enriched in EVs as compared to fungal cells (Johansson et al., 2018).

Host immune factors, such as antibodies, can induce changes in the composition of EVs (Matos Baltazar et al., 2016; Baltazar et al., 2018). Incubation of *H. capsulatum* yeasts with antibodies against HSP60 (heat-shock protein 60), a protein enriched in cell wall and EVs, significantly changed the EVs cargo (Matos Baltazar et al., 2016; Baltazar et al., 2018). This treatment led to an increase in protein content and the virulence factor urease, suggesting a counteraction of the fungal resistance mechanisms against the host defenses (Matos Baltazar et al., 2016). Moreover, the EVs from antibody-treated *H. capsulatum* have an inhibitory effect on phagocytosis by macrophages (Baltazar et al., 2018). EVs from *S. brasiliensis* enhanced phagocytosis of the respective cells and increased the fungal burden in dendritic cells (Ikeda et al., 2018).

As potential targets for immunotherapies, EVs from *C. neoformans, C. albicans and A. flavus* have been shown to elicit at least a partial protection in the moth *Galleria mellonella* or in mice (Vargas et al., 2015; Colombo et al., 2019; Brauer et al., 2020; Rizzo et al., 2020c). The EVs cargo seems to be crucial to modulate this response. The presence of GXM and sterylglucosides in EVs reduces the protective effects of EVs in *G. mellonella* (Colombo et al., 2019). In macrophages, *C. neoformans* EVs containing GXM trigger a lower antifungal immunological response compared to EVs from a strain with reduced GXM production (Oliveira et al., 2010). *C. albicans* EVs activate murine macrophages and dendritic cells, inducing a protective immune response in immunosuppressed mice (Vargas et al., 2015; Vargas et al., 2020). Proteomic analysis revealed several candidates that could be involved with this response, including immunogenic proteins MP65 and Bgl2 (Nisini et al., 2001; Gil-Bona et al., 2015). Purified MP65 induces the expression of the antigen presentation protein MHC-II and co-stimulatory molecules, such as CD86, in dendritic cells (Pietrella et al., 2006). Similarly, the endo-β-1,3-glucanase Bgl2 has been tested as a vaccine candidate with promising results (Gil-Bona et al., 2015). These studies highlight the potential of fungal EVs as candidates for vaccine development.

## EXTRACELLULAR VESICLES AND BIOMASS DEGRADATION

Biomass degradation by fungi is of major importance for agriculture and for biotechnological purposes. In agriculture, fungal infections and degradation of plants can cause major losses, while in biotechnology fungi can be used to convert biomass into biofuels or other bioproducts of economic value (Almeida et al., 2019; Oh and Jin, 2020). *F. oxysporum* is an environmental fungus that attacks cotton crops leading to significant losses in productivity (Gordon, 2017). *F. oxysporum* EVs induce damage on leaves from cotton and *Nicotiana benthamiana*, a close relative of tobacco, but its spores or hyphae do not cause the same damage. This effect, however, is not intrinsic to fungal EVs, since EVs from *S. cerevisiae* are not phytotoxic (Bleackley et al., 2019). Similarly, the wheat pathogen *Zymoseptoria tritici* (Hill and Solomon, 2020) produces EVs, which may be a part of its transition from apoplastic, non-symptomatic growth to necrotrophy of wheat. While some carbon-active enzymes associated with EVs were produced under media-grown conditions, additional *Zymoseptoria* effectors are expected *in planta*.

In biotechnology, degradation of cellulose from plants releases carbon for biofuel production (Oh and Jin, 2020). Soluble sugars released by fungal enzymes are built into ethanol, lipids, secondary metabolites, or other bioproducts accessible by fungal fermentation. *T. reesei* is a fungus that produces large amounts of cellulolytic enzymes. The proteomic analysis of EVs produced by this fungus revealed that enzymes with cellulase activity are exported through EVs. The cellulolytic activity of EVs from *T. reesei* is induced when the fungus is cultivated in the presence of cellulose, suggesting that cargo contained in fungal EVs is altered according to the environment (de Paula et al., 2019). Engineering yeasts for direct cellulosic degradation into ethanol producing yeasts could lead to important gains in biofuel production (Oh and Jin, 2020). In further support of environmental cues regulating EVs, growth conditions of submerged media vs. solid state fermentation (SSF) have shown that *Aspergilli* produce different secreted protein profiles. *A. oryzae* (Oda et al., 2006) secreted 4-6x more protein in SSF. *A. brasiliensis* ATCC9642 in SSF produces several differentially expressed proteins which lack secretion signals, suggesting an alternative route of secretion such as EVs (Volke-Sepulveda et al., 2016). Dissecting the molecular trafficking of EVs formation would create a novel compartment for enzyme delivery, or bioproduct collection from a culture without specialized transport proteins.

## NEW METHODOLOGIES AND INSTRUMENTATION

The small size and scarce amount of material obtainable in preparations of EVs represent a major analytical challenge. However, advances in omics technologies have immensely improved the sensitivity, throughput, and robustness of the measurements, leading to a more comprehensive characterization of the EVs molecular composition. For instance, back in 2008, EVs proteomic analysis in *C. neoformans* led to identification of 76 proteins (Rodrigues et al., 2008). Current high-resolution tandem mass spectrometry (HR-MS/MS)-based approaches (Lesur and Domon, 2015), including nanoflow liquid chromatography coupled to HR-MS/MS (nanoLC-HR-MS/MS; Sanders and Edwards, 2020), allow to identify and quantify over 2,000 proteins in fungal EVs (Zhao et al., 2019; Cleare et al., 2020). In this section, we will cover recent technological advances and their current impact and perspectives in analyzing fungal EVs.

### RNA-seq

The RNA yield recovered from EVs is quite variable when we compare samples from different origins and that have been isolated using distinct protocols. For fungal cells, there are many media and growth conditions that can affect the number of EVs obtained and, consequently, the amount of RNA. Usually, the EVs RNA yield for *C. neoformans*, *C. albicans*, *P. brasiliensis*, and *H. capsulatum* ranges from 1 to 15 ng when the EVs are isolated from culture supernatant after two ultracentrifugation steps, corresponding to EVs enrichment and washing (Peres et al., 2015). Growing fungi in solid media, instead liquid cultures, can improve the EV RNA recovery yield to up to 50 ng, as shown for *C. neoformans* and *C. gattii* preparations (Reis et al., 2019).

In the recent years, the next-generation sequencing has emerged as a robust tool to resolve the diversity of RNA sequences in EVs. RNA-seq has enabled major advances in the analysis of EVs RNAs, allowing the identification of low input amounts of distinct RNA populations (Kim et al., 2017). Such technology allows the comparison of EVs RNA across samples generated under a variety of experimental designs, such as different growth conditions, stresses, or even interspecies studies (Mateescu et al., 2017; Yeri et al., 2018). Given that small RNAs are highly enriched in EVs, most of the library construction protocols focus on the fractioning of this RNA population. There are many kits available, but most of them follow similar procedures, involving multiple steps for the small RNA purification (Giraldez et al., 2018). Overall, we believe that RNA-seq will have a major impact in identifying EVs RNAs that could have a role in the host-pathogen interaction. Especially those sequences regulating expression of mRNAs in recipient cells and consequently, affecting the host immune response or the fungal pathogenicity.

### Recent Advances in Sample Preparation for Mass-Spectrometry-Based (multi)Omic Analysis

The eternal challenge of an analytical chemist is to improve sensitivity, precision, and speed of the instrumentation, enabling the accurate analysis of trace amounts of samples in large scale. In this context, the recent developments in single-cell analysis might have a major impact in analysis of EVs as most of the procedures are easily adaptable for analyzing EVs. An important concept is to keep the volumes and contact surfaces as small as possible during the sample preparation and analysis to reduce losses due to contact absorption. Online sample preparations can eliminate losses associated with pipetting and sample transfer. One of such

techniques, called SNaPP (simplified nanoproteomics platform for reproducible global proteomics), allows the preparation and analysis of proteomic samples from nanograms of proteins (Huang et al., 2016). SNaPP has been successfully used to analyze Nipah virus-like particles (Johnston et al., 2019), which are secreted structures that share many characteristics with EVs, including size and the presence of a lipid membrane. Therefore, SNaPP has a great potential to be used for preparing EVs samples. Another technique to prepare small scale samples, but that has not yet been applied to study EVs, is the nanoPOTS (nanodroplet processing in one pot for trace samples; Zhu et al., 2018). nanoPOTS uses a microfluidic robot to prepare samples in nanoliter volumes, virtually eliminating any losses associated with absorption of proteins and peptides to the walls of pipettes and tubes. This technique has enabled to identify and quantify up to 2,500 proteins in single-cell proteomics analysis (Tsai et al., 2020), therefore, it might have an impact on analyzing EVs in the future.

Another analytical challenge in EVs studies is to obtain multiple omics measurements from the same samples. Having multiple measurements from the same samples is highly desirable since it decreases variability between datasets and efforts/costs associated with EVs preparation. Solvent phase separation-based extraction of metabolites, lipids, and proteins have been developed (Coman et al., 2016; Nakayasu et al., 2016). While the simultaneous metabolite, protein, lipid extraction (SIMPLEX) approach is based on methyl tert-butyl ether, methanol, and water; the metabolite, protein, and lipid extraction (MPLEx) technique uses a mixture of chloroform, methanol, and water for phase separation. To our knowledge, there are no papers in the literature reporting the analysis of EVs using SIMPLEX to extract the samples, even though its potential has been highlighted in a few review articles (Rosa-Fernandes et al., 2017; Ramirez et al., 2018). MPLEx, on the other hand, has been successfully used to analyze EVs from *H. capsulatum*, *C. auris*, and from the Gram-positive bacterium *Listeria monocytogenes* (Coelho et al., 2019; Cleare et al., 2020; Zamith-Miranda et al., 2020).

## Small Scale and Increased Throughput of Lipidomics, Metabolomics and Proteomics

When compared to conventional LC-MS/MS, advanced multidimensional analytical platforms such as LC-ion mobility spectrometry (IMS)-MS/MS, which combines liquid chromatography, ion mobility spectrometry, and tandem mass spectrometry, offers improvements in separation peak capacity, dynamic range, number of analytes detected, and quality of mass spectra (Baker et al., 2010; Rainville et al., 2017). Thus, the addition of IMS allows for improved coverage of metabolites, lipids, and proteins in EVs. Ion mobility also allows to separate isobaric molecules and enables detailed characterization of lipid molecular structure, including double-bond location, *cis*/*trans* orientation, and *sn*-positions of alkyl and/or acyl chains, none of which are possible using conventional LC-MS technologies (Zheng et al., 2018). Poad et al. (2018) have previously demonstrated the successful incorporation of the online ozonolysis into the existing LC-IMS-MS/MS instrumentation to enable characterization of double bonds in lipid standards. The IMS dimension significantly improves assignment of the ozonolysis products to their precursor ions.

One important challenge of the metabolomics and lipidomics fields is the reliable identification of molecules. Most of the identifications to date are validated by comparing MS/MS spectra and chromatographic elution profiles to *bona fide* standards. The inclusion of IMS separations opens a new perspective in the identification of small molecules without standards. The separation in IMS can be predicted with high precision (<5% error; Colby et al., 2019). Therefore, using multiple pieces of information, i.e., high mass accuracy, ion mobility separation, and tandem mass fragmentation might allow the identification of molecules without the need to validate them against standards, which would make metabolomic and lipidomic analyses much faster. A major bottleneck for metabolite identification is the reliance on reference databases that are constructed from authentic reference materials. This approach is highly limiting due to the cost and availability of authentic standards. Both commercial and publicly available reference databases currently only represent a small fraction of the possible metabolites that exist in biological systems (Dobson, 2004). The recent emergence of standards-free methods based on *in silico* methods such as quantum chemistry, machine learning, and deep learning have enabled accelerated building of very large reference libraries (Colby et al., 2020) with better coverage of the known chemical space, which was not feasible using authentic standards. Standards-free metabolomic approaches can therefore provide comprehensive coverage of the metabolome in EVs analyses, accelerating the discovery of small molecules and improving our understanding of their biological functions.

The low throughput of EVs analysis is a major challenge to conduct clinical studies to assess the EVs potential as a disease biomarker or to better understand their function in human health. One way to increase the throughput of analysis is by multiplexing samples. For proteomic analysis, isotope labeling or isobaric labeling can be used to multiplex samples. In this context, isobaric tags, such as isobaric tags for relative and absolute quantitation (iTRAQ) and tandem mass tag (TMT), allow multiplexing up to 16 samples (Li et al., 2020), thus increasing the speed of sample analysis. Another advantage of multiplexing is the relative increase in sample amount, as compared to analyzing them individually, providing gains in sensitivity. However, such techniques are not available for lipidomic and metabolomic analyses. Therefore, the gain in throughput of such analyses relies on reducing the time needed for the analysis of each sample. Ion mobility spectrometry is not only fast (milliseconds per scan), but also adds another dimension of separation to the LC-MS analysis, which allows to shorten the chromatographic separation time without compromising the depth of coverage. One of such concepts has been developed by the Evosep company (Odense, Denmark). The Evosep chromatographic system performs offline sample solid-phase extraction (SPE) and elution gradient, which is accumulated in a sample loop. During the analysis, the flow passes through the loop carrying the samples to an analytical column and subsequently to the mass spectrometer. This process

eliminates the time needed for column regeneration and re-equilibration, consequently increasing the number of samples that can be analyzed in a day. The Evosep system coupled to IMS-MS allows to analyze up to 60 samples a day with a coverage that exceeds 5,000 proteins or more than 1,000 proteins with 5-min separation gradients (Meier et al., 2018; Bekker-Jensen et al., 2020). For the analysis of lipids and metabolites, due to the overall lower intrinsic complexity of their structures and MS fragmentation profile, as compared to peptides, it is possible to even reduce the time needed for the analysis each sample by an IMS-MS-based approach, as recently proposed by Zhang et al. (2016). These authors used an automated SPE platform coupled to IMS-MS to analyze metabolites and xenobiotics in the human urine (Zhang et al., 2016). They employed six SPE columns with a broad range of chemical properties that allowed them to capture distinct sets of molecules. In this configuration, each sample is run 6 times, or 12 times if analyzed in both positive and negative modes, but each cycle only takes 10 s. Therefore, each sample only takes about 2 min to be analyzed, allowing the analysis of hundreds of samples a day.

## Integration of Multi-Omics Data

The combination of multiple omics measurements allows to obtain a much deeper view of the EVs composition. Despite the challenges associated with processing large amounts of data, there are excellent tools to handle such tasks, which we will not cover in this review. Integrating the results of each omics measurement provides an opportunity to understand much deeper details and the processes occurring in EVs and cells. Despite some interpretations might be limited in EVs, the integration of proteomics with metabolomics and lipidomics, for instance, might show consistent changes in the levels of enzymes, substrates, and products. We have applied this approach to study whole cells of the multidrug-resistant fungus *C. auris*, which showed consistent changes in the levels of enzymes with their respective lipid and metabolite products in drug-resistant strains (Zamith-Miranda et al., 2019). Further integration of such data with RNA-seq results may further provide insights into the transcriptional and post-transcriptional regulation of genes. Nematode parasites, for instance, secrete miRNAs via EVs that target and modulate the expression of host immune factors (Buck et al., 2014).

# CONCLUDING REMARKS, MAJOR KNOWLEDGE GAPS AND PERSPECTIVES

Omic approaches have been contributing to the field of fungal EVs since the initial characterization of these extracellular "organelles." Proteomic, lipidomic, metabolomic and RNA-seq technologies enabled a detailed characterization of the molecular composition of the fungal EVs, bringing new insights into their biogenesis, and biological and pathophysiological functions. Multiple omics analyses of mutant strains defective in different secretory pathway components have shown the participation of the Golgi complex in EVs biogenesis. In terms of EVs functions, omics analyses have played a key role in showing the participation of EVs in cell-wall remodeling and downstream function in antifungal resistance. Fungal EVs have also been shown to carry a variety of virulence factors, which opens new perspectives on how they are delivered to and interact with the host. Here are some major knowledge gaps in the field and how omics can contribute to close such gaps:

- *Biogenesis and EVs populations*. Despite numerous efforts of the field, the question regarding different EVs populations, their composition, and biogenesis processes is still open. Novel EVs purification techniques, such as differential centrifugation and affinity purification, might allow to separate different EVs populations, which in combination with genetics can lead to the identification of biogenesis pathways. Omics analyses can contribute by comprehensively characterizing the composition of different EVs populations.
- *Markers and biomarkers*. The absence of markers and biomarkers is a major impairment to perform cell biology and clinical studies on EVs, respectively. Ideally, would be to perform omics analysis of dozens of fungal species to identify EVs markers, whereas for developing clinically relevant biomarkers it often requires the analysis of hundreds to thousands of samples from multiple cohorts. Therefore, faster and more sensitive techniques will empower such studies.
- *Mechanisms of virulence*. Omics analysis will continue to detect or identify new virulence factors and their mechanisms. A major bottleneck is to study their mechanisms of action in host cells and animal models. We believe that techniques, such as co-affinity purification, followed by nanoLC-HR-MS/MS or an orthogonal analytical approach such as IMS-MS/MS, may have a pivotal role in identifying targets of virulence factors in the host cells, leading to a better understanding of the pathogenic mechanisms.
- *Structure-function relationship of fungal EVs molecules*. Structure-function relationship is another major gap in fungal EVs research. Thus far, most aforementioned studies that have identified fungal EVs molecules by proteomic, lipidomic, transcriptomic, and other omic approaches are descriptive in nature. In the coming years, investigators in the field should make greater strides to study the structure-function relationship of some of these fungal molecules, particularly those that have known or potential bioactivity, based on published data on fungi or other pathogen(s). This would require considerable improvement in (a) gene expression and knockout techniques for fungi; (b) purification and structural analysis of fungal molecules; and (c) chemical and/or enzymatic synthesis of fungus-specific molecular targets such as lipids, glycoconjugates, and metabolites.

Technological advances and Science are highly dependent on each other to progress, which is not different for EVs biology and

omic analyses. We foresee that advances in omic technologies will continue having major impact in studying EVs biology.

## AUTHOR CONTRIBUTIONS

All authors contributed to the literature review, writing of the manuscript and revision, and editing of the manuscript. All authors revised and approved the final version of the manuscript.

## FUNDING

DZ-M, JN, and EN were supported by NIH R21 AI124797. EB and EN were supported by a Laboratory Directed Research and Development project from Pacific Northwest National Laboratory (PNNL). AC was supported in part by NIH grants AI052733, AI15207, and HL059842. We were also very grateful to the Biomolecule Analysis and Omics Unit (formerly, Biomolecule Analysis Core Facility), supported by the NIH/NIMHD grants 2G12MD007592-21 and 5U54MD007592 (to Robert A. Kirken), for the access to mass spectrometry systems and other analytical instruments used in several of the studies described here. IA was partially supported by NIH/NIMHD grant 5U54MD007592, and a Special Visiting Researcher of the CNPq/Science Without Borders Science Program, Brazil. EC and AC were funded by the Johns Hopkins Malaria Research Institute Pilot Grant Casadevall_123. RPu was supported by the Brazilian funding agencies FAPESP, CAPES, and CNPq. RPe and CC were funded by Medical Research Council Centre for Medical Mycology at University of Exeter (MR/N006364/2). Parts of this work were performed in the Environmental Molecular Science Laboratory, a United States Department of Energy (DOE) national scientific user facility at PNNL in Richland, WA, United States. EC and AC are funded by NIAID R01 AI052733. EN was also supported by R01 AI127465 from NIAID.

## ACKNOWLEDGMENTS

This review is dedicated to the memory of Luiz Rodolpho Raja Gabaglia Travassos (1938–2020; Universidade Federal de Sao Paulo, Escola Paulista de Medicina, Brazil), who made numerous seminal contributions to the fields of fungal molecular and cellular biology, and glycobiology. Travassos was one the first investigators who proposed extracellular vesicles as carriers of fungal cell wall, plasma membrane, and intracellular bioactive molecules that could eventually modulate mammalian host infection and immune response, and fungal immunoevasion mechanisms.

## REFERENCES

Albuquerque, P. C., Nakayasu, E. S., Rodrigues, M. L., Frases, S., Casadevall, A., Zancope-Oliveira, R. M., et al. (2008). Vesicular transport in *Histoplasma capsulatum*: an effective mechanism for trans-cell wall transfer of proteins and lipids in ascomycetes. *Cell. Microbiol.* 10, 1695–1710. doi: 10.1111/j.1462-5822.2008.01160.x

Almeida, A. J., Cunha, C., Carmona, J. A., Sampaio-Marques, B., Carvalho, A., Malavazi, I., et al. (2009). Cdc42p controls yeast-cell shape and virulence of *Paracoccidioides brasiliensis*. *Fungal Genet. Biol.* 46, 919–926. doi: 10.1016/j.fgb.2009.08.004

Almeida, F., Rodrigues, M. L., and Coelho, C. (2019). The still underestimated problem of fungal diseases worldwide. *Front. Microbiol.* 10:214. doi: 10.3389/fmicb.2019.00214

Alves, L. R., Peres da Silva, R., Sanchez, D. A., Zamith-Miranda, D., Rodrigues, M. L., Goldenberg, S., et al. (2019). Extracellular vesicle-mediated RNA Release in *Histoplasma capsulatum*. *mSphere* 4, e00176-19. doi: 10.1128/mSphere.00176-19

Andreu, Z., and Yáñez-Mó, M. (2014). Tetraspanins in extracellular vesicle formation and function. *Front. Immunol.* 5:442. doi: 10.3389/fimmu.2014.00442

Baker, E. S., Livesay, E. A., Orton, D. J., Moore, R. J., Danielson, W. F. III., Prior, D. C., et al. (2010). An LC-IMS-MS platform providing increased dynamic range for high-throughput proteomic studies. *J. Proteome Res.* 9, 997–1006. doi: 10.1021/pr900888b

Baltazar, L. M., Zamith-Miranda, D., Burnet, M. C., Choi, H., Nimrichter, L., Nakayasu, E. S., et al. (2018). Concentration-dependent protein loading of extracellular vesicles released by *Histoplasma capsulatum* after antibody treatment and its modulatory action upon macrophages. *Sci. Rep.* 8:8065. doi: 10.1038/s41598-018-25665-5

Beauvais, A., and Latgé, J. P. (2018). Special issue: fungal cell wall. *J. Fungi* 4:91. doi: 10.3390/jof4030091

Bekker-Jensen, D. B., Martínez-Val, A., Steigerwald, S., Rüther, P., Fort, K. L., Arrey, T. N., et al. (2020). A compact quadrupole-orbitrap mass spectrometer with FAIMS interface improves proteome coverage in short LC gradients. *Mol. Cell Proteom.* 19, 716–729. doi: 10.1074/mcp.TIR119.001906

Bielska, E., Sisquella, M. A., Aldeieg, M., Birch, C., O'Donoghue, E. J., and May, R. C. (2018). Pathogen-derived extracellular vesicles mediate virulence in the fatal human pathogen *Cryptococcus gattii*. *Nat. Commun.* 9:1556. doi: 10.1038/s41467-018-03991-6

Bleackley, M. R., Samuel, M., Garcia-Ceron, D., McKenna, J. A., Lowe, R. G. T., Pathan, M., et al. (2019). Extracellular vesicles from the cotton pathogen *Fusarium oxysporum* f. sp. vasinfectum induce a phytotoxic response in plants. *Front. Plant Sci.* 10:1610. doi: 10.3389/fpls.2019.01610

Brauer, V. S., Pessoni, A. M., Bitencourt, T. A., de Paula, R. G., de Oliveira Rocha, L., Goldman, G. H., et al. (2020). Extracellular vesicles from *Aspergillus flavus* Induce M1 polarization in vitro. *mSphere* 5, e00190-20. doi: 10.1128/mSphere.00190-20

Buck, A. H., Coakley, G., Simbari, F., McSorley, H. J., Quintana, J. F., Le Bihan, T., et al. (2014). Exosomes secreted by nematode parasites transfer small RNAs to mammalian cells and modulate innate immunity. *Nat. Commun.* 5:5488. doi: 10.1038/ncomms6488

Bürgel, P. H., Marina, C. L., Saavedra, P. H. V., Albuquerque, P., de Oliveira, S. A. M., and Veloso Janior, P. H. H. (2020). *Cryptococcus neoformans* secretes small molecules that inhibit IL-1β inflammasome-dependent secretion. *Mediators Inflamm.* 2020:3412763. doi: 10.1155/2020/3412763

Burnet, M. C., Zamith-Miranda, D., Heyman, H. M., Weitz, K. K., Bredeweg, E. L., Nosanchuk, J. D., et al. (2020). Remodeling of the *Histoplasma capsulatum* membrane induced by monoclonal antibodies. *Vaccines* 8:269. doi: 10.3390/vaccines8020269

Carneiro, A. B., Iaciura, B. M., Nohara, L. L., Lopes, C. D., Veas, E. M., Mariano, V. S., et al. (2013). Lysophosphatidylcholine triggers TLR2- and TLR4-mediated signaling pathways but counteracts LPS-induced NO synthesis in peritoneal macrophages by inhibiting NF-κB translocation and MAPK/ERK phosphorylation. *PLoS One* 8:e76233. doi: 10.1371/journal.pone.0076233

Childers, D. S., Avelar, G. M., Bain, J. M., Pradhan, A., Larcombe, D. E., Netea, M. G., et al. (2020). Epitope shaving promotes fungal immune evasion. *mBio* 11:e00984-20. doi: 10.1128/mBio.00984-20

Cleare, L. G., Zamith, D., Heyman, H. M., Couvillion, S. P., Nimrichter, L., Rodrigues, M. L., et al. (2020). Media matters! alterations in the loading and release of *Histoplasma capsulatum* extracellular vesicles in response to different nutritional milieus. *Cell Microbiol.* 22:e13217. doi: 10.1111/cmi.13217

Coelho, C., Brown, L., Maryam, M., Vij, R., Smith, D. F. Q., Burnet, M. C., et al. (2019). Listeria monocytogenes virulence factors, including listeriolysin O, are secreted in biologically active extracellular vesicles. *J. Biol. Chem.* 294, 1202–1217. doi: 10.1074/jbc.RA118.006472

Colby, S. M., Nuñez, J. R., Hodas, N. O., Corley, C. D., and Renslow, R. R. (2020). Deep learning to generate in silico chemical property libraries and candidate molecules for small molecule identification in complex samples. *Anal. Chem.* 92, 1720–1729. doi: 10.1021/acs.analchem.9b02348

Colby, S. M., Thomas, D. G., Nuñez, J. R., Baxter, D. J., Glaesemann, K. R., Brown, J. M., et al. (2019). ISiCLE: a quantum chemistry pipeline for establishing in silico collision cross section libraries. *Anal. Chem.* 91, 4346–4356. doi: 10.1021/acs.analchem.8b04567

Colombo, A. C., Rella, A., Normile, T., Joffe, L. S., Tavares, P. M., Araújo, G. R. S., et al. (2019). *Cryptococcus neoformans* glucuronoxylomannan and sterylglucoside are required for host protection in an animal vaccination model. *mBio* 10:e02909-18. doi: 10.1128/mBio.02909-18

Coman, C., Solari, F. A., Hentschel, A., Sickmann, A., Zahedi, R. P., and Ahrends, R. (2016). Simultaneous metabolite, protein, lipid extraction (SIMPLEX): a combinatorial multimolecular omics approach for systems biology. *Mol. Cell Proteom.* 15, 1453–1466. doi: 10.1074/mcp.M115.053702

Cortes-Serra, N., Mendes, M. T., Mazagatos, C., Segui-Barber, J., Ellis, C. C., Ballart, C., et al. (2020). Plasma-derived extracellular vesicles as potential biomarkers in heart transplant patient with chronic chagas disease. *Emerg. Infect. Dis.* 26, 1846–1851. doi: 10.3201/eid2608.191042

da Silva, T. A., Roque-Barreira, M. C., Casadevall, A., and Almeida, F. (2016). Extracellular vesicles from *Paracoccidioides brasiliensis* induced M1 polarization in vitro. *Sci. Rep.* 6:35867. doi: 10.1038/srep35867

Dawson, C. S., Garcia-Ceron, D., Rajapaksha, H., Faou, P., Bleackley, M. R., and Anderson, M. A. (2020). Protein markers for *Candida albicans* EVs include claudin-like Sur7 family proteins. *J. Extracell. Vesicles* 9:1750810. doi: 10.1080/20013078.2020.1750810

De Oliveira Frazão, S., De Sousa, H. R., Silva, L. G. D., Folha, J. D. S., Gorgonha, K. C. M., Oliveira, G. P. Jr., et al. (2020). Laccase affects the rate of *Cryptococcus neoformans* nonlytic exocytosis from macrophages. *mBio* 11:e02085-20. doi: 10.1128/mBio.02085-20

de Paula, R. G., Antoniêto, A. C. C., Nogueira, K. M. V., Ribeiro, L. F. C., Rocha, M. C., Malavazi, I., et al. (2019). Extracellular vesicles carry cellulases in the industrial fungus *Trichoderma reesei*. *Biotechnol. Biofuels* 12:146. doi: 10.1186/s13068-019-1487-7

Detre, C., Kiss, E., Varga, Z., Ludányi, K., Pászty, K., Enyedi, A., et al. (2006). Death or survival: membrane ceramide controls the fate and activation of antigen-specific T-cells depending on signal strength and duration. *Cell. Signal.* 18, 294–306. doi: 10.1016/j.cellsig.2005.05.012

Dobson, C. M. (2004). Chemical space and biology. *Nature* 432, 824–828. doi: 10.1038/nature03192

Eisenman, H. C., Frases, S., Nicola, A. M., Rodrigues, M. L., and Casadevall, A. (2009). Vesicle-associated melanization in *Cryptococcus neoformans*. *Microbiology* 155(Pt. 12), 3860–3867. doi: 10.1099/mic.0.032854-0

Erb-Downward, J. R., Noggle, R. M., Williamson, P. R., and Huffnagle, G. B. (2008). The role of laccase in prostaglandin production by *Cryptococcus neoformans*. *Mol. Microbiol.* 68, 1428–1437. doi: 10.1111/j.1365-2958.2008.06245.x

Fu, M. S., Coelho, C., De Leon-Rodriguez, C. M., Rossi, D. C. P., Camacho, E., Jung, E. H., et al. (2018). *Cryptococcus neoformans* urease affects the outcome of intracellular pathogenesis by modulating phagolysosomal PH. *PLoS Pathog.* 14:e1007144. doi: 10.1371/journal.ppat.1007144

Gazos-Lopes, F., Oliveira, M. M., Hoelz, L. V., Vieira, D. P., Marques, A. F., Nakayasu, E. S., et al. (2014). Structural and functional analysis of a platelet-activating lysophosphatidylcholine of *Trypanosoma cruzi*. *PLoS Negl. Trop. Dis.* 8:e3077. doi: 10.1371/journal.pntd.0003077

Gehrmann, U., Qazi, K. R., Johansson, C., Hultenby, K., Karlsson, M., Lundeberg, L., et al. (2011). Nanovesicles from *Malassezia sympodialis* and host exosomes induce cytokine responses – novel mechanisms for host-microbe interactions in atopic eczema. *PLoS One* 6:e21480. doi: 10.1371/journal.pone.0021480

Gil-Bona, A., Llama-Palacios, A., Parra, C. M., Vivanco, F., Nombela, C., Monteoliva, L., et al. (2015). Proteomics unravels extracellular vesicles as carriers of classical cytoplasmic proteins in *Candida albicans*. *J. Proteome Res.* 14, 142–153. doi: 10.1021/pr5007944

Giraldez, M. D., Spengler, R. M., Etheridge, A., Godoy, P. M., Barczak, A. J., Srinivasan, S., et al. (2018). Comprehensive multi-center assessment of small RNA-seq methods for quantitative miRNA profiling. *Nat. Biotechnol.* 36, 746–757. doi: 10.1038/nbt.4183

Gordon, T. R. (2017). Fusarium oxysporum and the fusarium wilt syndrome. *Annu. Rev. Phytopathol.* 55, 23–39. doi: 10.1146/annurev-phyto-080615-095919

Hansakon, A., Ngamskulrungroj, P., and Angkasekwinai, P. (2020). Contribution of laccase expression to immune response against *Cryptococcus gattii* infection. *Infect. Immun.* 88, e712–e719. doi: 10.1128/iai.00712-19

Hill, E. H., and Solomon, P. S. (2020). Extracellular vesicles from the apoplastic fungal wheat pathogen *Zymoseptoria tritici*. *Fungal Biol. Biotechnol.* 7:13. doi: 10.1186/s40694-020-00103-2

Hoshino, A., Kim, H. S., Bojmar, L., Gyan, K. E., Cioffi, M., Hernandez, J., et al. (2020). Extracellular vesicle and particle biomarkers define multiple human cancers. *Cell* 182, 1044-1061.e18. doi: 10.1016/j.cell.2020.07.009

Huang, E. L., Piehowski, P. D., Orton, D. J., Moore, R. J., Qian, W. J., Casey, C. P., et al. (2016). SNaPP: simplified nanoproteomics platform for reproducible global proteomic analysis of nanogram protein quantities. *Endocrinology* 157, 1307–1314. doi: 10.1210/en.2015-1821

Huang, S. H., Wu, C. H., Chang, Y. C., Kwon-Chung, K. J., Brown, R. J., and Jong, A. (2012). *Cryptococcus neoformans*-derived microvesicles enhance the pathogenesis of fungal brain infection. *PLoS One* 7:e48570. doi: 10.1371/journal.pone.0048570

Ikeda, M. A. K., de Almeida, J. R. F., Jannuzzi, G. P., Cronemberger-Andrade, A., Torrecilhas, A. C. T., Moretti, N. S., et al. (2018). Extracellular vesicles from *Sporothrix brasiliensis* are an important virulence factor that induce an increase in fungal burden in experimental sporotrichosis. *Front. Microbiol.* 9:2286. doi: 10.3389/fmicb.2018.02286

Johansson, H. J., Vallhov, H., Holm, T., Gehrmann, U., Andersson, A., Johansson, C., et al. (2018). Extracellular nanovesicles released from the commensal yeast *Malassezia sympodialis* are enriched in allergens and interact with cells in human skin. *Sci. Rep.* 8:9182. doi: 10.1038/s41598-018-27451-9

Johnston, G. P., Bradel-Tretheway, B., Piehowski, P. D., Brewer, H. M., Lee, B. N. R., Usher, N. T., et al. (2019). Nipah virus-like particle egress is modulated by cytoskeletal and vesicular trafficking pathways: a validated particle proteomics analysis. *mSystems* 4, e00194-19. doi: 10.1128/mSystems.00194-19

Karkowska-Kuleta, J., Kulig, K., Karnas, E., Zuba-Surma, E., Woznicka, O., Pyza, E., et al. (2020). Characteristics of extracellular vesicles released by the pathogenic yeast-like fungi *Candida glabrata*, *Candida parapsilosis* and *Candida tropicalis*. *Cells* 9:1722.

Kim, K. M., Abdelmohsen, K., Mustapic, M., Kapogiannis, D., and Gorospe, M. (2017). RNA in extracellular vesicles. *Wiley Interdiscip. Rev. RNA* 8:e1413. doi: 10.1002/wrna.1413

Konečná, K., Klimentová, J., Benada, O., Němečková, I., Janďourek, O., Jílek, P., et al. (2019). A comparative analysis of protein virulence factors released via extracellular vesicles in two *Candida albicans* strains cultivated in a nutrient-limited medium. *Microb. Pathog.* 136:103666. doi: 10.1016/j.micpath.2019.103666

Koselny, K., Mutlu, N., Minard, A. Y., Kumar, A., Krysan, D. J., and Wellington, M. (2018). A Genome-wide screen of deletion mutants in the filamentous *Saccharomyces cerevisiae* background identifies ergosterol as a direct trigger of macrophage pyroptosis. *mBio* 9:e01204-18. doi: 10.1128/mBio.01204-18

Leidal, A. M., Huang, H. H., Marsh, T., Solvik, T., Zhang, D., Ye, J., et al. (2020). The LC3-conjugation machinery specifies the loading of RNA-binding proteins into extracellular vesicles. *Nat. Cell Biol.* 22, 187–199. doi: 10.1038/s41556-019-0450-y

Leitão, N. P. Jr. (2017). *Characterization of Extracellular Vesicles Isolated From Pathogenic Fungi Cultivated Under Stress and Their Role in Cell Communication*. Ph.D. thesis, Federal University of São Paulo, São Paulo.

Leone, F., Bellani, L., Muccifora, S., Giorgetti, L., Bongioanni, P., Simili, M., et al. (2018). Analysis of extracellular vesicles produced in the biofilm by the dimorphic yeast *Pichia fermentans*. *J. Cell Physiol.* 233, 2759–2767. doi: 10.1002/jcp.25885

Lesur, A., and Domon, B. (2015). Advances in high-resolution accurate mass spectrometry application to targeted proteomics. *Proteomics* 15, 880–890. doi: 10.1002/pmic.201400450

Li, J., Van Vranken, J. G., Pontano Vaites, L., Schweppe, D. K., Huttlin, E. L., Etienne, C., et al. (2020). TMTpro reagents: a set of isobaric labeling mass tags enables simultaneous proteome-wide measurements across 16 samples. *Nat. Methods* 17, 399–404. doi: 10.1038/s41592-020-0781-4

Longo, L. V. G., da Cunha, J. P. C., Sobreira, T., and Puccia, R. (2014). Proteome of cell wall-extracts from pathogenic *Paracoccidioides brasiliensis*: comparison among morphological phases, isolates, and reported fungal extracellular vesicle proteins. *EuPA Open Proteom.* 3, 216–228. doi: 10.1016/j.euprot.2014.03.003

Marcilla, A., Martin-Jaular, L., Trelis, M., de Menezes-Neto, A., Osuna, A., Bernal, D., et al. (2014). Extracellular vesicles in parasitic diseases. *J. Extracell. Vesicles* 3:25040. doi: 10.3402/jev.v3.25040

Marcos, C. M., Silva, J. F., Oliveira, H. C., Assato, P. A., Singulani, J. L., Lopez, A. M., et al. (2016). Decreased expression of 14-3-3 in *Paracoccidioides brasiliensis* confirms its involvement in fungal pathogenesis. *Virulence* 7, 72–84. doi: 10.1080/21505594.2015.1122166

Martínez-López, R., Luisa Hernáez, M., Redondo, E., Calvo, G., Radau, S., Gil, C., et al. (2020). Small extracellular vesicles secreted by *Candida albicans* hyphae have highly diverse protein cargoes that include virulence factors and stimulate macrophages. *bioRxiv [Preprint]* doi: 10.1101/2020.10.02.323774

Martin-Jaular, L., Nakayasu, E. S., Ferrer, M., Almeida, I. C., and Del Portillo, H. A. (2011). Exosomes from *Plasmodium yoelii*-infected reticulocytes protect mice from lethal infections. *PLoS One* 6:e26588. doi: 10.1371/journal.pone.0026588

Mateescu, B., Kowal, E. J., van Balkom, B. W., Bartel, S., Bhattacharyya, S. N., Buzás, E. I., et al. (2017). Obstacles and opportunities in the functional analysis of extracellular vesicle RNA - an ISEV position paper. *J. Extracell. Vesicles* 6:1286095. doi: 10.1080/20013078.2017.1286095

Matos Baltazar, L., Nakayasu, E. S., Sobreira, T. J., Choi, H., Casadevall, A., Nimrichter, L., et al. (2016). Antibody binding alters the characteristics and contents of extracellular vesicles released by *Histoplasma capsulatum*. *mSphere* 1, e00085-15. doi: 10.1128/mSphere.00085-15

Meier, F., Brunner, A. D., Koch, S., Koch, H., Lubeck, M., Krause, M., et al. (2018). Online parallel accumulation-serial fragmentation (PASEF) with a novel trapped ion mobility mass spectrometer. *Mol. Cell Proteom.* 17, 2534–2545. doi: 10.1074/mcp.TIR118.000900

Menné, C., Lauritsen, J. P., Dietrich, J., Kastrup, J., Wegener, A. M., Odum, N., et al. (2000). Ceramide-induced TCR up-regulation. *J. Immunol.* 165, 3065–3072. doi: 10.4049/jimmunol.165.6.3065

Miura, N., and Ueda, M. (2018). Evaluation of unconventional protein secretion by *Saccharomyces cerevisiae* and other fungi. *Cells* 7:128. doi: 10.3390/cells7090128

Nakayasu, E. S., Nicora, C. D., Sims, A. C., Burnum-Johnson, K. E., Kim, Y. M., Kyle, J. E., et al. (2016). MPLEx: a robust and universal protocol for single-sample integrative proteomic, metabolomic, and lipidomic analyses. *mSystems* 1, e00043-16. doi: 10.1128/mSystems.00043-16

Nicolás, F. E., and Ruiz-Vázquez, R. M. (2013). Functional diversity of RNAi-associated sRNAs in fungi. *Int. J. Mol. Sci.* 14, 15348–15360. doi: 10.3390/ijms140815348

Nimrichter, L., de Souza, M. M., Del Poeta, M., Nosanchuk, J. D., Joffe, L., Tavares, P. M., et al. (2016). Extracellular vesicle-associated transitory cell wall components and their impact on the interaction of fungi with host cells. *Front. Microbiol.* 7:1034. doi: 10.3389/fmicb.2016.01034

Nisini, R., Romagnoli, G., Gomez, M. J., La Valle, R., Torosantucci, A., Mariotti, S., et al. (2001). Antigenic properties and processing requirements of 65-kilodalton mannoprotein, a major antigen target of anti-Candida human T-cell response, as disclosed by specific human T-cell clones. *Infect. Immun.* 69, 3728–3736. doi: 10.1128/iai.69.6.3728-3736.2001

Noverr, M. C., Williamson, P. R., Fajardo, R. S., and Huffnagle, G. B. (2004). CNLAC1 is required for extrapulmonary dissemination of *Cryptococcus neoformans* but not pulmonary persistence. *Infect. Immun.* 72, 1693–1699. doi: 10.1128/iai.72.3.1693-1699.2004

Oda, K., Kakizono, D., Yamada, O., Iefuji, H., Akita, O., and Iwashita, K. (2006). Proteomic analysis of extracellular proteins from *Aspergillus oryzae* grown under submerged and solid-state culture conditions. *Appl. Environ. Microbiol.* 72, 3448–3457. doi: 10.1128/aem.72.5.3448-3457.2006

Oh, E. J., and Jin, Y. S. (2020). Engineering of *Saccharomyces cerevisiae* for efficient fermentation of cellulose. *FEMS Yeast Res.* 20:foz089. doi: 10.1093/femsyr/foz089

Oliveira, D. L., Nakayasu, E. S., Joffe, L. S., Guimaraes, A. J., Sobreira, T. J., Nosanchuk, J. D., et al. (2010). Characterization of yeast extracellular vesicles: evidence for the participation of different pathways of cellular traffic in vesicle biogenesis. *PLoS One* 5:e11113. doi: 10.1371/journal.pone.0011113

Oliveira, F. F. M., Paes, H. C., Peconick, L. D. F., Fonseca, F. L., Marina, C. L. F., Bocca, A. L., et al. (2020). Erg6 affects membrane composition and virulence of the human fungal pathogen *Cryptococcus neoformans*. *Fungal Genet. Biol.* 140:103368. doi: 10.1016/j.fgb.2020.103368

Panepinto, J., Komperda, K., Frases, S., Park, Y. D., Djordjevic, J. T., Casadevall, A., et al. (2009). Sec6-dependent sorting of fungal extracellular exosomes and laccase of *Cryptococcus neoformans*. *Mol. Microbiol.* 71, 1165–1176. doi: 10.1111/j.1365-2958.2008.06588.x

Pang, B., Zhu, Y., Ni, J., Thompson, J., Malouf, D., Bucci, J., et al. (2020). Extracellular vesicles: the next generation of biomarkers for liquid biopsy-based prostate cancer diagnosis. *Theranostics* 10, 2309–2326. doi: 10.7150/thno.39486

Parente-Rocha, J. A., Parente, A. F., Baeza, L. C., Bonfim, S. M., Hernandez, O., McEwen, J. G., et al. (2015). Macrophage interaction with *Paracoccidioides brasiliensis* yeast cells modulates fungal metabolism and generates a response to oxidative stress. *PLoS One* 10:e0137619. doi: 10.1371/journal.pone.0137619

Park, Y. D., Chen, S. H., Camacho, E., Casadevall, A., and Williamson, P. R. (2020). Role of the ESCRT pathway in laccase trafficking and virulence of *Cryptococcus neoformans*. *Infect. Immun.* 88, e00954-19. doi: 10.1128/iai.00954-19

Peres da Silva, R., Heiss, C., Black, I., Azadi, P., Gerlach, J. Q., Travassos, L. R., et al. (2015). vesicles from *Paracoccidioides* pathogenic species transport polysaccharide and expose ligands for DC-SIGN receptors. *Sci. Rep.* 5:14213. doi: 10.1038/srep14213

Peres da Silva, R., Longo, L. G. V., Cunha, J. P. C., Sobreira, T. J. P., Rodrigues, M. L., Faoro, H., et al. (2019). Comparison of the RNA content of extracellular vesicles derived from *Paracoccidioides brasiliensis* and *Paracoccidioides lutzii*. *Cells* 8:765. doi: 10.3390/cells8070765

Peres, R., Martins, S. T., Rizzo, J., Dos Reis, F. C. G., Joffe, L. S., Vainstein, M., et al. (2018). Golgi reassembly and stacking protein (GRASP) participates in vesicle-mediated RNA export in *Cryptococcus neoformans*. *Genes* 9:400. doi: 10.3390/genes9080400

Peres, R., Puccia, R., Rodrigues, M. L., Oliveira, D. L., Joffe, L. S., Cesar, G. V., et al. (2015). Extracellular vesicle-mediated export of fungal RNA. *Sci. Rep.* 5:7763. doi: 10.1038/srep07763

Pietrella, D., Bistoni, G., Corbucci, C., Perito, S., and Vecchiarelli, A. (2006). *Candida albicans* mannoprotein influences the biological function of dendritic cells. *Cell Microbiol.* 8, 602–612. doi: 10.1111/j.1462-5822.2005.00651.x

Poad, B. L. J., Zheng, X., Mitchell, T. W., Smith, R. D., Baker, E. S., and Blanksby, S. J. (2018). Online Ozonolysis Combined with ion mobility-mass spectrometry provides a new platform for lipid isomer analyses. *Anal. Chem.* 90, 1292–1300. doi: 10.1021/acs.analchem.7b04091

Rainville, P. D., Wilson, I. D., Nicholson, J. K., Isaac, G., Mullin, L., Langridge, J. I., et al. (2017). Ion mobility spectrometry combined with ultra performance liquid chromatography/mass spectrometry for metabolic phenotyping of urine: effects of column length, gradient duration and ion mobility spectrometry on metabolite detection. *Anal. Chim. Acta* 982, 1–8. doi: 10.1016/j.aca.2017.06.020

Ramirez, M. I., Amorim, M. G., Gadelha, C., Milic, I., Welsh, J. A., Freitas, V. M., et al. (2018). Technical challenges of working with extracellular vesicles. *Nanoscale* 10, 881–906. doi: 10.1039/c7nr08360b

Raposo, G., and Stoorvogel, W. (2013). Extracellular vesicles: exosomes, microvesicles, and friends. *J. Cell Biol.* 200, 373–383. doi: 10.1083/jcb.201211138

Rayamajhi, S., and Aryal, S. (2020). Surface functionalization strategies of extracellular vesicles. *J. Mater. Chem. B* 8, 4552–4569. doi: 10.1039/d0tb00744g

Rayner, S., Bruhn, S., Vallhov, H., Andersson, A., Billmyre, R. B., and Scheynius, A. (2017). Identification of small RNAs in extracellular vesicles from the commensal yeast *Malassezia sympodialis*. *Sci. Rep.* 7:39742. doi: 10.1038/srep39742

Reis, F. C. G., Borges, B. S., Jozefowicz, L. J., Sena, B. A. G., Garcia, A. W. A., Medeiros, L. C., et al. (2019). A Novel protocol for the isolation of fungal extracellular vesicles reveals the participation of a putative scramblase in polysaccharide export and capsule construction in *Cryptococcus gattii*. *mSphere* 4, e00080-19. doi: 10.1128/mSphere.00080-19

Riquelme, M., Bredeweg, E. L., Callejas-Negrete, O., Roberson, R. W., Ludwig, S., Beltrán-Aguilar, A., et al. (2014). The *Neurospora crassa* exocyst complex tethers Spitzenkörper vesicles to the apical plasma membrane during polarized growth. *Mol. Biol. Cell* 25, 1312–1326. doi: 10.1091/mbc.E13-06-0299

Rittershaus, P. C., Kechichian, T. B., Allegood, J. C., Merrill, A. H. Jr., Hennig, M., Luberto, C., et al. (2006). Glucosylceramide synthase is an essential regulator of pathogenicity of *Cryptococcus neoformans*. *J. Clin. Invest.* 116, 1651–1659. doi: 10.1172/JCI27890

Rizzo, J., Chaze, T., Miranda, K., Roberson, R. W., Gorgette, O., Nimrichter, L., et al. (2020a). Characterization of extracellular vesicles produced by *Aspergillus fumigatus* protoplasts. *mSphere* 5, e00476-20. doi: 10.1128/mSphere.00476-20

Rizzo, J., Rodrigues, M. L., and Janbon, G. (2020b). Extracellular vesicles in fungi: past, present, and future perspectives. *Front. Cell Infect. Microbiol.* 10:346. doi: 10.3389/fcimb.2020.00346

Rizzo, J., Wong, S. S. W., Gazi, A. D., Moyrand, F., Chaze, T., Commere, P. H., et al. (2020c). New insights into *Cryptococcus* extracellular vesicles suggest a new structural model and an antifungal vaccine strategy. *bioRxiv* [Preprint] doi: 10.1101/2020.08.17.253716

Rizzoli, S. O., and Jahn, R. (2007). Kiss-and-run, collapse and 'readily retrievable' vesicles. *Traffic* 8, 1137–1144. doi: 10.1111/j.1600-0854.2007.00614.x

Rodrigues, M. L., Franzen, A. J., Nimrichter, L., and Miranda, K. (2013). Vesicular mechanisms of traffic of fungal molecules to the extracellular space. *Curr. Opin. Microbiol.* 16, 414–420. doi: 10.1016/j.mib.2013.04.002

Rodrigues, M. L., Nakayasu, E. S., Oliveira, D. L., Nimrichter, L., Nosanchuk, J. D., Almeida, I. C., et al. (2008). Extracellular vesicles produced by *Cryptococcus neoformans* contain protein components associated with virulence. *Eukaryot. Cell* 7, 58–67. doi: 10.1128/EC.00370-07

Rodrigues, M. L., Nimrichter, L., Oliveira, D. L., Frases, S., Miranda, K., Zaragoza, O., et al. (2007). Vesicular polysaccharide export in *Cryptococcus neoformans* is a eukaryotic solution to the problem of fungal trans-cell wall transport. *Eukaryot. Cell* 6, 48–59. doi: 10.1128/EC.00318-06

Rodrigues, M. L., Travassos, L. R., Miranda, K. R., Franzen, A. J., Rozental, S., de Souza, W., et al. (2000). Human antibodies against a purified glucosylceramide from *Cryptococcus neoformans* inhibit cell budding and fungal growth. *Infect. Immun.* 68, 7049–7060. doi: 10.1128/iai.68.12.7049-7060.2000

Rosa-Fernandes, L., Rocha, V. B., Carregari, V. C., Urbani, A., and Palmisano, G. (2017). A perspective on extracellular vesicles proteomics. *Front. Chem.* 5:102. doi: 10.3389/fchem.2017.00102

Sanders, K. L., and Edwards, J. L. (2020). Nano-liquid chromatography-mass spectrometry and recent applications in omics investigations. *Anal. Methods* 12, 4404–4417. doi: 10.1039/d0ay01194k

Soehnlein, O., Lindbom, L., and Weber, C. (2009). Mechanisms underlying neutrophil-mediated monocyte recruitment. *Blood* 114, 4613–4623. doi: 10.1182/blood-2009-06-221630

Son, H., Park, A. R., Lim, J. Y., Shin, C., and Lee, Y. W. (2017). Genome-wide exonic small interference RNA-mediated gene silencing regulates sexual reproduction in the homothallic fungus *Fusarium graminearum*. *PLoS Genet.* 13:e1006595. doi: 10.1371/journal.pgen.1006595

Souza, J. A. M., Baltazar, L. M., Carregal, V. M., Gouveia-Eufrasio, L., Oliveira, A. G. D., Dias, W. G., et al. (2019). Characterization of *Aspergillus fumigatus* extracellular vesicles and their effects on macrophages and neutrophils functions. *Front. Microbiol.* 10:2008. doi: 10.3389/fmicb.2019.02008

Stahl, P. D., and Raposo, G. (2019). Extracellular vesicles: exosomes and microvesicles, integrators of homeostasis. *Physiology* 34, 169–177. doi: 10.1152/physiol.00045.2018

Strimbu, K., and Tavel, J. A. (2010). What are biomarkers? *Curr. Opin. HIV AIDS* 5, 463–466. doi: 10.1097/COH.0b013e32833ed177

Szempruch, A. J., Dennison, L., Kieft, R., Harrington, J. M., and Hajduk, S. L. (2016). Sending a message: extracellular vesicles of pathogenic protozoan parasites. *Nat. Rev. Microbiol.* 14, 669–675. doi: 10.1038/nrmicro.2016.110

Takeo, K., Uesaka, I., Uehira, K., and Nishiura, M. (1973). Fine structure of *Cryptococcus neoformans* grown in vivo as observed by freeze-etching. *J. Bacteriol.* 113, 1449–1454. doi: 10.1128/jb.113.3.1449-1454.1973

Tamayo, D., Muñoz, J. F., Almeida, A. J., Puerta, J. D., Restrepo, Á, Cuomo, C. A., et al. (2017). *Paracoccidioides* spp. catalases and their role in antioxidant defense against host defense responses. *Fungal Genet. Biol.* 100, 22–32. doi: 10.1016/j.fgb.2017.01.005

Tamayo, D., Muñoz, J. F., Lopez, Á, Urán, M., Herrera, J., Borges, C. L., et al. (2016). Identification and Analysis of the role of superoxide dismutases isoforms in the pathogenesis of *Paracoccidioides* spp. *PLoS Negl. Trop. Dis.* 10:e0004481. doi: 10.1371/journal.pntd.0004481

Torres, I., Hernandez, O., Tamayo, D., Muñoz, J. F., Leitão, N. P. Jr., García, A. M., et al. (2013). Inhibition of PbGP43 expression may suggest that gp43 is a virulence factor in *Paracoccidioides brasiliensis*. *PLoS One* 8:e68434. doi: 10.1371/journal.pone.0068434

Toyofuku, M., Nomura, N., and Eberl, L. (2019). Types and origins of bacterial membrane vesicles. *Nat. Rev. Microbiol.* 17, 13–24. doi: 10.1038/s41579-018-0112-2

Tsai, C. F., Zhao, R., Williams, S. M., Moore, R. J., Schultz, K., Chrisler, W. B., et al. (2020). An improved boosting to amplify signal with isobaric labeling (iBASIL) strategy for precise quantitative single-cell proteomics. *Mol. Cell Proteom.* 19, 828–838. doi: 10.1074/mcp.RA119.001857

Vallejo, M. C., Nakayasu, E. S., Longo, L. V., Ganiko, L., Lopes, F. G., Matsuo, A. L., et al. (2012a). Lipidomic analysis of extracellular vesicles from the pathogenic phase of *Paracoccidioides brasiliensis*. *PLoS One* 7:e39463. doi: 10.1371/journal.pone.0039463

Vallejo, M. C., Nakayasu, E. S., Matsuo, A. L., Sobreira, T. J., Longo, L. V., Ganiko, L., et al. (2012b). Vesicle and vesicle-free extracellular proteome of *Paracoccidioides brasiliensis*: comparative analysis with other pathogenic fungi. *J. Proteome Res.* 11, 1676–1685. doi: 10.1021/pr200872s

Vargas, G., Honorato, L., Guimaraes, A. J., Rodrigues, M. L., Reis, F. C. G., Vale, A. M., et al. (2020). Protective effect of fungal extracellular vesicles against murine candidiasis. *Cell. Microbiol.* 22:e13238. doi: 10.1111/cmi.13238

Vargas, G., Rocha, J. D., Oliveira, D. L., Albuquerque, P. C., Frases, S., Santos, S. S., et al. (2015). Compositional and immunobiological analyses of extracellular vesicles released by *Candida albicans*. *Cell Microbiol.* 17, 389–407. doi: 10.1111/cmi.12374

Volke-Sepulveda, T., Salgado-Bautista, D., Bergmann, C., Wells, L., Gutierrez-Sanchez, G., and Favela-Torres, E. (2016). Secretomic insight into glucose metabolism of *Aspergillus brasiliensis* in solid-state fermentation. *J. Proteome Res.* 15, 3856–3871. doi: 10.1021/acs.jproteome.6b00663

Walker, L., Sood, P., Lenardon, M. D., Milne, G., Olson, J., Jensen, G., et al. (2018). The viscoelastic properties of the fungal cell wall allow traffic of ambisome as intact liposome vesicles. *mBio* 9:e02383-17. doi: 10.1128/mBio.02383-17

Woith, E., Fuhrmann, G., and Melzig, M. F. (2019). Extracellular vesicles-connecting kingdoms. *Int. J. Mol. Sci.* 20:5695. doi: 10.3390/ijms20225695

Wolf, J. M., Espadas-Moreno, J., Luque-Garcia, J. L., and Casadevall, A. (2014). Interaction of *Cryptococcus neoformans* extracellular vesicles with the cell wall. *Eukaryot. Cell* 13, 1484–1493. doi: 10.1128/ec.00111-14

Yeri, A., Courtright, A., Danielson, K., Hutchins, E., Alsop, E., Carlson, E., et al. (2018). Evaluation of commercially available small RNASeq library preparation kits using low input RNA. *BMC Genomics* 19:331. doi: 10.1186/s12864-018-4726-6

Zamith-Miranda, D., Heyman, H. M., Cleare, L. G., Couvillion, S. P., Clair, G., Bredeweg, E. L., et al. (2019). Multi-omics signature of *Candida auris*, an emerging and multidrug-resistant pathogen. *mSystems* 4, e257–e219. doi: 10.1128/mSystems.00257-19

Zamith-Miranda, D., Heyman, H. M., Couvillion, S. P., Cordero, R. J. B., Rodrigues, M. L., Nimrichter, L., et al. (2020). Comparative molecular and immunoregulatory analysis of extracellular vesicles from *Candida albicans* and *Candida auris*. *bioRxiv* [Preprint] doi: 10.1101/2020.11.04.368472

Zarnowski, R., Sanchez, H., Covelli, A. S., Dominguez, E., Jaromin, A., Bernhardt, J., et al. (2018). *Candida albicans* biofilm–induced vesicles confer drug resistance through matrix biogenesis. *PLoS Biol.* 16:e2006872. doi: 10.1371/journal.pbio.2006872

Zhang, X., Lan, Y., Xu, J., Quan, F., Zhao, E., Deng, C., et al. (2019). CellMarker: a manually curated resource of cell markers in human and mouse. *Nucleic Acids Res.* 47, D721–D728. doi: 10.1093/nar/gky900

Zhang, X., Romm, M., Zheng, X., Zink, E. M., Kim, Y. M., Burnum-Johnson, K. E., et al. (2016). SPE-IMS-MS: an automated platform for sub-sixty second surveillance of endogenous metabolites and xenobiotics in biofluids. *Clin. Mass Spectrom.* 2, 1–10. doi: 10.1016/j.clinms.2016.11.002

Zhao, K., Bleackley, M., Chisanga, D., Gangoda, L., Fonseka, P., Liem, M., et al. (2019). Extracellular vesicles secreted by *Saccharomyces cerevisiae* are involved in cell wall remodelling. *Commun. Biol.* 2:305. doi: 10.1038/s42003-019-0538-8

Zheng, X., Smith, R. D., and Baker, E. S. (2018). Recent advances in lipid separations and structural elucidation using mass spectrometry combined with ion mobility spectrometry, ion-molecule reactions and fragmentation approaches. *Curr. Opin. Chem. Biol.* 42, 111–118. doi: 10.1016/j.cbpa.2017.11.009

Zhu, Y., Piehowski, P. D., Zhao, R., Chen, J., Shen, Y., Moore, R. J., et al. (2018). Nanodroplet processing platform for deep and quantitative proteome profiling of 10-100 mammalian cells. *Nat. Commun.* 9:882. doi: 10.1038/s41467-018-03367-w

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

frontiers in Genetics

ORIGINAL RESEARCH
published: 18 May 2021
doi: 10.3389/fgene.2021.648701

# Multiomics Analysis Reveals Molecular Abnormalities in Granulosa Cells of Women With Polycystic Ovary Syndrome

*Rusong Zhao*[1,2,3,4], *Yonghui Jiang*[1,2,3,4], *Shigang Zhao*[1,2,3,4]* *and Han Zhao*[1,2,3,4]*

[1] Center for Reproductive Medicine, Cheeloo College of Medicine, Shandong University, Jinan, China, [2] National Research Center for Assisted Reproductive Technology and Reproductive Genetics, Shandong University, Jinan, China, [3] Key Laboratory of Reproductive Endocrinology of Ministry of Education, Shandong University, Jinan, China, [4] Shandong Provincial Clinical Medicine Research Center for Reproductive Health, Shandong University, Jinan, China

Polycystic ovary syndrome (PCOS) is the most common complex endocrine and metabolic disease in women of reproductive age. It is characterized by anovulatory infertility, hormone disorders, and polycystic ovarian morphology. Regarding the importance of granulosa cells (GCs) in the pathogenesis of PCOS, few studies have investigated the etiology at a single "omics" level, such as with an mRNA expression array or methylation profiling assay, but this can provide only limited insights into the biological mechanisms. Here, genome-wide DNA methylation together with lncRNA-miRNA-mRNA profiles were simultaneously detected in GCs of PCOS cases and controls. A total of 3579 lncRNAs, 49 miRNAs, 669 mRNAs, and 890 differentially methylated regions (DMR)-associated genes were differentially expressed between PCOS cases and controls. Pathway analysis indicated that these differentially expressed genes were commonly associated with steroid biosynthesis and metabolism-related signaling, such as glycolysis/gluconeogenesis. In addition, we constructed ceRNA networks and identified some known ceRNA axes, such as lncRNAs-miR-628-5p-CYP11A1/HSD17B7. We also identified many new ceRNA axes, such as lncRNAs-miR-483-5p-GOT2. Interestingly, most ceRNA axes were also closely related to steroid biosynthesis and metabolic pathways. These findings suggest that it is important to systematically consider the role of reproductive and metabolic genes in the pathogenesis of PCOS.

**Keywords: polycystic ovary syndrome, methylome, transcriptome, metabolism, steroid biosynthesis**

OPEN ACCESS

***Edited by:***
*Sanjay Kumar Banerjee, National Institute of Pharmaceutical Education and Research, India*

***Reviewed by:***
*Nicola Bernabò, University of Teramo, Italy*
*Leda Torres, National Institute of Pediatrics, Mexico*

***Correspondence:**
*Shigang Zhao*
*zsg0108@126.com*
*Han Zhao*
*hanzh80@yahoo.com*

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

***Received:*** *01 January 2021*
***Accepted:*** *06 April 2021*
***Published:*** *18 May 2021*

***Citation:***
*Zhao RS, Jiang YH, Zhao SG and Zhao H (2021) Multiomics Analysis Reveals Molecular Abnormalities in Granulosa Cells of Women With Polycystic Ovary Syndrome. Front. Genet. 12:648701. doi: 10.3389/fgene.2021.648701*

## INTRODUCTION

Polycystic ovary syndrome (PCOS) is a life-long reproductive, neuroendocrine, and metabolic disorder that affects up to 6–15% of women of reproductive age (Risal et al., 2019). Its main clinical manifestations are ovulatory dysfunction, hyperandrogenemia, and polycystic ovaries, which can lead to infertility (Fauser et al., 2012). In addition to the above reproductive disorders, PCOS is often accompanied by metabolic abnormalities, such as insulin resistance (IR). IR can increase pituitary luteinizing hormone (LH) secretion, testosterone secretion in theca cells, and P450scc activity in

granulosa cells (GCs), which interferes with follicle maturation and leads to the development of PCOS (Li et al., 2019). Studies have shown that the abnormal ovarian hormone production is mainly attributed to the hypertrophy of follicular theca cells and the altered expression of key steroid biosynthesis enzymes in GCs (Aste et al., 1998).

As the most abundant cells in the ovary, GCs are closely associated with the development of oocytes and play an essential role in both normal folliculogenesis and steroidogenesis (Hummitzsch et al., 2015). Previous studies have shown that cumulus and mural GCs contribute to the process of oocyte maturation by tight regulation and controlled changes in steroid hormones in the pathogenesis of PCOS (Holesh et al., 2020). Oocytes lack the capacity to carry out some metabolic processes, such as glycolysis and amino acid uptake. They rely on GCs to deliver nutrients and remove waste. In addition, the metabolic profile of GCs is associated with the fate of their accompanying oocyte (Gioacchini et al., 2018; Yilmaz et al., 2018; Fontana et al., 2020).

Previous studies have separately screened differentially expressed mRNAs, miRNAs (DEMs) and lncRNAs (DELs) in GCs to explore the regulatory mechanism of PCOS. Few studies have indicated that genome-wide DNA methylation changes may affect the expression of different genes in PCOS ovaries, as revealed by methylated DNA immunoprecipitation (MeDIP) experiments (Yu et al., 2015; Xu et al., 2016). However, to date, no studies have been performed to identify the whole transcriptome and methylome in same GCs of women with PCOS. In this study, the lncRNA-miRNA-mRNA expression profiles and DNA methylation of GCs in PCOS were comprehensively analyzed. Our goal was to integrate multiomics data to identify differentially expressed genes (DEGs) and differentially methylated regions (DMRs) in PCOS and to construct molecular networks that could help us to better understand the etiology of PCOS.

# MATERIALS AND METHODS

## Sample Selection

Five women with PCOS and five age/body mass index (BMI)-matched control subjects from the Center for Reproductive Medicine, Shandong University, Jinan, China, were included in this study. All patients gave informed consent, and the study was approved by the institutional review board of the Reproductive Hospital Affiliated to Shandong University. The definition of PCOS was based on the 2003 Rotterdam criteria (Rotterdam ESHRE/ASRM-Sponsored PCOS Consensus Workshop Group, 2004). Women considered as suffering from PCOS had at least two of the following three characteristics: polycystic ovaries on ultrasound, irregularity/absence of menses, and hyperandrogenism. Cases with congenital adrenal hyperplasia, androgen-secreting tumors, Cushing's syndrome, thyroid disease, and hyperprolactinemia were excluded. The control group was selected from healthy women who attended the center for IVF with their husbands due to a male factor. All relevant clinical information was obtained from the Electronic Medical Records System. BMI was calculated as weight (kg)/height$^2$ (m). Peripheral blood was collected on the 3rd to 5th days of the menstrual cycle to measure serum hormone levels. The levels of follicle-stimulating hormone (FSH), LH, estrogen (E2), prolactin (PRL), and testosterone (T) were measured with a chemiluminescence analyzer (Beckman Coulter, United States). Type B ultrasound was used to determine the antral follicle counts (AFC).

## Retrieval of GCs

GCs were collected from follicular fluid obtained via ultrasound-guided transvaginal oocyte retrieval after informed consent had been given by the patients who received the long gonadotropin-releasing hormone agonist protocol. Oocyte retrieval was performed 36 h after human chorionic gonadotropin (hCG) injection by transvaginal ultrasound-guided needle puncture for follicles >15 mm in diameter. At the time of oocyte retrieval, follicular fluid aspirates were collected in sterile tubes and centrifuged. GCs were isolated and purified from the follicular fluid with Ficoll-Percoll (Solarbio, Beijing, China) as previously described (Iwase et al., 2009), and then immediately stored at −80°C for further analysis.

## RNA−Seq Analysis and Quality Assessment

Total RNA was extracted using TRIzol Reagent (Invitrogen, CA, United States) and purified using an RNeasy Mini Kit (Qiagen, CA, United States). The quality of RNA was assessed using an Agilent 2100 Bioanalyzer (Agilent, Palo Alto, CA, United States). The rRNA-depleted RNA samples were further processed in accordance with the Illumina protocol (New England Biolabs, Massachusetts, United States). After cDNA synthesis, the samples were sequenced with an Illumina HiSeq 2500 using the paired-end (PE) sequencing strategy. The raw data were recorded. The overall quality of the RNA−seq data was evaluated by FastQC. Clean reads were aligned to the reference genome (Ensembl release 95, *Homo sapiens*) using TopHat2 (v2.0.14) (Kim et al., 2013) with the default parameters.

## MeDIP-Seq Library Construction

MeDIP is a method for immunoprecipitating the methylated portion of the genome using an antibody capable of recognizing 5mC (Wilson and Beck, 2016). Following the manufacturer's instructions, MeDIP was performed to analyze genome-wide methylation using the Zymo Research DNA Methylation IP Kit (Cat #D5101; Zymo Research, CA, United States). Immunoprecipitated DNA was PCR-amplified, purified, quantified, and sequenced on the Illumina HiSeq 2500 platform. MeDIP−seq reads were mapped to the human genome using BWA software (Li et al., 2012). MACS2 was used to call peaks. To study the DNA methylation differences between two groups, DMRs were identified using the Cummerbund (Trapnell et al., 2012) and ChIPpeakAnno (Zhu et al., 2010) packages in R. Briefly, DMRs were assigned to genomic regions based on gene annotations available from JGI and in-house repeat annotation in GFF3 format. The following gene regions were included:

3'UTR, 5'UTR, promoter, coding DNA sequence (CDS), intron, upstream 1 kb, and downstream 1 kb.

## Screening and Clustering Analysis of Differentially Expressed mRNAs, DELs, and DEMs

Data preprocessing and follow-up analysis were performed in the R programming environment (version 3.6.1), and Bioconductor packages were applied for the analysis of DEGs. The lists of DEGs, DELs, and DEMs between controls and PCOS cases were generated using the edgeR package (version 3.32.0) (Robinson et al., 2010). To normalize the raw data, log-fold change $|(\log FC)| > 1.2$ (mRNA and miRNA), $|(\log FC)| > 2.0$ (lncRNA) and $p$-value $<0.05$ were considered to indicate statistically significant differences between the PCOS and control groups. To generate an overview of the lncRNA, miRNA and mRNA expression profiles and compare them between the two groups, hierarchical clustering analysis was performed based on the expression levels of all transcripts and significantly differentially expressed transcripts using the pheatmap R package based on Euclidean distance.

## lncRNA, miRNA, and mRNA Prediction and Coexpression Network Construction

The miRNA target genes were predicted using the prediction results of the TargetScan and miRcode (Jeggari et al., 2012) databases. The potential target genes transcribed within a 10-kb region upstream or downstream of the lncRNAs were paired and predicted using the UCSC Genome Browser[1] (Song et al., 2019). The expression of differentially expressed mRNAs, DEMs and DELs was analyzed by Pearson's correlation coefficient using the stats and pcaPP R packages. The miRNA-mRNA network and the lncRNA-mRNA coexpression network were constructed based on analysis of the correlations among the differentially expressed mRNAs, DEMs, and DELs. A p-value of $<0.05$ for the miRNA-mRNA network and one of $<0.01$ for the lncRNA-mRNA network were considered statistically significant. The target genes that overlapped with the DEGs were then identified and used to construct the miRNA-mRNA network and lncRNA-mRNA network using Cytoscape software (version 3.8.1) (Saito et al., 2012).

## Functional Enrichment Analysis

The Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analysis in each module and network was conducted using the Database for Annotation, Visualization and Integrated Discover (DAVID). DEGs and enriched pathways were mapped using KEGG pathway annotation with KOBAS3.0[2]. The top 10 KEGG pathways were selected and ranked by the enrichment factor. To perform literature-based functional analysis, a total of 370 follicle development- and 437 steroid metabolism-related genes were obtained from the Ovarian Kaleidoscope Database (OKDB)[3]. To identify key transcription factors (TFs), a total of 1496 human TFs were obtained from the Human Protein Atlas Database (HPA)[4]. Subsequently, the Venn diagram tool was used to help identify the common genes that were the focus of this work.

## Statistical Analysis

Regarding the clinical characteristics of PCOS patients and controls, quantitative variables are expressed as the mean ± SD. $P < 0.05$ was considered significant. The clinical data analyses were performed with Statistical Package for Social Science (SPSS 25.0; IBM Corp, Armonk, NY, United States).

# RESULTS

## Clinical Features

**Table 1** presents the basic statistics of both PCOS and control subjects regarding the most important characteristics, such as FSH, LH, $E_2$, T, P, PRL, and AFC levels, as well as age and BMI. Significant differences between the two groups were found for LH, T, and AFC, all of which had higher levels in PCOS cases ($P < 0.05$).

## Differential Expression Analysis

To identify the DEGs, GCs from five healthy women and five women with PCOS were studied. As indicated in **Figure 1A**, a correlation plot was used to determine the correlation between samples and to verify the homogeneity between biological replicates. As presented in the histogram in **Figure 1B**, 669 mRNAs, 49 miRNAs and 3579 lncRNAs were differentially expressed between the PCOS and control groups. Among them, 546 and 123 mRNAs, 31 and 18 miRNAs, and 2226 and 1353 lncRNAs were upregulated and downregulated in PCOS, respectively. Hierarchical clustering heatmaps of the differentially expressed RNAs are shown in **Figures 1C,E,G**. All differentially expressed mRNAs, DEMs and DELs are

[1]http://genome.ucsc.edu/

[2]http://kobas.cbi.pku.edu.cn/kobas3/

[3]http://okdb.appliedbioinfo.net/

[4]https://www.proteinatlas.org/

**TABLE 1 |** Clinical characteristics of women with polycystic ovary syndrome and controls.

| | PCOS | Control | *P*-value |
|---|---|---|---|
| Age, years | 29 ± 1.0 | 28.4 ± 2.07 | 0.576 |
| BMI, kg/m$^2$ | 22.23 ± 1.87 | 22.26 ± 2.00 | 0.984 |
| FSH, IU/L | 6.53 ± 0.87 | 6.46 ± 0.51 | 0.874 |
| LH, IU/L | 12.20 ± 3.59 | 4.32 ± 2.43 | 0.004 |
| $E_2$, pg/ml | 55.14 ± 18.34 | 34.35 ± 14.27 | 0.08 |
| PRL, ng/ml | 28.50 ± 26.10 | 18.31 ± 10.51 | 0.442 |
| T, ng/dl | 52.42 ± 8.14 | 15.07 ± 9.40 | 0.001 |
| AFC, *n* | 38.0 ± 16.40 | 17.00 ± 1.41 | 0.045 |

*Data are presented as the mean ± SD.*
*BMI, body mass index; LH, luteinizing hormone; FSH, follicle-stimulating hormone; T, testosterone; AFC, antral follicle counts.*

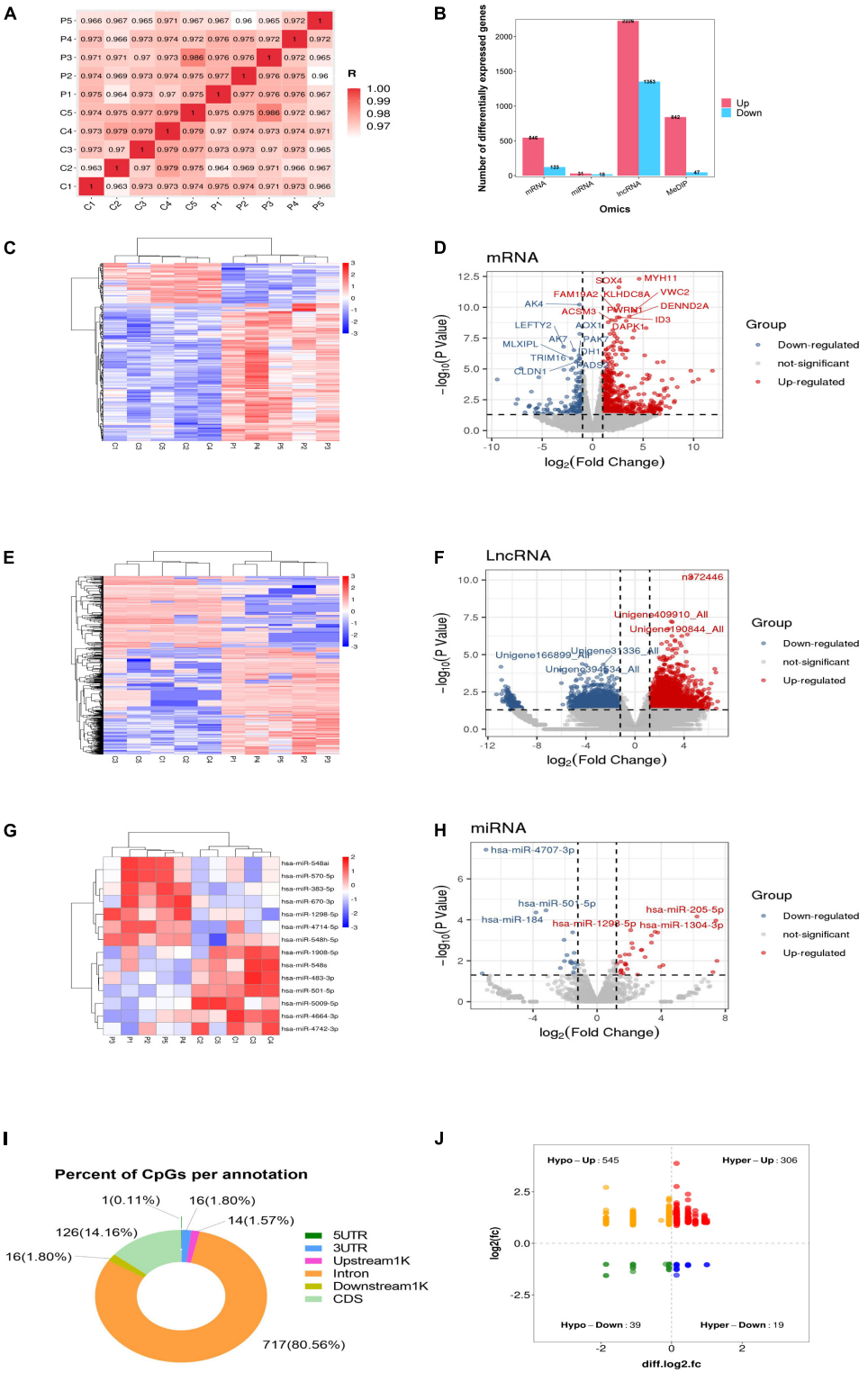



**FIGURE 1 |** Global differentially expressed mRNAs, lncRNAs, miRNAs, and differentially methylated regions (DMRs) identified in PCOS and control granulosa cells. **(A)** Correlation heatmap between PCOS and control samples. **(B)** The numbers of differentially expressed mRNAs, differentially expressed lncRNAs (DELs) and differentially expressed miRNAs (DEMs). **(C)** Hierarchical clustering presentation of DEGs in the PCOS and control groups. **(D)** Volcano plot of DEGs in the PCOS and control groups. **(E)** Hierarchical clustering presentation of DELs in the PCOS and control groups. **(F)** Volcano plot of DELs in the PCOS and control groups. **(G)** Hierarchical clustering presentation of DEMs in the PCOS and control groups. **(H)** Volcano plot of DEMs in the PCOS and control groups. **(I)** The distribution of DMRs. **(J)** The correlation of DMR-associated genes and mRNA expression.

listed in **Supplementary Table 1**. The volcano plots showed the differential expression of mRNAs, miRNAs and lncRNAs between the PCOS group and control group (**Figures 1D,F,H**). DNA methylation analysis of the MeDIP-seq data showed 890 CpG sites that were differentially methylated in PCOS GCs compared with control GCs (**Supplementary Table 1**). In terms of the gene structures associated with the CpG sites, the proportions of CDS, intron, downstream 1 kb, upstream 1 kb, 3'UTR, and 5'UTR were 126 (14.16%), 717 (80.56%), 16 (1.8%), 14 (1.57%), 16 (1.8%), and 1 (0.11%), respectively (**Figure 1I**). We identified 545 hypomethylated and upregulated genes, 19 hypermethylated and downregulated genes, 306 hypermethylated and upregulated genes, and 39 hypomethylated and downregulated genes by integrating the DNA methylation and gene expression data (**Figure 1J**). Moreover, the chromosomal locations of the DMRs were examined, and they were found to be present on all chromosomes except the Y chromosome (**Supplementary Figure 1**).

## Functional Enrichment Analysis of DEGs, DELs, DEMs, and DMR-Associated Genes

To investigate the key pathways, the DEGs, DELs, DEMs and DMR-associated genes were evaluated and compared in terms of potential functional pathways in the KEGG database (**Supplementary Table 2**). As shown in **Figure 2A**, the results revealed that the DEGs were mainly involved in steroid biosynthesis and many metabolism-related pathways, such as type II diabetes mellitus, glycolysis/gluconeogenesis, carbon metabolism, biosynthesis of amino acids, HIF-1 signaling. Combining the known genes with human TFs, we found that the expression of FOXA1, HIF3A, and STMN1 was upregulated in PCOS GCs (**Supplementary Figure 2A**). The most significantly enriched biological functions of these TFs were the TGF-beta signaling pathway, IL-17 signaling pathway, endocrine resistance, human T-cell leukemia virus 1 infection, estrogen signaling pathway, Toll-like receptor signaling pathway, IR, GnRH secretion and TNF signaling pathway (**Supplementary Figure 2B**). For the known genes related to follicle development and steroid metabolism, AMH, FSHR, ESR2, DDX4, and SMAD9 expression was upregulated in the GCs of PCOS patients, while INHA, SOD2, and CYP11A1 expression was downregulated (**Supplementary Figures 2C,D**). All these functions and pathways have been proven to be closely correlated with the pathogenesis of PCOS.

To further study the role and potential mechanisms of DELs, we identified 124 of their target mRNAs. KEGG analysis identified a total of 40 significantly enriched pathways, including metabolic pathways and steroid biosynthesis (**Figure 2B**). Notably, the metabolic pathways included carbon metabolism, biosynthesis of amino acids, IR, HIF-1 signaling pathway, terpenoid backbone biosynthesis and glycolysis/gluconeogenesis. Analysis of DEM target genes also identified a number of pathways. Further investigation by KEGG revealed that these miRNAs participated in the regulation of metabolism and steroid synthesis, such as 2-oxocarboxylic acid metabolism, glycerolipid metabolism, arginine and proline metabolism, the FoxO signaling pathway, ovarian steroidogenesis and steroid hormone biosynthesis (**Figure 2C**). The DMR-associated genes were also involved in metabolic pathways and steroid pathways. Specifically, these genes were associated with glycosaminoglycan biosynthesis, collecting duct acid secretion, bacterial invasion of epithelial cells, adherens junction, steroid biosynthesis, biosynthesis of unsaturated fatty acids and the notch signaling pathway (**Figure 2D**). Notably, all four omics enrichment analyses identified the steroid biosynthesis pathway. In addition, all three RNA omics analyses revealed the enrichment of metabolic pathways, which are key players in steroidogenesis by acting as a source of energy and substrate for steroid production. These multiomics enrichment results suggest a common etiology of abnormal metabolism and abnormal ovarian steroid formation in GCs of women with PCOS.

## Construction of a lncRNA-miRNA-mRNA ceRNA Network

According to the predicted correlations among lncRNAs, miRNAs and mRNAs, a competing endogenous RNA (ceRNA) network was constructed using ceRNA mechanism analysis. The miRNA-mRNA coexpression network was constructed based on the correlation analysis between the DEGs and DEMs. A total of 67 differentially expressed target genes were predicted for 13 DEMs, which were used to construct the miRNA-mRNA coexpression network (**Figure 3A**). This network included 10 interactions and was associated with metabolic pathways, as determined by searching the KEGG database. The interactions included hsa-miR-548i-SOD2/IDH1, hsa-miR-500a-5p-NSDHL, hsa-miR-483-5p-GOT2, and hsa-miR-214-5p-BRCA1/MKI67. Among all DEM interactions in this regulatory network, FOXO1-hsa-miR-324-5p to DGKA-hsa-miR-148b-5p, FAM160A1-hsa-miR-628-5p, and HOMER2-hsa-miR-130b-5p-PRLR were revealed to represent continuous network connections. Similar to the miRNA-mRNA network, the lncRNA-mRNA coexpression network was constructed based on analysis of the correlation between the DELs and DEGs. In total, 34 lncRNAs and 112 mRNAs involved in 326 interactions were selected to generate the network map (**Figure 3B**).

There were 217 nodes in the ceRNA network, which consisted of 79 lncRNAs, 6 miRNAs and 11 mRNAs, forming 8 pathways (**Supplementary Table 3** and **Figure 3C**). The top five pathways of lncRNAs, miRNAs and mRNAs are displayed in **Figure 3D**, indicating the important biological significance of these molecules. KEGG pathway analysis was also performed to determine the involvement of coexpressed genes in different biological pathways. Five pathways overlapped with the enriched genes in the integrated ceRNA network, namely, glycerolipid metabolism, metabolic pathways, biosynthesis of unsaturated fatty acids, steroid biosynthesis, and peroxisome. For example, the AY603498-hsa-miR-628-5p-CYP11A1 and BC036229-hsa-miR-628-5p-HSD17B7 ceRNA axes, which contribute to steroid hormone biosynthesis, were downregulated in PCOS. The AK097578-hsa-miR-548i-IDH1 and AK128202-hsa-miR-483-5p-GOT2 networks were also identified to be associated with

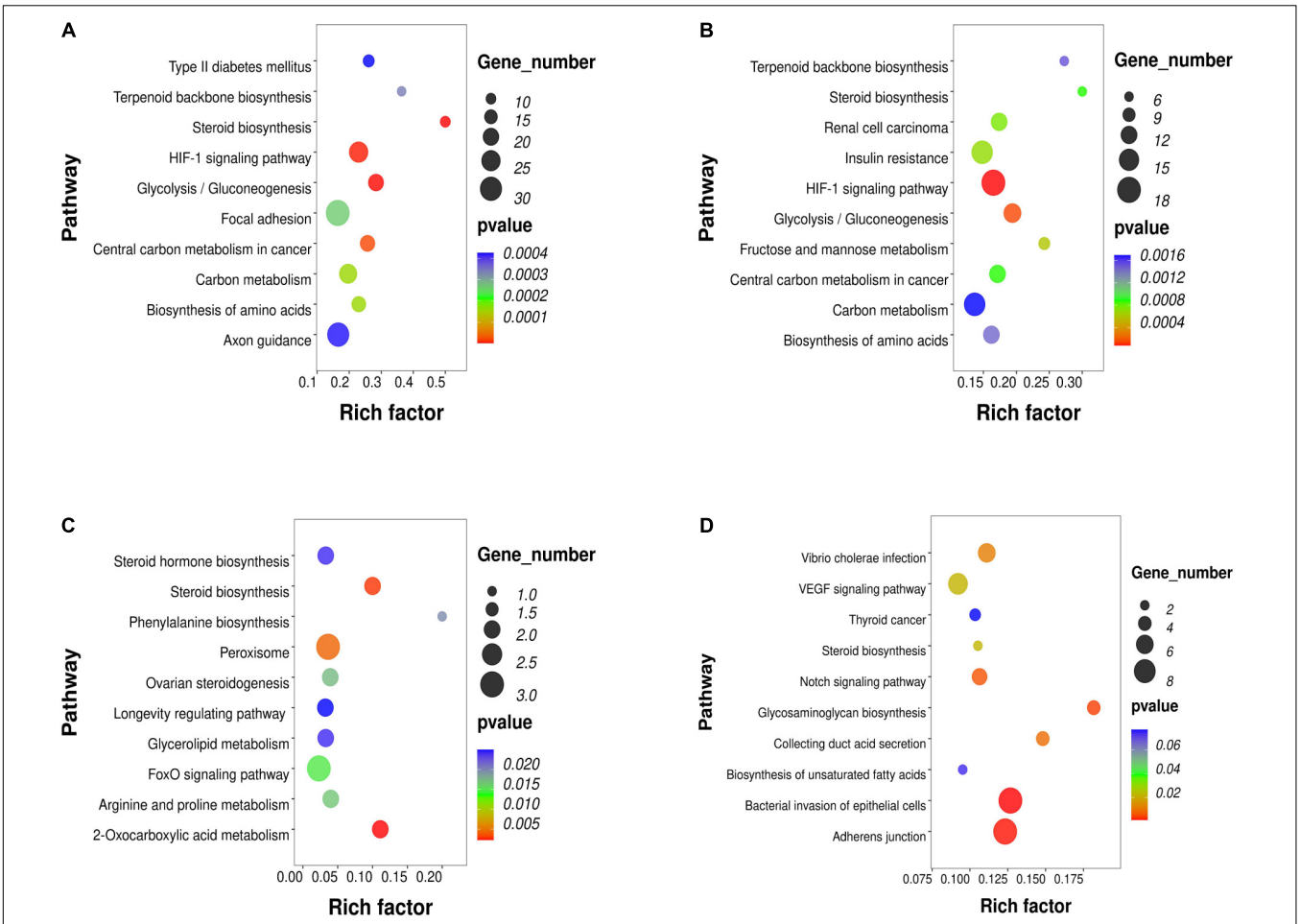



**FIGURE 2 |** KEGG pathway analysis of RNA-seq and methylation results in PCOS and control GCs. The top 10 affected biofunctions are grouped by disease. **(A)** Enriched KEGG pathways of DEGs. **(B)** Enriched KEGG pathways of DELs. **(C)** Enriched KEGG pathways of DEMs. **(D)** Enriched KEGG pathways of DMRs.

metabolic pathways. These analyses identified coexpressed genes that were associated with PCOS development.

## DISCUSSION

In the present study, we systematically investigated the differences in the mRNA-miRNA-lncRNA transcriptome and methylation modifications in control and PCOS GCs. Previous studies have focused on only single methylation modifications (Xu et al., 2016; Sagvekar et al., 2019) or single transcriptomics (Jones et al., 2015; Lan et al., 2015) in GCs. In addition, there were some multiomics studies on whole blood (Li et al., 2017), follicular fluid (Naji et al., 2018) and adipose tissue (Kokosar et al., 2016; Pan et al., 2018). However, few studies have performed multiomics analyses of GCs in patients with PCOS. Although many factors have been proven to play important roles in PCOS development in recent decades, no multiomics study has been performed in ovarian GCs. In this study, we systematically investigated control and PCOS ovarian GC mRNA-miRNA-lncRNA-DNA profiles and their potential regulatory networks.

In fact, some studies have shown that in GCs of PCOS patients, there is notable disruption of the entrainment of hormones and metabolic rhythms during the menstrual cycle (Wawrzkiewicz-Jalowiecka et al., 2020). The interaction of glucose/lipid metabolism and steroid synthesis shows obvious effects on both the development and the clinical manifestations of PCOS, mainly by increasing androgen availability, changing the function of GCs and disrupting follicle development (Pasquali et al., 2006). Follicle development depends on the synchronization of oocyte maturation and GC proliferation and differentiation. At the same time, the maturation of oocytes relies on the steroids and nutrients provided by GCs (Sutton-McDowall et al., 2010). From the perspective of follicle development, some of the DEGs identified in this study were related to androgen excess, impaired ovulation, and oxidative stress. Examples of such DEGs include CYP11A1, HSD17B7, and FOXO1, which have been identified to participate in the occurrence and development of PCOS (Sagvekar et al., 2019). PCOS is also characterized by IR and hyperinsulinemia. In PCOS, we also identified the involvement of some metabolic genes, such as IRS1 and INSR, showing increased expression, and IDH1 and GOT2, showing

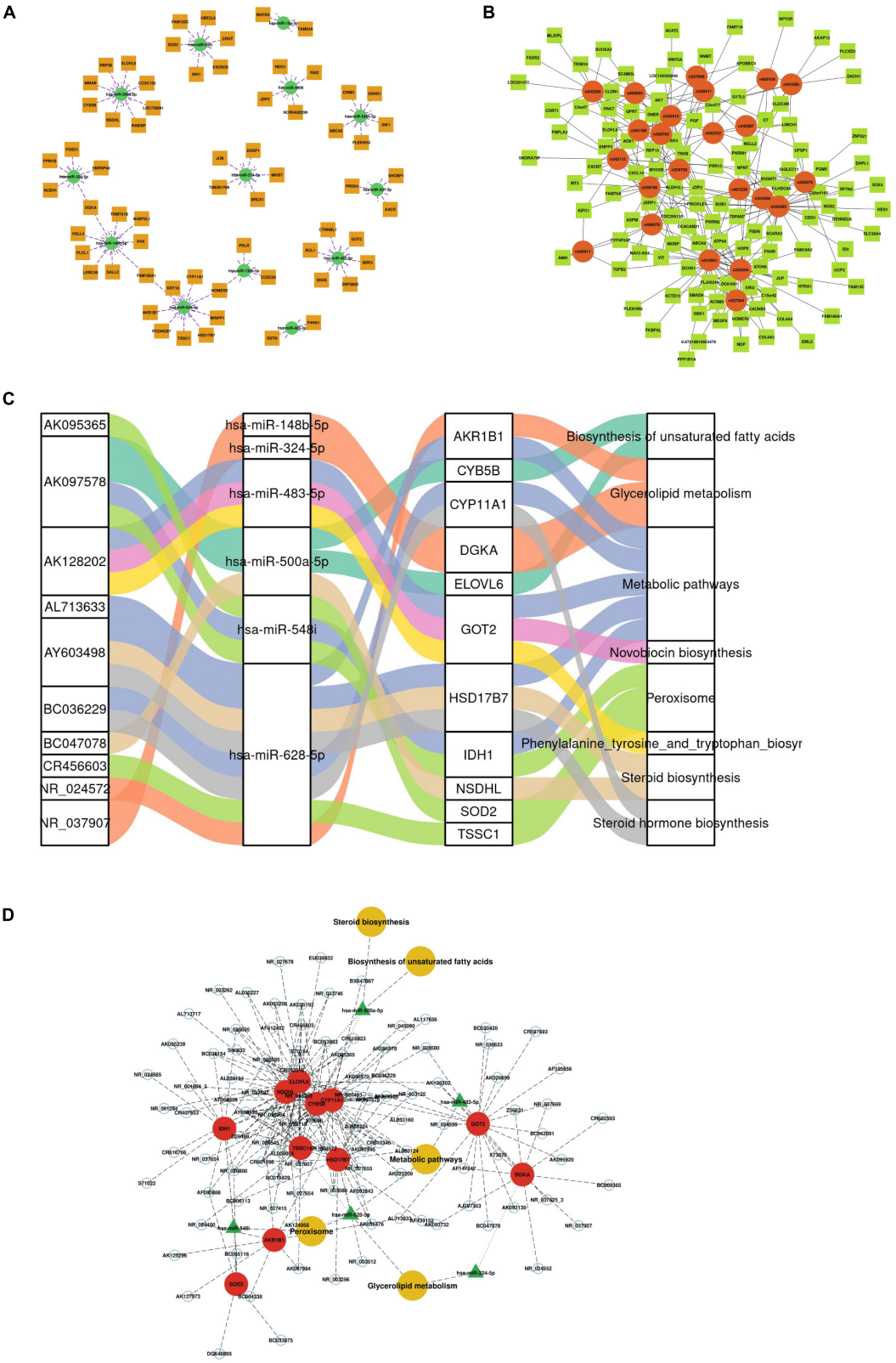



**FIGURE 3 |** Construction of the competing endogenous (ceRNA) regulatory network. **(A)** miRNA-mRNA interaction network. **(B)** lncRNA-mRNA coexpression network. **(C)** Sankey diagram of integrative network analysis of multi-RNA-seq data. **(D)** ceRNA interaction network of miRNA-mRNA-lncRNA interactions. This plot shows the potential regulatory linkage of different RNAs and biological pathways. The four modules represent lncRNAs, miRNAs, mRNAs, and pathways.

decreased expression at the transcriptional level, which may confer a genetic predisposition to developing this condition (Feng et al., 2015). Of note, TF analysis showed a higher level of the ZBTB16 gene in PCOS GCs, which is consistent with the findings of recent PCOS susceptibility gene studies. A large-scale genome-wide meta-analysis of PCOS patients of European ancestry suggests that a variant at the ZBTB16 locus was strongly associated with ovulatory dysfunction and polycystic ovarian morphology (Day et al., 2018). The SNP rs1784692 in the ZBTB16 gene was associated with PCOS and BMI levels in Han Chinese women (Yang et al., 2020).

PCOS is a complex and heterogeneous condition that results from the interaction of diverse genetic and environmental factors (Rosenfield and Ehrmann, 2016). In our study, we identified some key common enriched pathways, including glycolysis/gluconeogenesis, steroid biosynthesis, and IR, in the KEGG analysis of lncRNA-miRNA-mRNA interactions and DMR-associated genes. By constructing a ceRNA network, we also observed that the most highly enriched ceRNA axes might play a role in regulating metabolic pathways and steroid biosynthesis in the development of PCOS. Among them, both lncRNA-miR-628-5p-CYP11A1 and lncRNA-miR-628-5p-HSD17B7 ceRNA regulatory axes are associated with steroid hormone biosynthesis and metabolic pathways. A differential expression study showed that an increase in miR-628-5p serum levels at 20 weeks of gestation was observed in women who developed severe preeclampsia (Martinez-Fierro et al., 2019). In this study, the miR-628-5p-CYP11A1/HSD17B7 network was downregulated in PCOS GCs. A randomized clinical trial showed that frozen embryo transfer resulted in an increased risk of preeclampsia in twin pregnancy in women with PCOS (Zhang et al., 2018). Other preliminary evidence described that women with PCOS and the highest maternal testosterone levels in the early second trimester had the highest risk of developing preeclampsia (Valdimarsdottir et al., 2020). A transcriptome study showed that HSD17B7 and CYP11A1 expression was repressed in DHT-treated ovaries and that the dysregulation of HSD17B7 and CYP11A1 expression was associated with the biosynthesis and metabolism of steroids and cholesterol and lipids (Salilew-Wondim et al., 2015). Therefore, miR-628-5p may be an important hub regulator of HSD17B7 and CYP11A1 and may increase the risk of pregnancy complications by affecting steroid hormone biosynthesis and metabolic pathways in PCOS patients.

We also observed a downregulated axis consisting of lncRNAs-miR-483-5p and GOT2 associated with metabolic pathways in the ceRNA network. A previous study reported that miR-483-5p expression was significantly decreased in cumulus cells of PCOS patients (Shi et al., 2015). Other miRNA expression profiles revealed that miR-483-5p can regulate Notch3/MAPK3 expression (Xu et al., 2015) and progesterone concentrations (Sang et al., 2013) in cumulus GCs and follicular fluid of PCOS patients. Although few studies have focused on the function of GOT2 in PCOS, an important conclusion was reached by Yang et al. (2015), who found that GOT2 participates in mitochondrial metabolism through acetylation (Borst, 2020). Moreover, miR-483-5p is associated with future onset of both diabetes and cardiovascular disease (Gallo et al., 2018) and increases hepatic LDL receptor levels by inhibiting PCSK9 production (Dong et al., 2020). According to these results, it can be speculated that miR-483-5p may regulate GOT2 to contribute to the IR of PCOS and is worthy of further investigation.

In summary, the results show that women with PCOS have multiple transcriptional and epigenetic changes in GCs that are related to steroid hormone synthesis and metabolic pathways. Several genes and pathways, such as lncRNAs-miR-628-5p-CYP11A1/HSD17B7 and lncRNAs-miR-483-5p-GOT2, play important roles in the etiology of PCOS and may be novel candidate biomarkers or treatment targets for PCOS.

## DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: NCBI GEO GSE168404.

## ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Institutional review board of the Reproductive Hospital Affiliated to Shandong University. The patients/participants provided their written informed consent to participate in this study.

## AUTHOR CONTRIBUTIONS

HZ contributed to conception and design of the study. SGZ revised the manuscript. RSZ performed bioinformatics analysis and wrote the manuscript. YHJ supervised the bioinformatics analysis and edited the manuscript. All authors contributed to manuscript revision, read, and approved the submitted version.

## FUNDING

This research was supported by the National Key Research and Development Program of China (2018YFC1004303), the Basic Science Center Program (31988101) and the National Program (82071606, 31871509) of NSFC, and the Foundation for Distinguished Young Scholars of Shandong Province (JQ201816).

## ACKNOWLEDGMENTS

We sincerely thank the patients for their participation. We thank openbiox community and Hiplot team (https://hiplot.com.cn) for providing technical assistance and valuable tools for data analysis and visualization. We also thank Honghui Zhang and Yu

Zhang from Shandong University, China for their proofreading of this manuscript.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.648701/full#supplementary-material

**Supplementary Figure 1 |** Chromosome graph of differentially methylated regions (DMRs) between PCOS and control GCs. Different colors represent different types of DMR loci.

**Supplementary Figure 2 |** Specific pathway analysis of differentially expressed genes (DEGs). **(A)** Heatmap of DEGs associated with transcription factor regulation. **(B)** KEGG pathways of DEGs associated with transcription factor regulation. **(C)** Heatmap of DEGs associated with follicle development. **(D)** Heatmap of DEGs associated with steroid metabolism.

**Supplementary Table 1 |** Differentially expressed miRNA-mRNA-lncRNA-CpG-associated genes.

**Supplementary Table 2 |** KEGG pathways of miRNA-mRNA-lncRNA-CpG-associated genes.

**Supplementary Table 3 |** Integrated analysis of identified signature genes in the miRNA-mRNA-lncRNA network.

## REFERENCES


Aste, N., Pau, M., and Biggio, P. (1998). Kerion celsi: a clinical epidemiological study. *Mycoses* 41, 169–173. doi: 10.1111/j.1439-0507.1998.tb00319.x

Borst, P. (2020). The malate-aspartate shuttle (Borst cycle): how it started and developed into a major metabolic pathway. *IUBMB Life* 72, 2241–2259. doi: 10.1002/iub.2367

Day, F., Karaderi, T., Jones, M. R., Meun, C., He, C., Drong, A., et al. (2018). Large-scale genome-wide meta-analysis of polycystic ovary syndrome suggests shared genetic architecture for different diagnosis criteria. *PLoS Genet.* 14:e1007813. doi: 10.1371/journal.pgen.1007813

Dong, J., He, M., Li, J., Pessentheiner, A., Wang, C., Zhang, J., et al. (2020). microRNA-483 ameliorates hypercholesterolemia by inhibiting PCSK9 production. *JCI Insight* 5:e143812. doi: 10.1172/jci.insight.143812

Fauser, B. C., Tarlatzis, B. C., Rebar, R. W., Legro, R. S., Balen, A. H., Lobo, R., et al. (2012). Consensus on women's health aspects of polycystic ovary syndrome (PCOS): the Amsterdam ESHRE/ASRM-Sponsored 3rd PCOS consensus workshop group. *Fertil. Steril.* 97, 28–38. doi: 10.1016/j.fertnstert.2011.09.024

Feng, C., Lv, P. P., Yu, T. T., Jin, M., Shen, J. M., Wang, X., et al. (2015). The association between polymorphism of INSR and polycystic ovary syndrome: a meta-analysis. *Int. J. Mol. Sci.* 16, 2403–2425. doi: 10.3390/ijms16022403

Fontana, J., Martinkova, S., Petr, J., Zalmanova, T., and Trnka, J. (2020). Metabolic cooperation in the ovarian follicle. *Physiol. Res.* 69, 33–48. doi: 10.33549/physiolres.934233

Gallo, W., Esguerra, J., Eliasson, L., and Melander, O. (2018). miR-483-5p associates with obesity and insulin resistance and independently associates with new onset diabetes mellitus and cardiovascular disease. *PLoS One* 13:e206974. doi: 10.1371/journal.pone.0206974

Gioacchini, G., Notarstefano, V., Sereni, E., Zaca, C., Coticchio, G., Giorgini, E., et al. (2018). Does the molecular and metabolic profile of human granulosa cells correlate with oocyte fate? New insights by Fourier transform infrared microspectroscopy analysis. *Mol. Hum. Reprod.* 24, 521–532. doi: 10.1093/molehr/gay035

Holesh, J. E., Bass, A. N., and Lord, M. (2020). *Physiology, Ovulation*. Treasure Island, FL: StatPearls Publishing.

Hummitzsch, K., Anderson, R. A., Wilhelm, D., Wu, J., Telfer, E. E., Russell, D. L., et al. (2015). Stem cells, progenitor cells, and lineage decisions in the ovary. *Endocr. Rev.* 36, 65–91. doi: 10.1210/er.2014-1079

Iwase, A., Goto, M., Harata, T., Takigawa, S., Nakahara, T., Suzuki, K., et al. (2009). Insulin attenuates the insulin-like growth factor-I (IGF-I)-Akt pathway, not IGF-I-extracellularly regulated kinase pathway, in luteinized granulosa cells with an increase in PTEN. *J. Clin. Endocrinol. Metab.* 94, 2184–2191. doi: 10.1210/jc.2008-1948

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

Jones, M. R., Brower, M. A., Xu, N., Cui, J., Mengesha, E., Chen, Y. D., et al. (2015). Systems genetics reveals the functional context of PCOS loci and identifies genetic and molecular mechanisms of disease heterogeneity. *PLoS Genet.* 11:e1005455. doi: 10.1371/journal.pgen.1005455

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S. L. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* 14:R36. doi: 10.1186/gb-2013-14-4-r36

Kokosar, M., Benrick, A., Perfilyev, A., Fornes, R., Nilsson, E., Maliqueo, M., et al. (2016). Epigenetic and transcriptional alterations in human adipose tissue of polycystic ovary syndrome. *Sci. Rep.* 6:22883. doi: 10.1038/srep22883

Lan, C. W., Chen, M. J., Tai, K. Y., Yu, D. C., Yang, Y. C., Jan, P. S., et al. (2015). Functional microarray analysis of differentially expressed genes in granulosa cells from women with polycystic ovary syndrome related to MAPK/ERK signaling. *Sci. Rep.* 5:14994. doi: 10.1038/srep14994

Li, M., Wu, H., Luo, Z., Xia, Y., Guan, J., Wang, T., et al. (2012). An atlas of DNA methylomes in porcine adipose and muscle tissues. *Nat .Commun.* 3:850. doi: 10.1038/ncomms1854

Li, S., Zhu, D., Duan, H., Ren, A., Glintborg, D., Andersen, M., et al. (2017). Differential DNA methylation patterns of polycystic ovarian syndrome in whole blood of Chinese women. *Oncotarget* 8, 20656–20666. doi: 10.18632/oncotarget.9327

Li, X., Zhu, Q., Wang, W., Qi, J., He, Y., Wang, Y., et al. (2019). Elevated chemerin induces insulin resistance in human granulosa-lutein cells from polycystic ovary syndrome patients. *FASEB J.* 33, 11303–11313. doi: 10.1096/fj.201802829R

Martinez-Fierro, M. L., Carrillo-Arriaga, J. G., Luevano, M., Lugo-Trampe, A., Delgado-Enciso, I., Rodriguez-Sanchez, I. P., et al. (2019). Serum levels of miR-628-3p and miR-628-5p during the early pregnancy are increased in women who subsequently develop preeclampsia. *Pregnancy Hypertens.* 16, 120–125. doi: 10.1016/j.preghy.2019.03.012

Naji, M., Nekoonam, S., Aleyasin, A., Arefian, E., Mahdian, R., Azizi, E., et al. (2018). Expression of miR-15a, miR-145, and miR-182 in granulosa-lutein cells, follicular fluid, and serum of women with polycystic ovary syndrome (PCOS). *Arch. Gynecol. Obstet.* 297, 221–231. doi: 10.1007/s00404-017-4570-y

Pan, J. X., Tan, Y. J., Wang, F. F., Hou, N. N., Xiang, Y. Q., Zhang, J. Y., et al. (2018). Aberrant expression and DNA methylation of lipid metabolism genes in PCOS: a new insight into its pathogenesis. *Clin. Epigenetics* 10:6. doi: 10.1186/s13148-018-0442-y

Pasquali, R., Gambineri, A., and Pagotto, U. (2006). The impact of obesity on reproduction in women with polycystic ovary syndrome. *BJOG* 113, 1148–1159. doi: 10.1111/j.1471-0528.2006.00990.x

Risal, S., Pei, Y., Lu, H., Manti, M., Fornes, R., Pui, H. P., et al. (2019). Prenatal androgen exposure and transgenerational susceptibility to polycystic ovary syndrome. *Nat. Med.* 25, 1894–1904. doi: 10.1038/s41591-019-0666-1

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

Rosenfield, R. L., and Ehrmann, D. A. (2016). The Pathogenesis of Polycystic Ovary Syndrome (PCOS): the hypothesis of PCOS as functional ovarian hyperandrogenism revisited. *Endocr. Rev.* 37, 467–520. doi: 10.1210/er.2015-1104

Rotterdam ESHRE/ASRM-Sponsored PCOS Consensus Workshop Group (2004). Revised 2003 consensus on diagnostic criteria and long-term health risks related

to polycystic ovary syndrome (PCOS). *Hum. Reprod.* 19, 41–47. doi: 10.1093/humrep/deh098

Sagvekar, P., Kumar, P., Mangoli, V., Desai, S., and Mukherjee, S. (2019). DNA methylome profiling of granulosa cells reveals altered methylation in genes regulating vital ovarian functions in polycystic ovary syndrome. *Clin. Epigenetics* 11:61. doi: 10.1186/s13148-019-0657-6

Saito, R., Smoot, M. E., Ono, K., Ruscheinski, J., Wang, P. L., Lotia, S., et al. (2012). A travel guide to Cytoscape plugins. *Nat. Methods* 9, 1069–1076. doi: 10.1038/nmeth.2212

Salilew-Wondim, D., Wang, Q., Tesfaye, D., Schellander, K., Hoelker, M., Hossain, M. M., et al. (2015). Polycystic ovarian syndrome is accompanied by repression of gene signatures associated with biosynthesis and metabolism of steroids, cholesterol and lipids. *J. Ovarian Res.* 8:24. doi: 10.1186/s13048-015-0151-5

Sang, Q., Yao, Z., Wang, H., Feng, R., Wang, H., Zhao, X., et al. (2013). Identification of microRNAs in human follicular fluid: characterization of microRNAs that govern steroidogenesis in vitro and are associated with polycystic ovary syndrome in vivo. *J. Clin. Endocrinol. Metab.* 98, 3068–3079. doi: 10.1210/jc.2013-1715

Shi, L., Liu, S., Zhao, W., and Shi, J. (2015). miR-483-5p and miR-486-5p are down-regulated in cumulus cells of metaphase II oocytes from women with polycystic ovary syndrome. *Reprod. Biomed. Online* 31, 565–572. doi: 10.1016/j.rbmo.2015.06.023

Song, J., Lu, Y., Sun, W., Han, M., Zhang, Y., and Zhang, J. (2019). Changing expression profiles of lncRNAs, circRNAs and mRNAs in esophageal squamous carcinoma. *Oncol. Lett.* 18, 5363–5373. doi: 10.3892/ol.2019.10880

Sutton-McDowall, M. L., Gilchrist, R. B., and Thompson, J. G. (2010). The pivotal role of glucose metabolism in determining oocyte developmental competence. *Reproduction* 139, 685–695. doi: 10.1530/REP-09-0345

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R., et al. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* 7, 562–578. doi: 10.1038/nprot.2012.016

Valdimarsdottir, R., Wikstrom, A. K., Kallak, T. K., Elenis, E., Axelsson, O., Preissl, H., et al. (2020). Pregnancy outcome in women with polycystic ovary syndrome in relation to second-trimester testosterone levels. *Reprod. Biomed. Online* 42, 217–225. doi: 10.1016/j.rbmo.2020.09.019

Wawrzkiewicz-Jalowiecka, A., Kowalczyk, K., Trybek, P., Jarosz, T., Radosz, P., Setlak, M., et al. (2020). In search of new therapeutics-molecular aspects of the PCOS pathophysiology: genetics, hormones, metabolism and beyond. *Int. J. Mol. Sci.* 21:7054. doi: 10.3390/ijms21197054

Wilson, G. A., and Beck, S. (2016). "Computational analysis and integration of MeDIP-seq methylome data," in *Next Generation Sequencing: Advances, Applications and Challenges*, ed. J. K. Kulski (Rijeka: InTech), 159–169.

Xu, B., Zhang, Y. W., Tong, X. H., and Liu, Y. S. (2015). Characterization of microRNA profile in human cumulus granulosa cells: identification of microRNAs that regulate Notch signaling and are associated with PCOS. *Mol. Cell Endocrinol.* 404, 26–36. doi: 10.1016/j.mce.2015.01.030

Xu, J., Bao, X., Peng, Z., Wang, L., Du, L., Niu, W., et al. (2016). Comprehensive analysis of genome-wide DNA methylation across human polycystic ovary syndrome ovary granulosa cell. *Oncotarget* 7, 27899–27909. doi: 10.18632/oncotarget.8544

Yang, H., Zhou, L., Shi, Q., Zhao, Y., Lin, H., Zhang, M., et al. (2015). SIRT3-dependent GOT2 acetylation status affects the malate-aspartate NADH shuttle activity and pancreatic tumor growth. *EMBO J.* 34, 1110–1125. doi: 10.15252/embj.201591041

Yang, J., Zhao, R., Li, L., Li, G., Yang, P., Ma, J., et al. (2020). Verification of a ZBTB16 variant in polycystic ovary syndrome patients. *Reprod. Biomed. Online* 41, 724–728. doi: 10.1016/j.rbmo.2020.05.005

Yilmaz, B., Vellanki, P., Ata, B., and Yildiz, B. O. (2018). Metabolic syndrome, hypertension, and hyperlipidemia in mothers, fathers, sisters, and brothers of women with polycystic ovary syndrome: a systematic review and meta-analysis. *Fertil. Steril.* 109, 356–364. doi: 10.1016/j.fertnstert.2017.10.018

Yu, Y. Y., Sun, C. X., Liu, Y. K., Li, Y., Wang, L., and Zhang, W. (2015). Genome-wide screen of ovary-specific DNA methylation in polycystic ovary syndrome. *Fertil. Steril.* 104, 145–153. doi: 10.1016/j.fertnstert.2015.04.005

Zhang, B., Wei, D., Legro, R. S., Shi, Y., Li, J., Zhang, L., et al. (2018). Obstetric complications after frozen versus fresh embryo transfer in women with polycystic ovary syndrome: results from a randomized trial. *Fertil. Steril.* 109, 324–329. doi: 10.1016/j.fertnstert.2017.10.020

Zhu, L. J., Gazin, C., Lawson, N. D., Pages, H., Lin, S. M., Lapointe, D. S., et al. (2010). ChIPpeakAnno: a bioconductor package to annotate ChIP-seq and ChIP-chip data. *BMC Bioinformatics* 11:237. doi: 10.1186/1471-2105-11-237

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

frontiers in Genetics

ORIGINAL RESEARCH
published: 20 May 2021
doi: 10.3389/fgene.2021.636441

# Machine Learning Assisted Prediction of Prognostic Biomarkers Associated With COVID-19, Using Clinical and Proteomics Data

*Rahila Sardar*[1,2†], *Arun Sharma*[1†] *and Dinesh Gupta*[1*]

[1] Translational Bioinformatics Group, International Centre for Genetic Engineering and Biotechnology, New Delhi, India, [2] Department of Biochemistry, Jamia Hamdard, New Delhi, India

OPEN ACCESS

**Edited by:**
*Amit Kumar Yadav, Translational Health Science and Technology Institute (THSTI), India*

**Reviewed by:**
*Arjun Ray, Indraprastha Institute of Information Technology Delhi, India*
*R. Shyama Prasad Rao, Yenepoya University, India*

***Correspondence:**
*Dinesh Gupta*
*dinesh@icgeb.res.in*

[†] *These authors have contributed equally to this work*

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

**Received:** *01 December 2020*
**Accepted:** *18 March 2021*
**Published:** *20 May 2021*

**Citation:**
*Sardar R, Sharma A and Gupta D (2021) Machine Learning Assisted Prediction of Prognostic Biomarkers Associated With COVID-19, Using Clinical and Proteomics Data. Front. Genet. 12:636441. doi: 10.3389/fgene.2021.636441*

With the availability of COVID-19-related clinical data, healthcare researchers can now explore the potential of computational technologies such as artificial intelligence (AI) and machine learning (ML) to discover biomarkers for accurate detection, early diagnosis, and prognosis for the management of COVID-19. However, the identification of biomarkers associated with survival and deaths remains a major challenge for early prognosis. In the present study, we have evaluated and developed AI-based prediction algorithms for predicting a COVID-19 patient's survival or death based on a publicly available dataset consisting of clinical parameters and protein profile data of hospital-admitted COVID-19 patients. The best classification model based on clinical parameters achieved a maximum accuracy of 89.47% for predicting survival or death of COVID-19 patients, with a sensitivity and specificity of 85.71 and 92.45%, respectively. The classification model based on normalized protein expression values of 45 proteins achieved a maximum accuracy of 89.01% for predicting the survival or death, with a sensitivity and specificity of 92.68 and 86%, respectively. Interestingly, we identified 9 clinical and 45 protein-based putative biomarkers associated with the survival/death of COVID-19 patients. Based on our findings, few clinical features and proteins correlate significantly with the literature and reaffirm their role in the COVID-19 disease progression at the molecular level. The machine learning–based models developed in the present study have the potential to predict the survival chances of COVID-19 positive patients in the early stages of the disease or at the time of hospitalization. However, this has to be verified on a larger cohort of patients before it can be put to actual clinical practice. We have also developed a webserver CovidPrognosis, where clinical information can be uploaded to predict the survival chances of a COVID-19 patient. The webserver is available at http://14.139.62.220/covidprognosis/.

**Keywords: machine learning, biomarkers discovery, COVID-19, feature selection, proteomics and bioinformatics**

## INTRODUCTION

In December 2019, the COVID-19 disease initiated as an outbreak caused by SARS-CoV-2, which quickly snowballed into a catastrophic worldwide healthcare crisis (Srivastava et al., 2020). On March 11, 2020, the World Health Organization (WHO) declared COVID-19 a global pandemic with more than 118,000 cases in 114 countries and over 4,000 deaths, much more than the morbidity and mortality caused by related viruses such as SARS and MERS. As of March 14, 2021, the pandemic has caused more than 119 million confirmed COVID-19 cases and ∼2.64 million deaths worldwide[1].

Compared to other respiratory diseases such as influenza, the COVID-19 human-to-human transmission is facilitated through respiratory droplets (particles > 5–10 nm in diameter) from coughing and sneezing. The clinical symptoms associated with COVID-19 patients vary from asymptomatic or symptomatic forms (Cascella et al., 2020). A study published in *JAMA* consists of data from 72,314 cases, including records from confirmed, suspected, diagnosed, and asymptomatic COVID-19 patients, shared by the Chinese Center for Disease Control and Prevention (China CDC), demonstrating the epidemiologic curve of the Chinese outbreak. As per this report, the mortality of critically ill patients was 49.0% in contrast to 2.3% for the overall COVID-19 patients. The mortality was also higher for patients with various comorbidities such as cardiovascular disease, diabetes, chronic respiratory disease, and oncological diseases, whereas patients with the age of 9 or younger did not have any fatal cases (Wu and McGoogan, 2020).

At present, no SARS-CoV-2 specific drug or reliable prognostic biomarker is available for COVID-19 treatment (González-Pacheco et al., 2020; Pandey et al., 2020). Various therapeutic measures to enhance the immune systems by immune modulators have been proposed (Zhong et al., 2020). Recommended preventive measures include social distancing, proper health, and hygiene management (Al-Rohaimi and Al Otaibi, 2020). It is also known that the severity of COVID-19 largely depends on the host and viral factors. The latter highlights the importance of identifying the host features associated with the disease severity at the molecular level (Zhang et al., 2020). Given the facts enumerated above, it is desirable to have the correct prognostic assessment of patients for proper clinical management.

Artificial intelligence (AI) is being employed to meet new healthcare requirements, in view of the pandemic, for example, tracking the SARS-CoV-2 virus spread and quickly identifying high-risk patients (Sharma et al., 2020). Machine learning (ML) methods have been exploited to analyze various kinds of biological datasets such as proteomics data, NGS data, and metabolomics data to predict the biomarkers for classification of samples and genes associated with a particular disease state (Dumancas et al., 2017; Cambiaghi et al., 2018). The mitigation potential of AI technology has been extensively demonstrated for various pandemics and infectious diseases, for example, SARS, Ebola, HIV, and COVID-19 (Lalmuanawma et al., 2020; Overmyer et al., 2020).

To date, there are several reports on clinical biomarkers associated with the disease prognosis. However, there are only a few published articles on protein-based biomarkers, and hence, further research is required to confirm the existing findings (Graziani et al., 2020; Kaur et al., 2020; Kermali et al., 2020). Integrated data analysis on COVID-19 genomes has been performed to identify several crucial factors involved in host–pathogen interaction. However, limited attempts have been made to integrate high throughput datasets (Sardar et al., 2020). Yan et al. (2020b) developed a machine learning model with more than 90% accuracy on 485 COVID positive patients to predict the clinical biomarkers associated with individual patients' mortality. Another study by Yao H. et al. (2020) aimed to predict the disease severity among the patients by utilizing the data on 137 COVID-19 infected patients using an ML-based model on the blood and urine examination parameters. However, these methods are not free from errors, limitations, and challenges, rendering them unfit to be used in real-world problems.

Motivated by the availability of appropriate clinical datasets, we used such a dataset for training ML algorithms to exploit its potential for the prognosis of COVID-19 positive patients. We designed a pipeline to predict features, namely proteins and clinical parameters, associated with the disease severity and survival of the COVID-19 patients. Interestingly, we have identified 9 clinical features and 45 proteins related to the survival/death of COVID-19 patients. Few of the identified clinical features and proteins correlate well with the literature and reaffirm their role in the COVID-19 disease progression at the molecular level (Shen et al., 2020; Wynants et al., 2020; Yan et al., 2020a). The potential role of identified proteins in various pathways, their native functions, potential to be a drug target, etc., are described in the subsequent sections. The ML-based models developed in the present study possess an immense potential to predict the survival chances of COVID-19 positive patients in the early stages of the disease or at the time of hospitalization.

## MATERIALS AND METHODS

### Data Source

We downloaded the clinical and normalized protein expression profile data for 306 COVID-19 patients and 78 other patients (control subjects) from the Olink website (Filbin et al.). We downloaded three files, namely "MGH_COVID_OLINK_NPX.txt," "MGH_COVID_Clinical_Info.txt," and "variable_descriptions.xlsx," containing protein data (with relative quantification values given in Olink's proprietary Normalized Protein expression (NPX) units), essential clinical data (associated with each sample), and a worksheet (with a description of the clinical variables presented), respectively. Although clinical and protein data were present in two different files, the data were linked based on the subject IDs.

[1]https://covid19.who.int/

## Data Preprocessing

Data preprocessing is essential for a machine learning study. Hence, we checked the data for any experimental impurities through semiautomated ways. As depicted in **Figure 1**, clinical and proteomic data were missing for a few patients. In the case of clinical data, we replaced missing values with "-1." Thus, we used the clinical data of 42 dead and 264 survivors (Whole dataset I) for training the "Clinical Information" based classification models for days 0–7. However, in the proteomics data, the protein expression values were missing for 165 and 248 patients for days 3 and 7, respectively. Therefore, we used only proteomics data for the Day 0 proteomics information-based classification model generation. For only one COVID-19 positive patient (who died within 28 days of hospitalization), protein expression values (for few of the 1,428 proteins) were missing, while protein expression values were missing for 15 patients among the survivors (for few of the 1,428 proteins); hence, we excluded these records from the study (**Figure 1**). Thus, we used the proteomics data (Whole dataset II) of 41 dead and 249 survivors to train and validate the machine learning–based models.

As evident from the downloaded data, the number of survivors and deaths in clinical as well as proteomics data were imbalanced. The survivor's data (for both clinical and proteomics data) were split into five, almost equal-sized, divisions (P1–P5). Furthermore, we trained and validated the models using each of the five divisions and the dataset of dead patients. The tools, techniques, and statistical measures used to evaluate the model performances and the retrieved results are given in the subsequent sections.

## Tools Used for the Development of Classification Models

WEKA (Frank et al., 2016), a popular and widely used data mining and machine learning tool, was used for training and validation of the various machine learning–based classification models developed in this study. All the techniques available with the WEKA (v3.8.2) were used to train and validate the classification models. For clinical data, five types of models are generated, i.e., the models based on (1) Day 0 clinical parameters, (2) Day 3 clinical parameters, (3) Day 7 clinical parameters, (4) Days 0–7 clinical parameters, and (5) Selected clinical parameters (out of Days 0–7 clinical parameters). On the other hand, for proteomic data, two types of models are generated, i.e., (1) Day 0, all 1428 protein parameters, and (2) Day 0 protein parameters based on feature selection.

We trained and evaluated 44 different types of ML classification algorithms available in WEKA (v3.8.2). However, several combinations of various parameters for these algorithms and the number of input parameters used (for the training and validation of classification models) resulted in thousands of models (for details, check http://14.139.62.220/covidprognosis/supple.php). For example, in the case of Day 0 clinical parameters-based model (using the P1 dataset), a total of 85 models were trained and evaluated using Day 0 all 33 clinical parameters. Thus, for P1–P5 splits, a total of 425 models (85 × 5) were developed to determine the best classification models.

## Feature Selection

In different machine learning–based classification studies, all the input features do not play an equally significant role in classification (Sharma et al., 2016; Jablonka et al., 2020; Kumar et al., 2020). Therefore, to identify the most significant clinical and proteomics features, all the feature selection techniques available with WEKA were applied to the Days 0–3 clinical features dataset (consisting of 33 clinical parameters) and Day 0 proteomics data (for the 1,428 proteins).

## Cross-Validation Techniques Used

The availability of enormous data is essential for preparing training and validation datasets during a machine learning–based study. However, due to limited patients' records, it was impossible to prepare separate training and validation datasets. Therefore, the leave-one-out cross-validation (LOOCV) technique was used to utilize the available information optimally. In the LOOCV technique, the models are trained and validated so that each record is used for training and testing. The LOOCV technique has widely been used to solve several classification problems (Mete et al., 2016; Nath and Subbiah, 2016; Jiang et al., 2019).

## Formulae Used to Evaluate Performance of the Models

The performance of the models was evaluated using statistical measures such as sensitivity, specificity, accuracy, and Mathew's correlation coefficient (MCC). The formulae used are given below:

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \times 100$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \times 100$$

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \times 100$$

$$\text{MCC} = \frac{(\text{TP})(\text{TN}) - (\text{FP})(\text{FN})}{\sqrt{[\text{TP} + \text{FP}][\text{TP} + \text{FN}][\text{TN} + \text{FP}][\text{TN} + \text{FN}]}} \times 100$$

where TP and TN are correctly predicted positive and negative examples, respectively. Similarly, FP and FN are wrongly predicted positive and negative examples, respectively. The models with the highest MCC value and almost equal sensitivity and specificity values are considered best prediction models.

## Pathway Analysis and Identification of Drug Targets

To understand the biological functions of the shortlisted proteins, pathway analysis was performed using the DAVID tool (Jiao et al., 2012). Targeting host proteins appears to be a promising approach in antiviral research. To identify the drugs against the selected proteins, all the drug target information was downloaded from the TTD database, and only validated and clinically proven drugs were used for the analysis (Wang et al., 2020). The drugs that have been withdrawn or not in use were removed from the drug-targets based analysis.

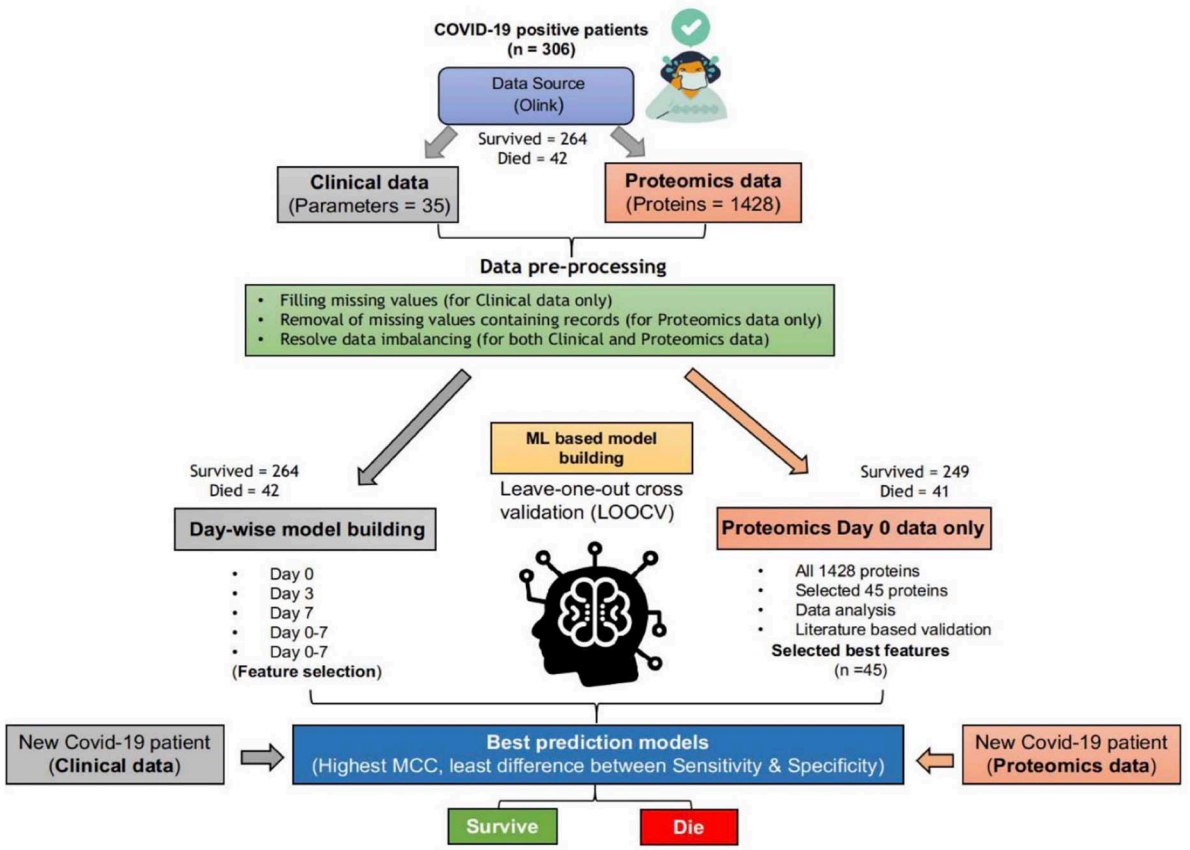



**FIGURE 1 |** ML-based pipeline to identify key features associated with survival based on clinical and proteomics data. (The figure images were generated using biorender.com).

## Webserver Development

The CovidPrognosis webserver has been developed using efficient and open-source Linux-Apache-MySQL-PHP/ Perl/Python (LAMP) server technologies. The user interface (UI) or web interface is developed using HTML, CSS, PHP (v7.1.28), and AJAX. Moreover, the predictions are performed using the WEKA-based machine learning models, trained and validated on clinical parameters.

# RESULTS

## Models Based on Whole Clinical Parameters

The classification models were developed using clinical information, as given in **Supplementary Table 1**. A total of five types of models (thousands in number; based on all available techniques in the WEKA package) were developed using the Day 0 (Sr. No. 3-21), Day 3 (Sr. No. 3-14 and 22-28), Day 7 (Sr. No. 3-14 and Sr. No. 29-35), and Days 0–7 (Sr. No. 3-35) clinical parameter values (**Supplementary Table 1**). However, two models achieved the highest performance using Day 0 and Days 0–7 information, while "Whole dataset I" based models showed a large difference between sensitivity and specificity values. This difference may be attributed to the imbalance between the number of records for survived and died patients. The Day 0 clinical parameters-based model (using the "IterativeClassifierOptimizer" technique) achieved a maximum accuracy of 87.37% with the highest sensitivity (%), specificity (%), MCC, and ROC values of 88.10, 86.79, 0.75, and 0.863, respectively (**Table 1**). Using "RandomForest" as the classification technique and Days 0–7 clinical parameters (33) as input features, a maximum accuracy of 89.47% was achieved with the highest sensitivity (%), specificity (%), MCC, and ROC values of 85.71, 92.45, 0.79, and 0.921, respectively (**Table 1**).

## Feature Selection for Clinical Parameters

For the clinical data, three clinical parameters, namely, age, absolute lymphocyte count (Day 0), and creatinine level (Day 0), and nine clinical parameters, i.e., age, absolute lymphocyte count (Day 0), creatinine level (Day 0), preexisting heart disease(s), preexisting hypertension, preexisting kidney disease(s), D-dimer level (Day 0), any GI-related symptoms at the time of hospital presentation, and cardiac event-Trop_72 (hs-cTn = > 100 within the first 72 h of presentation) clinical parameters or features were selected by the majority of the techniques[2]. Therefore,

[2]http://14.139.62.220/covidprognosis/supple.php

**TABLE 1 |** Performance of best models based on whole clinical parameters.

| Dataset (no. of clinical parameters used) | Day(s) | Sensitivity (%) | Specificity (%) | Accuracy (%) | MCC | ROC | WEKA technique used |
|---|---|---|---|---|---|---|---|
| Whole dataset I (19) | 0 | 50 | 94.7 | 88.56 | 0.48 | 0.806 | AttributeSelectedClassifier |
| **P1 (19)** | **0** | **88.1** | **86.79** | **87.37** | **0.75** | **0.863** | **IterativeClassifierOptimizer** |
| Average of P1–P5 splits (19) | 0 | 81.90 | 82.94 | 82.48 | 0.65 | 0.808 | IterativeClassifierOptimizer |
| Whole dataset I (33) | 0, 3, 7 | 47.62 | 96.21 | 89.54 | 0.51 | 0.739 | J48 |
| **P2 (33)** | **0, 3, 7** | **85.71** | **92.45** | **89.47** | **0.79** | **0.921** | **RandomForest (with -K 4)** |
| Average of P1–P5 splits (33) | 0, 3, 7 | 75.24 | 81.43 | 78.68 | 0.57 | 0.868 | RandomForest (with -K 4) |

these three clinical parameters (selected by CfsSubsetEval as "Attribute Evaluator" with BestFirst as "Search Method") and nine clinical parameters [selected by "InfoGainAttributeEval" as "Attribute Evaluator" with Ranker algorithm (attributes with ranking value $>$ 0 were selected)] have been used for the training and evaluation of the machine learning–based models.

## Models Based on Selected Clinical Parameters

From the analysis of the clinical data, it is found that the patients from the age group of 65–80+ years, with lower elevated lymphocyte count at Day 0 ($<$1.00), D-dimer $\geq$ 1,000 (units), are at a higher risk of death during hospitalization and require immediate treatment (**Figure 2**).

The "Whole dataset I"–based models showed a large difference between sensitivity and specificity values. A maximum accuracy of 87.37% was achieved with sensitivity (%), specificity (%), MCC, and ROC values of 85.71, 88.68, 0.74, and 0.845, from the three selected clinical features, respectively. While from the nine selected clinical parameters, a maximum accuracy of 86.32% was achieved with sensitivity (%), specificity (%), and MCC, and ROC values of 83.33, 88.68, 0.72, and 0.81, respectively, as shown in **Table 2**. The identified clinical features such as serum creatinine (Day 0), age, absolute lymphocyte count (Day 0), and D-dimer (Day 0) along with comorbidities such as preexisting heart disease(s), preexisting kidney disease(s), preexisting hypertension, GI symptoms at presentation, and Trop-72 can be highly useful in the classification of patients with survival or dying probabilities. These identified features can be evaluated as biomarkers that can help identify the patients who require immediate medical attention.

## Models Based on Whole NPX Proteomics Data

To understand the role of the protein expression profile in the classification of COVID-19 patients who survived vs. are dead, the expression values of 1428 proteins were used to develop machine learning–based classification models. The "Whole dataset II"–based models showed a large difference between sensitivity and specificity values. It is evident from **Table 3** that an accuracy of 83.52% was achieved (using the dataset P4) with a sensitivity (%), specificity (%), MCC, and ROC values of 82.93, 84, 0.67, and 0.868, respectively.

## Identification of Proteins Associated With Survival vs. Deaths

The feature selection technique was applied to determine the most significant proteins that are helpful for the classification of patients who survived COVID-19 vs. those who died. Therefore, for proteomics data, different feature selection techniques resulted in the selection of a different set of proteomic features (see text footnote 2). Thus, a total of 45 proteins were identified through WEKA using CfsSubsetEval as the "Attribute Evaluator" with BestFirst as the "Search Method" (**Supplementary Tables 2, 3**).

As evident from **Table 4**, an accuracy of 89.01% was achieved (using the dataset P2) with sensitivity (%), specificity (%), MCC, and ROC values of 92.68, 86, 0.78, and 0.953, respectively. On the other hand, "Whole dataset II"–based models showed a large difference between sensitivity and specificity values.

## Expression and Pathway Analysis of the Shortlisted Proteins

The shortlisted proteins include lipid metabolism proteins (APOM), a protease inhibitor (FETUB), serine protease (FA7, GGH), growth factors (EGFR, PDGFB, TGFA, and GDF8), chemokines, interleukins (IL8, IL17C), and others (**Supplementary Table 2**). Recent studies have shown that APOM is downregulated in severe COVID-19 patients (Shen et al., 2020). The dysregulation of APOM is also associated with hepatitis B virus (HBV) infected patients (Gu et al., 2011). Another important protein associated with survival is angiopoietin (AGP), which is recently reported to cause inflammatory intussusceptive angiogenesis and diffuse alveolar damage in COVID-19, and the progression of carcinogenetic events in cancer patients (Saha and Anirvan, 2020). Q96PL1_SG3A2 is highly expressed and shows antifibrotic activity in the lungs (Cai et al., 2014).

These shortlisted proteins were further analyzed to understand their role in human physiology and COVID-19 prognosis. From the pathway analysis, we found that the selected 45 proteins are associated with pathways such as the IFN-gamma pathway, IL5 and IL3 mediating signaling events, cytokine, chemokine, and VEGF signaling, as shown in **Figure 3**.

## Identification of Potential Drug Targets Among the Shortlisted Proteins

To date, no reliable drug has been approved to treat COVID-19. From the drug target database (**Supplementary Table 4**), we were

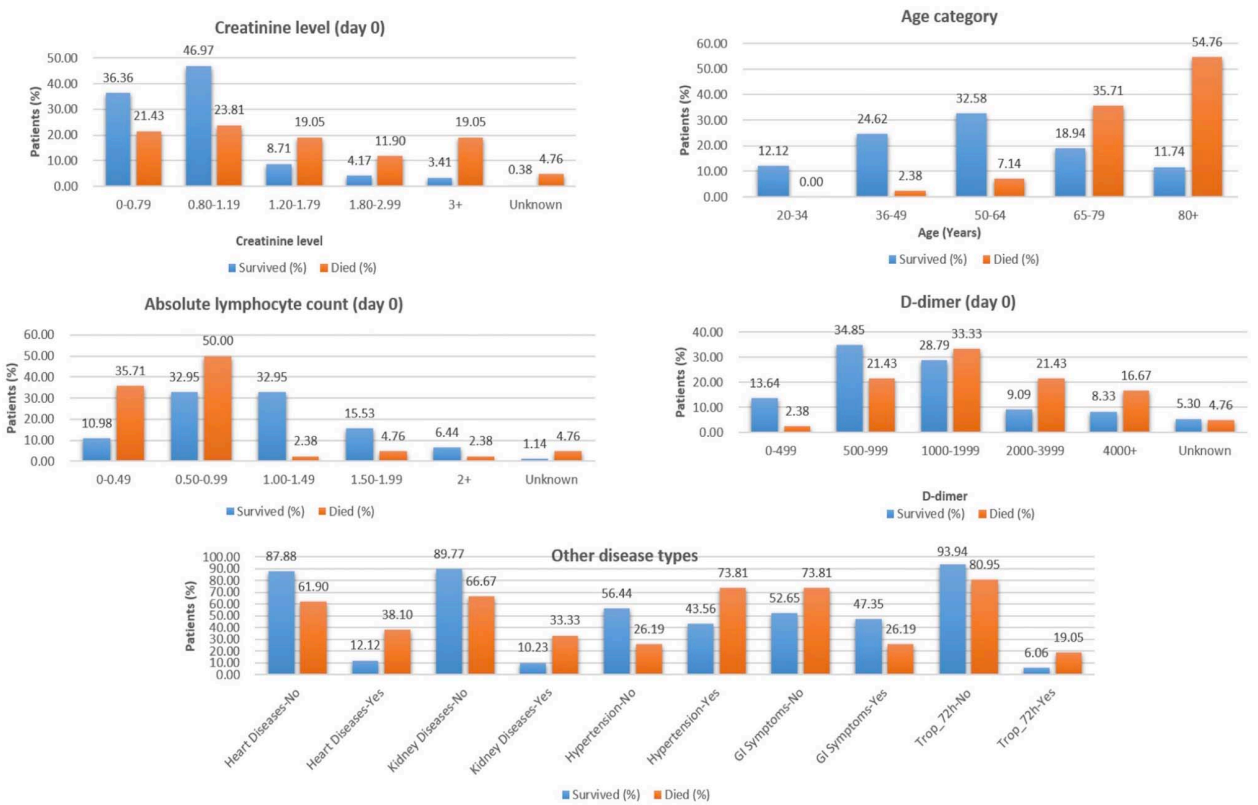



**FIGURE 2 |** Selected features from clinical data to classify COVID-19 patients who survived vs. those who died.

**TABLE 2 |** Performance of best models based on selected clinical parameter values.

| Dataset (no. of clinical parameters used) | Day(s) | Sensitivity (%) | Specificity (%) | Accuracy (%) | MCC | ROC | WEKA technique used |
|---|---|---|---|---|---|---|---|
| Whole dataset I (3) | 0 | 50 | 94.7 | 88.56 | 0.48 | 0.806 | J48 |
| **P2 (3)** | **0** | **85.71** | **88.68** | **87.37** | **0.74** | **0.845** | **RandomSubSpace** |
| Average of P1–P5 splits (3) | 0 | 83.33 | 80.31 | 81.64 | 0.63 | 0.831 | RandomSubSpace |
| Whole dataset I (9) | 0 | 50 | 94.7 | 88.56 | 0.48 | 0.806 | AttributeSelectedClassifier |
| **P2 (9)** | **0, 3** | **83.33** | **88.68** | **86.32** | **0.72** | **0.81** | **IterativeClassifierOptimizer** |
| Average of P1–P5 splits (9) | 0, 3 | 81.43 | 78.02 | 79.54 | 0.59 | 0.823 | IterativeClassifierOptimizer |

**TABLE 3 |** Performance of best models based on all 1428 proteins NPX values.

| Dataset (no. of proteins used) | Day(s) | Sensitivity (%) | Specificity (%) | Accuracy (%) | MCC | ROC | WEKA technique used |
|---|---|---|---|---|---|---|---|
| Whole Dataset II (1428) | 0 | 39.02 | 95.18 | 87.24 | 0.4 | 0.791 | AdaBoostM1 |
| P4 (1428) | 0 | 82.93 | 84 | 83.52 | 0.67 | 0.868 | LogitBoost |
| Average of P1–P5 splits (1428) | 0 | 69.76 | 71.90 | 70.94 | 0.42 | 0.755 | LogitBoost |

**TABLE 4 |** Performance of best models based on selected 45 protein NPX values.

| Dataset (No. of proteins used) | Day(s) | Sensitivity (%) | Specificity (%) | Accuracy (%) | MCC | ROC | WEKA technique used |
|---|---|---|---|---|---|---|---|
| Whole dataset II (45) | 0 | 80.49 | 92.77 | 91.03 | 0.67 | 0.948 | BayesNet |
| **P2 (45)** | **0** | **92.68** | **86** | **89.01** | **0.78** | **0.953** | **BayesNet** |
| Average of P1–P5 splits (45) | 0 | 82.44 | 82.72 | 82.59 | 0.65 | 0.902 | BayesNet |
| **P5 (45)** | **0** | **85.37** | **91.84** | **88.89** | **0.78** | **0.886** | **SMO; NormalizedPolyKernel** |
| Average of P1–P5 splits (45) | 0 | 83.42 | 79.97 | 81.51 | 0.63 | 0.817 | SMO; NormalizedPolyKernel |

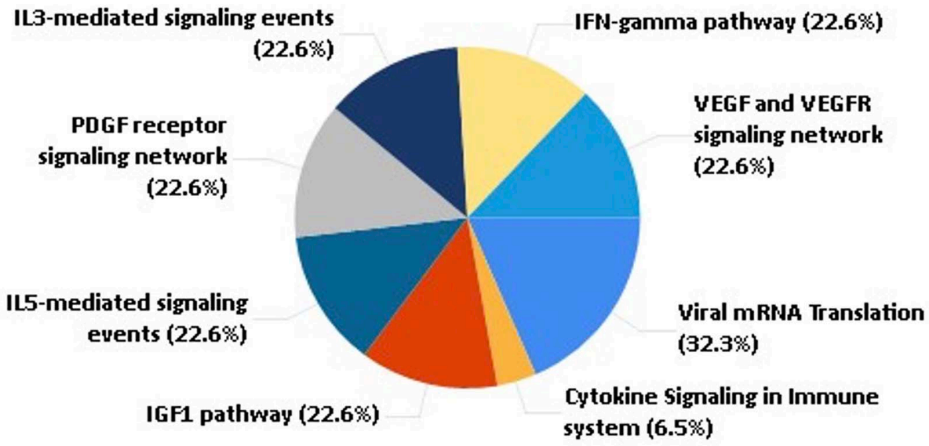



**FIGURE 3 |** Pathway analysis of the selected 45 proteins.

able to identify clinically used drugs that target 18 proteins among the shortlisted 45 proteins. The maximum number of drugs was found to target growth factor associated proteins, i.e., VGFR2 and EGFR, followed by FA7 and ANGP2 (**Supplementary Figure 1**). It is observed that during viral infection through respiratory viruses, EGFR gets activated via the NADPH oxidase signaling pathway in the airway epithelium. The activation of EGFR causes suppression of IFN regulatory factor (IRF) 1-dependent CXCL10 production showing their role in antiviral defense (Kalinowski et al., 2014).

## The Development and Utility of the CovidPrognosis Webserver

The utility of a machine learning–based method relies upon its ease of use. Therefore, to enhance the real-life usage of the developed prediction models by researchers or clinicians, we have developed the webserver CovidPrognosis. The webserver is freely available for scientific use and clinical validation at http://14.139.62.220/covidprognosis/. In the current version, the users can input three parameters for Day 0 or 33 parameters for Days 0, 3, and 7. The survival chances of the patient, represented by the input parameters, are predicted based on the user-supplied values. A detailed description of the clinical parameters is available on the CovidPrognosis webserver's website at http://14.139.62.220/covidprognosis/help.php. Day 0 denotes the day on which the patient was admitted to a hospital, while Days 3 and 7 represent the third and seventh day after hospitalization, respectively. The Day 0–based model helps in the early estimation of the seriousness of the case, while the days 0–7-based model may prove useful while monitoring the patient's health status at the time of hospital stay. **Figure 4** shows the prediction results by the CovidPrognosis webserver's three clinical parameters-based model using Day 0 clinical information of a COVID-19 patient. The webserver may prove to be a valuable resource for researchers and clinicians for independent validation and further improvement.

# DISCUSSION

COVID-19 is caused by the novel coronavirus SARS-CoV-2 that belongs to the SARS-CoV and MERS family of viruses. To date, the disease has led to millions of deaths worldwide. COVID-19 can be diagnosed by real-time PCR (RT-PCR), chest X-ray images, CT scan images, and serological blood tests (Augustine et al., 2020, p. 19). However, these diagnostic methods have low accuracy with a high false-positive rate of prediction (Surkova et al., 2020; To et al., 2020) and cannot help distinguish patients with different severity of illness. In addition to the respiratory illness, COVID-19 can cause many other illnesses such as kidney failure, heart disease, and venous thromboembolism and may damage the CNS leading to mortality (Kollias et al., 2020; Larsen et al., 2020; Shi et al., 2020; Wu et al., 2020).

The most common clinical abnormalities observed in COVID-19 positive patients are lymphopenia, leukopenia, thrombocytopenia, elevated CRP and inflammatory markers, elevated cardiac biomarkers, decreased albumin, and abnormal renal and liver function (Paranjpe et al., 2020; Zhu et al., 2020). The increase in SARS-CoV-2 spread and mortality has motivated researchers to develop vaccines or antiviral drugs. Similarly, clinicians too are trying different treatment strategies to improve prognosis, reduce treatment period, and alleviate the suffering of COVID-19 patients. Therefore, it is necessary to identify factors/biomarkers associated with the patients' mortality and survival on available patient datasets to reduce the mortality rate.

Based on clinical parameters, researchers have identified several biomarkers (using an ML-based approach) like using a multivariable logistic regression model. Yao Y. et al. (2020) showed that the value of D-dimer > 2mg/L was associated with mortality among COVID-19 patients. The group has observed a significant correlation between D-dimer levels and disease severity measured by the CT, oxygenation index, and clinical staging. Another group, Yan et al. (2020a), identified lactic dehydrogenase (LDH), lymphocyte, and high-sensitivity C-reactive protein (hs-CRP) that were associated with the survival of individual patients. Similarly, in the present study, we

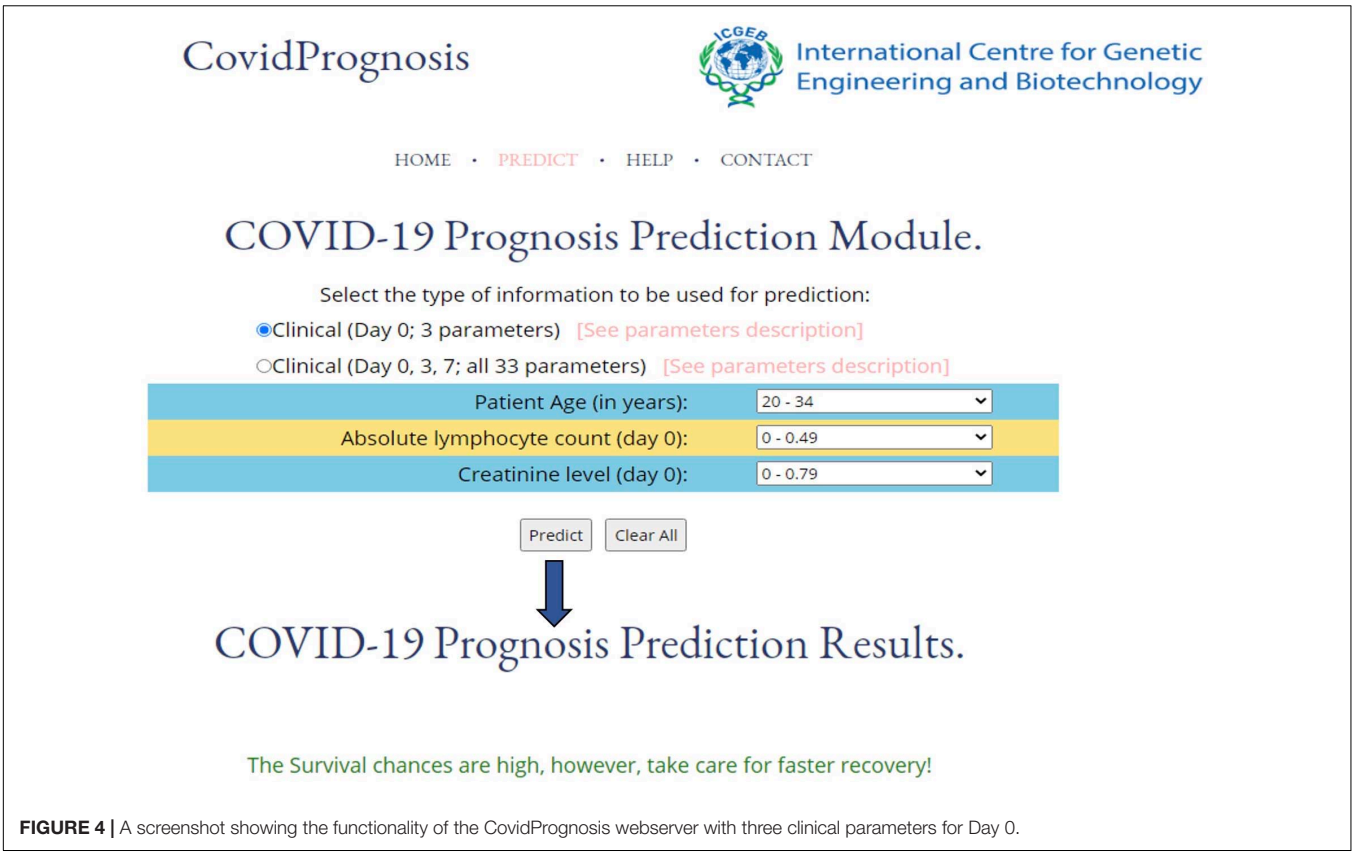



**FIGURE 4 |** A screenshot showing the functionality of the CovidPrognosis webserver with three clinical parameters for Day 0.

have applied ML-based prediction on a cohort of 306 COVID positive patients with 33 clinical parameters and 1,428 protein expression values. From the number of WEKA models on clinical data, RandomSubSpace and IterativeClassifierOptimizer perform best with the accuracy of 87.37 and 84.32%, respectively. These models identified nine shortlisted features from among 33 clinical parameters, namely, age category, absolute lymphocyte count (Day 0), creatinine level (Day 0), preexisting heart disease(s), preexisting hypertension, preexisting kidney disease(s), D-dimer level (Day 0), GI symptoms, and cardiac event-troponin level 72 h (hs-cTn = > 100 within the first 72 h of presentation). Of the nine shortlisted clinical parameters, D-dimer, lymphocyte count, and kidney disease are reported to play an important role in the survival prediction of COVID-19 patients, thus validating the findings of the present study (Cheng et al., 2020; Pan et al., 2020; Yan et al., 2020a). Moreover, some previously not identified clinical parameters such as creatinine, age, and cardiac troponin, along with GI symptoms, heart disease, and hypertension, could predict the COVID-19 prognosis and disease severity.

While employing LogitBoost on 1428 protein expression data, survival prediction models were able to achieve an accuracy of 83.52% with sensitivity (%), specificity (%), MCC, and ROC values of 82.93, 84, 0.67, and 0.868, respectively. However, the accuracy was further improved after applying the feature selection algorithms (available in WEKA), and the highest accuracy of 89.01% (with the balanced dataset) was achieved with sensitivity (%), specificity (%), MCC, and ROC values of 92.68, 86, 0.78, and 0.953, respectively. Thus, the model led to identifying 45 proteins enriched in various pathways such as angiogenesis, interleukin, cytokine, chemokine, and VEGF signaling. The enrichment of host immune system pathways suggested that SARS-CoV-2 uses the host immune system defense mechanism to hijack the body's mucous membrane cells.

Shen et al. have identified 93 proteins associated with the severity of COVID-19 disease based on the data of 46 COVID-positive patients using machine learning models (Bojkova et al., 2020; Qiu et al., 2020; Shen et al., 2020). Interestingly, some of the shortlisted 45 proteins, such as PROC, IL16, EGFR, ANGP2, APOP1, coagulation factor VII, and FEUTB (identified in the present study), are already well reported in the literature for their role in the disease prognosis and severity, thus validating the current findings (Bojkova et al., 2020; Qiu et al., 2020; Shen et al., 2020; Shu et al., 2020; Yin et al., 2020). In our analysis, other protein classes such as different growth factors and phospholipase factors are newly discovered, which can be explored further for their role in disease severity. The role of phospholipase A2 in the inhibition of coronavirus replication is well established by EM and confocal microscopy, which can also be confirmed for SARS-CoV-2 (Müller et al., 2017).

From the drug-target network construction, it is observed that FDA-approved drugs target growth factor associated proteins, i.e., VGFR2 and EGFR, followed by FA7 and ANGP2, suggesting their potential implication in drug repurposing.

From the present study, we show that the ML-based prediction/classification models can efficiently help in the

prognosis of COVID-19 patients based upon identified clinical and protein biomarkers associated with COVID-19 severity/survival. The clinicians and researchers can test new COVID-19 cases to predict the patients who are likely to survive within 28 days after hospitalization. The results obtained from the ML-based techniques may also lead to the biomarker discovery for COVID-19 for early prognosis, potentially reducing mortality rate and may also serve as useful drug targets.

To increase the utility of the present work, we have developed an easy-to-use CovidPrognosis webserver to assist researchers and clinicians in quickly evaluating the machine learning model or identifying the prognostic biomarkers associated with the survival or death of COVID-19 patients. The webserver is available at http://14.139.62.220/covidprognosis/. The current version of the model is a proof of concept that machine learning–based prognostic tools can be developed. The CovidPrognosis webserver will be regularly updated with the latest COVID-19 datasets in order to increase its efficiency, reliability, and utility.

## DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://www.olink.com/mgh-covid-study/.

## AUTHOR CONTRIBUTIONS

DG, AS, and RS conceptualized the study, analyzed the data, and prepared the manuscript. AS carried out the machine learning studies. All authors reviewed and approved the final version.

## FUNDING

This work was financially supported by the Department of Biotechnology (DBT), Government of India, grant no. BT/PR40151/BTIS/137/5/2021, awarded to DG. Financial support provided by the Indian Council of Medical Research (ICMR), India to RS as Senior Research Fellowship is duly acknowledged (2019-5850).

## ACKNOWLEDGMENTS

We acknowledge ICGEB for providing the necessary infrastructure and facilities for the research.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.636441/full#supplementary-material

**Supplementary Figure 1 |** Drug–target network of the top proteins among the selected 45 proteins.

**Supplementary Table 1 |** Clinical information or parameters used for the generation of clinical information-based models.

**Supplementary Table 2 |** Description of shortlisted 45 proteins useful in the classification of survived vs. died COVID-19 patients.

**Supplementary Table 3 |** NPX expression values for 45 shortlisted proteins.

**Supplementary Table 4 |** Drug–target interactions retrieved from TTD.

## REFERENCES

Al-Rohaimi, A. H., and Al Otaibi, F. (2020). Novel SARS-CoV-2 outbreak and COVID19 disease; a systemic review on the global pandemic. *Genes Dis.* 7, 491–501. doi: 10.1016/j.gendis.2020.06.004

Augustine, R., Das, S., Hasan, A., S, A., Abdul Salam, S., Augustine, P., et al. (2020). Rapid antibody-based COVID-19 mass surveillance: relevance, challenges, and prospects in a pandemic and post-pandemic world. *J. Clin. Med.* 9:3372. doi: 10.3390/jcm9103372

Bojkova, D., Klann, K., Koch, B., Widera, M., Krause, D., Ciesek, S., et al. (2020). Proteomics of SARS-CoV-2-infected host cells reveals therapy targets. *Nature* 583, 469–472. doi: 10.1038/s41586-020-2332-7

Cai, Y., Winn, M. E., Zehmer, J. K., Gillette, W. K., Lubkowski, J. T., Pilon, A. L., et al. (2014). Preclinical evaluation of human secretoglobin 3A2 in mouse models of lung development and fibrosis. *Am. J. Physiol. Lung Cell. Mol. Physiol.* 306, L10–L22. doi: 10.1152/ajplung.00037.2013

Cambiaghi, A., Díaz, R., Martinez, J. B., Odena, A., Brunelli, L., Caironi, P., et al. (2018). An innovative approach for the integration of proteomics and metabolomics data in severe septic shock patients stratified for mortality. *Sci. Rep.* 8:6681. doi: 10.1038/s41598-018-25035-1

Cascella, M., Rajnik, M., Cuomo, A., Dulebohn, S. C., and Di Napoli, R. (2020). *Features, Evaluation, and Treatment of Coronavirus (COVID-19)*. Treasure Island, FL: StatPearls Publishing.

Cheng, Y., Luo, R., Wang, K., Zhang, M., Wang, Z., Dong, L., et al. (2020). Kidney disease is associated with in-hospital death of patients with COVID-19. *Kidney Int.* 97, 829–838. doi: 10.1016/j.kint.2020.03.005

Dumancas, G. G., Adrianto, I., Bello, G., and Dozmorov, M. (2017). Current developments in machine learning techniques in biological data mining. *Bioinform. Biol. Insights* 11, 1–4. doi: 10.1177/1177932216687545

Filbin, M., Goldberg, M., and Hacohen, N. *Data Provided by the MGH Emergency Department COVID-19 Cohort with O-Link Proteomics*. Available online at: https://www.olink.com/mgh-covid-study/

Frank, E., Hall, M. A., Pal, C. J., and Witten, I. H. (2016). *The WEKA Workbench. Online Appendix for Data Mining: Practical Machine Learning Tools and Techniques*, 4th Edn. Burlington, MA: Mourghan Kaufamnn.

González-Pacheco, H., Amezcua-Guerra, L. M., Sandoval, J., and Arias-Mendoza, A. (2020). Potential usefulness of pentoxifylline, a non-specific phosphodiesterase inhibitor with anti-inflammatory, anti-thrombotic, antioxidant, and anti-fibrogenic properties, in the treatment of SARS-CoV-2. *Eur. Rev. Med. Pharmacol. Sci.* 24, 7612–7614. doi: 10.26355/eurrev_202007_21921

Graziani, D., Soriano, J. B., Del Rio-Bermudez, C., Morena, D., Díaz, T., Castillo, M., et al. (2020). Characteristics and prognosis of COVID-19 in patients with COPD. *J. Clin. Med.* 9:3259. doi: 10.3390/jcm9103259

Gu, J.-G., Zhu, C., Cheng, D., Xie, Y., Liu, F., and Zhou, X. (2011). Enchanced levels of apolipoprotein M during HBV infection feedback suppresses HBV replication. *Lipids Health Dis.* 10:154. doi: 10.1186/1476-511X-10-154

Jablonka, K. M., Ongari, D., Moosavi, S. M., and Smit, B. (2020). Big-data science in porous materials: materials genomics and machine learning. *Chem. Rev.* 120, 8066–8129. doi: 10.1021/acs.chemrev.0c00004

Jiang, M., Mieronkoski, R., Syrjälä, E., Anzanpour, A., Terävä, V., Rahmani, A. M., et al. (2019). Acute pain intensity monitoring with the classification of multiple physiological parameters. *J. Clin. Monit. Comput.* 33, 493–507. doi: 10.1007/s10877-018-0174-8

Jiao, X., Sherman, B. T., Huang, D. W., Stephens, R., Baseler, M. W., Lane, H. C., et al. (2012). DAVID-WS: a stateful web service to facilitate gene/protein list analysis. *Bioinformatics* 28, 1805–1806. doi: 10.1093/bioinformatics/bts251

Kalinowski, A., Ueki, I., Min-Oo, G., Ballon-Landa, E., Knoff, D., Galen, B., et al. (2014). EGFR activation suppresses respiratory virus-induced IRF1-dependent CXCL10 production. *Am. J. Physiol. Lung Cell. Mol. Physiol.* 307, L186–L196. doi: 10.1152/ajplung.00368.2013

Kaur, M., Tiwari, S., and Jain, R. (2020). Protein based biomarkers for non-invasive Covid-19 detection. *Sens. Bio Sens. Res.* 29:100362. doi: 10.1016/j.sbsr.2020.100362

Kermali, M., Khalsa, R. K., Pillai, K., Ismail, Z., and Harky, A. (2020). The role of biomarkers in diagnosis of COVID-19 – a systematic review. *Life Sci.* 254:117788. doi: 10.1016/j.lfs.2020.117788

Kollias, A., Kyriakoulis, K. G., Dimakakos, E., Poulakou, G., Stergiou, G. S., and Syrigos, K. (2020). Thromboembolic risk and anticoagulant therapy in COVID-19 patients: emerging evidence and call for action. *Br. J. Haematol.* 189, 846–847. doi: 10.1111/bjh.16727

Kumar, S. N., Saxena, P., Patel, R., Sharma, A., Pradhan, D., Singh, H., et al. (2020). Predicting risk of low birth weight offspring from maternal features and blood polycyclic aromatic hydrocarbon concentration. *Reprod. Toxicol.* 94, 92–100. doi: 10.1016/j.reprotox.2020.03.009

Lalmuanawma, S., Hussain, J., and Chhakchhuak, L. (2020). Applications of machine learning and artificial intelligence for Covid-19 (SARS-CoV-2) pandemic: a review. *Chaos Solitons Fractals* 139:110059. doi: 10.1016/j.chaos.2020.110059

Larsen, C. P., Bourne, T. D., Wilson, J. D., Saqqa, O., and Sharshir, M. A. (2020). Collapsing glomerulopathy in a patient with COVID-19. *Kidney Int. Rep.* 5, 935–939. doi: 10.1016/j.ekir.2020.04.002

Mete, M., Sakoglu, U., Spence, J. S., Devous, M. D., Harris, T. S., and Adinoff, B. (2016). Successful classification of cocaine dependence using brain imaging: a generalizable machine learning approach. *BMC Bioinformatics* 17(Suppl. 13):357. doi: 10.1186/s12859-016-1218-z

Müller, C., Hardt, M., Schwudke, D., Neuman, B. W., Pleschka, S., and Ziebuhr, J. (2017). Inhibition of cytosolic phospholipase A2α impairs an early step of coronavirus replication in cell culture. *J. Virol.* 92:JVI.01463-17. doi: 10.1128/JVI.01463-17

Nath, A., and Subbiah, K. (2016). Probing an optimal class distribution for enhancing prediction and feature characterization of plant virus-encoded RNA-silencing suppressors. *3 Biotech* 6:93. doi: 10.1007/s13205-016-0410-1

Overmyer, K. A., Shishkova, E., Miller, I. J., Balnis, J., Bernstein, M. N., Peters-Clarke, T. M., et al. (2020). Large-scale multi-omic analysis of COVID-19 severity. *Cell Syst.* 12, 23–40. doi: 10.1016/j.cels.2020.10.003

Pan, F., Yang, L., Li, Y., Liang, B., Li, L., Ye, T., et al. (2020). Factors associated with death outcome in patients with severe coronavirus disease-19 (COVID-19): a case-control study. *Int. J. Med. Sci.* 17, 1281–1292. doi: 10.7150/ijms.46614

Pandey, S. C., Pande, V., Sati, D., Upreti, S., and Samant, M. (2020). Vaccination strategies to combat novel corona virus SARS-CoV-2. *Life Sci.* 256:117956. doi: 10.1016/j.lfs.2020.117956

Paranjpe, I., Russak, A., De Freitas, J. K., Lala, A., Miotto, R., Vaid, A., et al. (2020). Clinical characteristics of hospitalized Covid-19 patients in New York city. *medRxiv* [Preprint] doi: 10.1101/2020.04.19.20062117

Qiu, Y., Wu, D., Ning, W., Zhang, J., Shu, T., Huang, C., et al. (2020). *Postmortem Tissue Proteomics Reveals The Pathogenesis of Multiorgan Injuries of COVID-19*. Durham, NC: Research Squrae. doi: 10.21203/rs.3.rs-38091/v1

Saha, A., and Anirvan, P. (2020). Cancer progression in COVID-19: integrating the roles of renin angiotensin aldosterone system, angiopoietin-2, heat shock protein-27 and epithelial mesenchymal transition. *Ecancermedicalscience* 14:1099. doi: 10.3332/ecancer.2020.1099

Sardar, R., Satish, D., Birla, S., and Gupta, D. (2020). Integrative analyses of SARS-CoV-2 genomes from different geographical locations reveal unique features potentially consequential to host-virus interaction, pathogenesis and clues for novel therapies. *Heliyon* 6:e04658. doi: 10.1016/j.heliyon.2020.e04658

Sharma, A., Gupta, P., Kumar, R., and Bhardwaj, A. (2016). dPABBs: a novel in silico approach for predicting and designing anti-biofilm peptides. *Sci. Rep.* 6:21839. doi: 10.1038/srep21839

Sharma, A., Rani, S., and Gupta, D. (2020). Artificial intelligence-based classification of chest X-ray images into COVID-19 and other infectious diseases. *Int. J. Biomed. Imaging* 2020, 1–10. doi: 10.1155/2020/8889023

Shen, B., Yi, X., Sun, Y., Bi, X., Du, J., Zhang, C., et al. (2020). Proteomic and metabolomic characterization of COVID-19 patient sera. *Cell* 182, 59–72.e15. doi: 10.1016/j.cell.2020.05.032

Shi, S., Qin, M., Shen, B., Cai, Y., Liu, T., Yang, F., et al. (2020). Association of cardiac injury with mortality in hospitalized patients with COVID-19 in Wuhan, China. *JAMA Cardiol.* 5:802. doi: 10.1001/jamacardio.2020.0950

Shu, T., Ning, W., Wu, D., Xu, J., Han, Q., Huang, M., et al. (2020). Plasma proteomics identify biomarkers and pathogenesis of COVID-19. *Immunity* 53, 1108–1122.e5. doi: 10.1016/j.immuni.2020.10.008

Srivastava, N., Baxi, P., Ratho, R. K., and Saxena, S. K. (2020). "Global trends in epidemiology of coronavirus disease 2019 (COVID-19)," in *Coronavirus Disease 2019 (COVID-19)* Medical Virology: From Pathogenesis to Disease Control, ed. S. K. Saxena (Singapore: Springer Singapore), 9–21. doi: 10.1007/978-981-15-4814-7_2

Surkova, E., Nikolayevskyy, V., and Drobniewski, F. (2020). False-positive COVID-19 results: hidden problems and costs. *Lancet Respir. Med.* 8, 1167–1168. doi: 10.1016/S2213-2600(20)30453-7

To, K. K.-W., Hung, I. F.-N., Ip, J. D., Chu, A. W.-H., Chan, W.-M., Tam, A. R., et al. (2020). Coronavirus disease 2019 (COVID-19) re-infection by a phylogenetically distinct severe acute respiratory syndrome coronavirus 2 strain confirmed by whole genome sequencing. *Clin. Infect. Dis.* ciaa1275. doi: 10.1093/cid/ciaa1275

Wang, Y., Zhang, S., Li, F., Zhou, Y., Zhang, Y., Wang, Z., et al. (2020). Therapeutic target database 2020: enriched resource for facilitating research and early development of targeted therapeutics. *Nucleic Acids Res.* 48, D1031–D1041. doi: 10.1093/nar/gkz981

Wu, Y., Xu, X., Chen, Z., Duan, J., Hashimoto, K., Yang, L., et al. (2020). Nervous system involvement after infection with COVID-19 and other coronaviruses. *Brain Behav. Immun.* 87, 18–22. doi: 10.1016/j.bbi.2020.03.031

Wu, Z., and McGoogan, J. M. (2020). Characteristics of and important lessons from the coronavirus disease 2019 (COVID-19) outbreak in China: summary of a report of 72 314 cases from the Chinese center for disease control and prevention. *JAMA* 323:1239. doi: 10.1001/jama.2020.2648

Wynants, L., Van Calster, B., Collins, G. S., Riley, R. D., Heinze, G., Schuit, E., et al. (2020). Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. *BMJ* 369:m1328. doi: 10.1136/bmj.m1328

Yan, L., Zhang, H.-T., Goncalves, J., Xiao, Y., Wang, M., Guo, Y., et al. (2020a). A machine learning-based model for survival prediction in patients with severe COVID-19 infection. *medRxiv* [Preprint]. doi: 10.1101/2020.02.27.20028027

Yan, L., Zhang, H.-T., Goncalves, J., Xiao, Y., Wang, M., Guo, Y., et al. (2020b). An interpretable mortality prediction model for COVID-19 patients. *Nat. Mach. Intell.* 2, 283–288. doi: 10.1038/s42256-020-0180-7

Yao, H., Zhang, N., Zhang, R., Duan, M., Xie, T., Pan, J., et al. (2020). Severity detection for the coronavirus disease 2019 (COVID-19) patients using a machine learning model based on the blood and urine tests. *Front. Cell Dev. Biol.* 8:683. doi: 10.3389/fcell.2020.00683

Yao, Y., Cao, J., Wang, Q., Shi, Q., Liu, K., Luo, Z., et al. (2020). D-dimer as a biomarker for disease severity and mortality in COVID-19 patients: a case control study. *J. Intensive Care* 8:49. doi: 10.1186/s40560-020-00466-z

Yin, X.-X., Zheng, X.-R., Peng, W., Wu, M.-L., and Mao, X.-Y. (2020). Vascular endothelial growth factor (VEGF) as a vital target for brain inflammation during the COVID-19 outbreak. *ACS Chem. Neurosci.* 11, 1704–1705. doi: 10.1021/acschemneuro.0c00294

Zhang, X., Tan, Y., Ling, Y., Lu, G., Liu, F., Yi, Z., et al. (2020). Viral and host factors related to the clinical outcome of COVID-19. *Nature* 583, 437–440. doi: 10.1038/s41586-020-2355-0

Zhong, J., Tang, J., Ye, C., and Dong, L. (2020). The immunology of COVID-19: is immune modulation an option for treatment? *Lancet Rheumatol.* 2, e428–e436. doi: 10.1016/S2665-9913(20)30120-X

Zhu, J., Zhong, Z., Ji, P., Li, H., Li, B., Pang, J., et al. (2020). Clinicopathological characteristics of 8697 patients with COVID-19 in China: a meta-analysis. *Fam. Med. Commun. Health* 8:e000406. doi: 10.1136/fmch-2020-000406

**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 Physiology

ORIGINAL RESEARCH
published: 24 May 2021
doi: 10.3389/fphys.2021.645212

# Exploration of Crucial Mediators for Carotid Atherosclerosis Pathogenesis Through Integration of Microbiome, Metabolome, and Transcriptome

*Lei Ji[1†], Siliang Chen[1†], Guangchao Gu[1], Jiawei Zhou[1], Wei Wang[1], Jinrui Ren[1], Jianqiang Wu[2], Dan Yang[3] and Yuehong Zheng[1*]*

[1] Department of Vascular Surgery, Peking Union Medical College Hospital, Chinese Academy of Medical Sciences, Peking Union Medical College, Beijing, China, [2] Medical Research Center, Peking Union Medical College Hospital, Chinese Academy of Medical Sciences, Peking Union Medical College, Beijing, China, [3] Department of Computational Biology and Bioinformatics, Chinese Academy of Medical Sciences, Peking Union Medical College, Institute of Medicinal Plant Development, Beijing, China

OPEN ACCESS

**Edited by:**
Sanjay Kumar Banerjee, National Institute of Pharmaceutical Education and Research (Guwahati), India

**Reviewed by:**
Padhmanand Sudhakar, KU Leuven, Belgium
Andreas Dräger, University of Tübingen, Germany

***Correspondence:**
Yuehong Zheng
yuehongzheng@yahoo.com

[†] These authors have contributed equally to this work and share the first authorship

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

**Received:** 22 December 2020
**Accepted:** 12 April 2021
**Published:** 24 May 2021

**Citation:**
Ji L, Chen S, Gu G, Zhou J, Wang W, Ren J, Wu J, Yang D and Zheng Y (2021) Exploration of Crucial Mediators for Carotid Atherosclerosis Pathogenesis Through Integration of Microbiome, Metabolome, and Transcriptome. Front. Physiol. 12:645212. doi: 10.3389/fphys.2021.645212

**Background:** Carotid atherosclerosis (CAS) is an important cause of stroke. Although interactions between the gut microbiome and metabolome have been widely investigated with respect to the pathogenesis of cardiovascular diseases, information regarding CAS remains limited.

**Materials and Methods:** We utilized 16S ribosomal DNA sequencing and untargeted metabolomics to investigate the alterations in the gut microbiota and plasma metabolites of 32 CAS patients and 32 healthy controls. The compositions of the gut microbiota differed significantly between the two groups, and a total of 11 differentially enriched genera were identified. In the metabolomic analysis, 11 and 12 significantly changed metabolites were screened in positive (POS) and negative (NEG) modes, respectively. α-N-Phenylacetyl-L-glutamine was an upregulated metabolite in CAS patients detected in both POS and NEG modes and had the highest $|\log_2(\text{fold change})|$ in POS mode. In addition, transcriptomic analysis was performed using the GSE43292 dataset.

**Results:** A total of 132 differentially expressed genes (DEGs) were screened. Among the upregulated DEGs in CAS patients, FABP4 exhibited the highest $|\log_2(\text{fold change})|$. Furthermore, FABP4 was positively associated with *Acidaminococcus* and had the highest Spearman's correlation coefficient and the most significant *p*-value among the microbiota–DEG pairs.

**Conclusion:** In this study, we investigated the potential "microbiota–metabolite–gene" regulatory axis that may act on CAS, and our results may help to establish a theoretical basis for further specialized study of this disease.

**Keywords: carotid atherosclerosis, microbiome, metabolome, transcriptome, correlation analysis**

## INTRODUCTION

Atherosclerosis (AS) is a diffuse, slowly progressing disease that affects large- and medium-sized arteries. Advanced atherosclerotic plaques can invade the arterial lumen, impeding blood flow, and resulting in tissue ischemia (Faxon et al., 2004; Libby et al., 2019). Carotid atherosclerosis (CAS) is a preventable cause of 20–30% of stroke, approximately 21% of people aged 30–79 years have carotid plaque, and 1.5% have carotid stenosis (Petty et al., 2000; Song et al., 2020). The pathophysiological features of AS are primarily linked to lipid accumulation, chronic inflammation, calcification, and thrombosis (Libby et al., 2019). Many studies have vastly improved our understanding of the pathogenesis of AS, but, despite these advances, we still lack definitive evidence to translate basic results to the bedside (Weber and Noels, 2011). Although AS is a systemic disease sharing common major risk factors, differences exist in the strength and impact per arterial site (Aboyans et al., 2018). Medical interventions that result in the prevention of CAS are especially centered on statins, but which are not targeted enough when CAS is regarded as a unique form of AS (Artom et al., 2014).

Findings from the past decade have suggested that the structure and composition of the gut microbiota are associated with AS in humans and animal models (Jonsson and Backhed, 2017). The contributions of the gut microbiota to AS can be divided into three main categories. First, local or distant infections might aggravate atherogenesis. Second, patients with AS have altered lipid metabolism, and bacterial taxa in the gut were observed to correlate with plasma cholesterol levels (Koren et al., 2011). Third, diet and specific components that are metabolized by gut microbiota can have various effects on AS. Metabolites filtered or produced by gut microbiota, such as trimethylamine-N-oxide, short-chain fatty acids (SCFAs), and secondary bile acids, have been observed to affect the development of AS (Wang et al., 2011; Wahlstrom et al., 2016; Chen et al., 2018). Most studies of the relationship between CAS and microbiota could be classified into the first category mentioned earlier. A wide variety of microbial DNA has accordingly been found in carotid atherosclerotic plaques in different populations (Ziganshina et al., 2016; Lindskog Jonsson et al., 2017). Bacteria observed in the atherosclerotic plaques are also detected at other body sites, predominantly the gut, which might thus serve as reservoirs of these potentially pathogenic microorganisms (Jonsson and Backhed, 2017). However, limited information is available focusing on the gut microbiota composition in CAS patients. With respect to metabolomics, several studies have found a number of metabolites associated with CAS on the different stages (Vojinovic et al., 2018; Lee T. H. et al., 2019), which were used as non-causal biomarkers, but further study is necessary to elucidate the pathogenesis of CAS. Also, considerable uncertainty remains concerning the relationship between CAS and metabolites.

Taken together, both human and animal studies have indicated that alterations of the gut microbiota and plasma metabolites might be involved in the progression of AS, but the details of these alterations in patients with CAS have not been fully characterized. To address this question, we performed multi-omics combined 16S ribosomal DNA (rDNA) gene sequencing using fecal samples and untargeted liquid chromatography–mass spectrometry using plasma samples from 32 CAS patients and 32 healthy controls with gene expression profiling from the Gene Expression Omnibus (GEO) database to characterize the gut microbial community and plasma metabolic profiles. Also, we performed an integrated analysis of the microbiome, metabolome, and transcriptome. These results may ultimately provide a more in-depth understanding of the "microbiota–metabolite–gene" axis in the pathogenesis of CAS.

**Abbreviations:** AS, atherosclerosis; CAS, carotid atherosclerosis; CKD, chronic kidney disease; DEG, differentially expressed gene; EPA, eicosapentaenoic acid; GEO, Gene Expression Omnibus; GO-BP, gene ontology-biological process; KEGG, Kyoto Encyclopedia of Genes and Genomes; LDA, linear discriminant analysis; LEfSe, linear discriminant analysis effect size; NEG, negative mode; NF-κB, nuclear factor kappa-B; OTU, operational taxonomic unit; PAGly, phenylacetylglycine; PCA, principal component analysis; PERMANOVA, permutational multivariate analysis of variance; PICRUSt, phylogenetic investigation of communities by reconstruction of unobserved states; POS, positive mode; QIIME, quantitative insights into microbial ecology; RF, random forest; ROC, receiver operating characteristic; SCFA, short-chain fatty acid; UHPLC-QTOFMS, ultra-high-performance liquid tandem chromatography/quadrupole time-of-flight mass spectrometry.

## MATERIALS AND METHODS

### Medical Ethics

The Ethics Committee of the Peking Union Medical College Hospital (PUMCH) has approved this study (institutional approval number: JS-2629). Each participant provided signed informed consent before participating in the present study.

### Patients Recruitment

CAS patients were recruited from the Department of Vascular Surgery, Peking Union Medical College Hospital. The inclusion criteria for recruitment were as follows: (1) Diagnosis with carotid atherosclerosis by ultrasound or CT angiography; (2) age $\geq$ 45 years; the exclusion criteria were applied to both CAS patients and healthy controls: (1) Antibiotic usage within 6 months; (2) probiotic usage within 6 months; (3) history of gastrointestinal diseases (such as inflammatory bowel disease); (4) history of abdominal surgery (such as gastrectomy); (5) major dietary change 1 week before sample collection.

We first recruited 71 CAS patients and 39 healthy controls. Next, 39 CAS patients were excluded due to antibiotic usage ($n$ = 14), probiotic usage ($n$ = 6), digestive disease ($n$ = 8), and abdominal surgery ($n$ = 11). Meanwhile, seven healthy controls were excluded due to antibiotic usage ($n$ = 4), probiotic usage ($n$ = 2), and abdominal surgery ($n$ = 1). Finally, each group had 32 subjects for further analysis.

### Sample Collection

Peripheral blood and stool samples were collected in the morning after an overnight fast ($\geq$ 8 h). Plasma samples were obtained by centrifugation at 3,000 rpm for 10 min at room temperature. All plasma and stool samples were rapidly frozen and stored at −80°C until analysis.

## Genomic DNA Extraction and 16S Ribosomal DNA Sequencing

Genomic DNA extraction was performed using QIAamp® Fast DNA Stool Mini Kit (Qiagen, Hilden, Germany) and examined using Thermo NanoDrop 2000 (Thermo Fisher Scientific, New York, NY, United States). The V3-V4 region of the bacterial 16S rDNA was amplified using KAPA HiFi Hotstart ReadyMix PCR Kit (KAPA Biosystems, Wilmington, MA, United States) with the primers 314F (CCTACGGGRSGCAGCAG) and 806R (GGACTACVVGGGTATCTAATC) and sequenced using an Illumina PE250 platform (Illumina, California, United States).

## Ultra-High-Performance Liquid Tandem Chromatography/Quadrupole Time-of-Flight Mass Spectrometry Metabolomic Profiling of Patient Plasma Samples

Plasma samples of patients were prepared for ultra-high-performance liquid tandem chromatography/quadrupole time-of-flight mass spectrometry (UHPLC-QTOFMS) analysis by application of validated protocols (Dunn et al., 2011). The UHPLC separation was carried out using a 1290 Infinity series UHPLC System (Agilent Technologies Inc., Santa Clara, California, United States), equipped with a UPLC BEH Amide column. The TripleTOF 6600 mass spectrometry (AB Sciex, Foster City, CA, United States) was used for its ability to acquire tandem mass spectrometry spectra on an information-dependent basis during a liquid chromatography–mass spectrometry experiment. Both positive ion mode (POS) and negative ion mode (NEG) were used to obtain maximal coverage for plasma metabolites.

## Transcriptomic Profiling of Atherosclerotic Samples From the Gene Expression Omnibus Database

To have a comprehensive understanding of CAS pathogenesis from a multi-omics perspective, we also acquired transcriptomic profiling data of CAS samples from the GEO database (GSE43292) (Edgar et al., 2002). The transcriptomic dataset was not measured from the same cohort of patients from whom the 16S and metabolomic datasets were generated. The probes in the series matrix file were annotated by gene symbols using the platform data table (GPL6244), and a gene expression matrix was obtained for further transcriptomic analysis.

## Statistical Analysis

Operational taxonomic units (OTUs) were obtained by ultra-fast sequence analysis (USEARCH) v11.0 with a sequence similarity of 0.97 (Edgar, 2013). α- and β-diversities were calculated using Quantitative Insights Into Microbial Ecology (QIIME, version 1.7.0) based on OTU counts (Caporaso et al., 2010). The "vegan" package in R version 3.6.2 was used to perform a permutational multivariate analysis of variance (PERMANOVA) to compare β-diversity between the two groups. Next, we performed a differential abundance analysis using the linear discriminant analysis (LDA) method on the LDA effect size (LEfSe) platform and the Wilcoxon rank-sum test (Segata et al., 2011). To determine the functional alterations in the gut microbiota of CAS patients, Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analysis was conducted through phylogenetic investigation of communities by reconstruction of unobserved states (PICRUSt) to predict the functional composition profiles of microbiota based on OTUs (Langille et al., 2013).

UHPLC-QTOFMS data were analyzed by SMICA (version 15.0.2, Sartorius Stedim Data Analytics AB, Umea, Sweden) to conduct multivariate statistical analysis. Differential metabolites were obtained by comparing CAS patients and healthy controls using a *t*-test. KEGG pathway analysis was also conducted for these different metabolites.

For transcriptomic profiling data, differentially expressed gene (DEG) analysis was performed based on gene expression matrix using the "limma" R package. A $|\log_2(\text{fold-change})|$ of $> 1$ and an adjusted $p$-value $< 0.01$ were selected as the threshold for DEG screening. In addition to KEGG pathway analysis, gene ontology-biological process (GO-BP) analysis was conducted using the Database for Annotation, Visualization and Integrated Discovery version 6.8[1]; the enrichment analysis was also conducted using Reactome version 75[2] to further demonstrate the biological functions of DEGs.

To integrate the multi-omics data, Spearman's correlation indices between differential omics data were calculated and visualized by heatmap (Shannon et al., 2003). Finally, receiver operating characteristic (ROC) analysis was performed using Statistical Product and Service Solutions version 25.0 (SPSS Inc., 2017, Chicago, IL, United States). Random forest (RF) analysis was conducted using the Biomarker analysis section of MetaboAnalyst version 3.0 (www.metaboanalyst.ca). The area under the curve was calculated to demonstrate the potential diagnostic value of differentially enriched genera, metabolites, and genes.

To further improve the accuracy of the analyses of microbiome and metabolome, the adjustment for covariates in differential genera and metabolites was performed. First, in the microbiome analysis, the associations between genera and clinical characteristics of CAS patients and healthy controls were evaluated using a generalized linear model, and $p < 0.05$ was considered to be statistically significant (Qian et al., 2018). Second, in the metabolome analysis, PERMANOVA was used to test the statistically significant differences between metabolic profiles and clinical characteristics. The $p$-value was corrected for multiple tests using a cutoff of 0.05.

# RESULTS

## Flowchart of Our Study

The workflow of our study is shown in **Figure 1**. A total of 32 CAS patients and 32 healthy controls were included in our study. Fecal and plasma samples were taken for

[1]https://david.ncifcrf.gov/

[2]https://reactome.org/

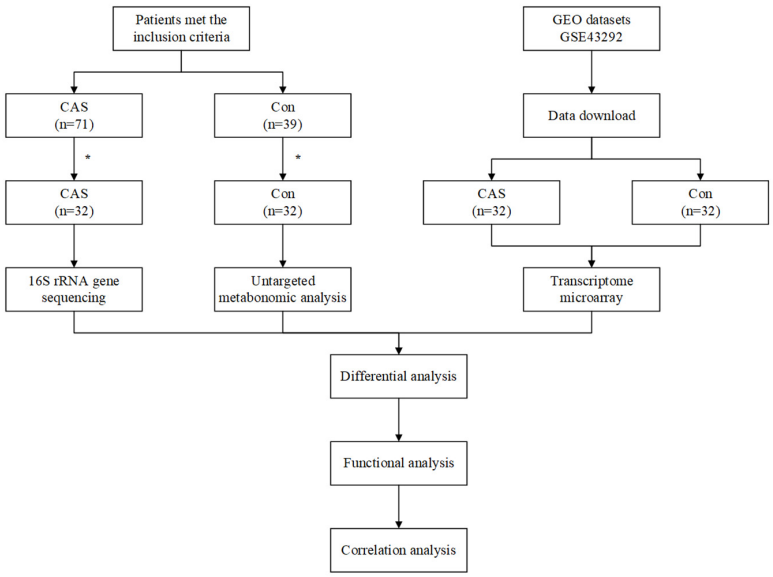



**FIGURE 1 |** Workflow of our study. *A total of 39 patients were exclude in CAS Group, due to antibiotic usage ($n$ = 14), probiotic usage ($n$ = 6), digestive disease ($n$ = 8), and abdominal surgery ($n$ = 11). A total of seven healthy controls were excluded due to antibiotic usage ($n$ = 4), probiotic usage ($n$ = 2), and abdominal surgery ($n$ = 1).

microbiome and metabolome analysis, respectively. Differentially enriched microbiota and metabolites were identified. KEGG pathways were predicted to show the functional composition profiles of differentially enriched microbiota and metabolites. Then, DEG and related functional annotation analyses were conducted based on a messenger RNA (mRNA) microarray dataset (GSE43292) to explore the differences between CAS patients and healthy controls from a transcriptomic level. Furthermore, correlation analyses were performed between differentially enriched microbiota, metabolites, and DEGs to integrate omics. Finally, the potential clinical significance of differentially enriched microbiota, metabolites, and DEGs was determined by ROC and RF analyses.

## Clinical Characteristics of Carotid Atherosclerosis Patients and Controls

For microbiome and metabolome analysis, 64 fecal and plasma samples were used for 16S rDNA sequencing and untargeted metabolomic analysis (UHPLC-QTOFMS). The baseline of our study cohort is shown in **Table 1**. Although the body mass index is marginally higher in the CAS group (24.7 ± 2.7 for CAS patients and 23.2 ± 2.2 for healthy controls, $p$ = 0.047), there were no significant differences in age and sex between CAS patients and healthy controls.

**TABLE 1 |** Characteristics of study cohort.

| | Control ($n$ = 32) | CAS ($n$ = 32) | $p$-value |
|---|---|---|---|
| Age (years) | 66.2 ± 4.8 | 64.5 ± 6.7 | 0.263 |
| Male (%) | 28 (87.5) | 28 (87.5) | > 0.999 |
| BMI (kg/m$^2$) | 23.2 ± 2.2 | 24.7 ± 2.7 | 0.047* |
| Hypertension (%) | 6 (18.8) | 12 (37.5) | 0.095 |
| Diabetes (%) | 3 (9.4) | 6 (18.8) | 0.281 |
| Coronary heart disease (%) | 0 (0) | 9 (28.1) | < 0.001* |
| White blood cell (× 10$^9$/L) | 6.5 ± 1.3 | 6.2 ± 1.5 | 0.420 |
| Monocyte (× 10$^9$/L) | 0.37 ± 0.15 | 0.38 ± 0.10 | 0.136 |
| Hcy (μmol/L) | 16.5 ± 7.0 | 16.6 ± 6.3 | 0.958 |
| TC (mmol/L) | 4.5 ± 1.3 | 3.2 ± 0.7 | < 0.001* |
| TG (mmol/L) | 1.3 ± 0.8 | 1.2 ± 0.6 | 0.754 |
| HDL-C (mmol/L) | 1.3 ± 0.3 | 1.0 ± 0.2 | < 0.001* |
| LDL-C (mmol/L) | 2.9 ± 0.6 | 1.8 ± 0.6 | < 0.001* |

*Normally distributed variables between two groups were analyzed by Student's t-test. Mann–Whitney U test was applied for data of this type that were not normally distributed. $\chi^2$-Square test or Fisher's exact test compared categorical variables. *$p < 0.05$.*

## Microbial Profiling of Carotid Atherosclerosis Patients and Controls

### Gut Microbiota Richness, Composition, and Diversity

We used 2,300,644 high-quality reads from 64 patients for downstream analysis. The rarefaction curves of richness (observed_species and chao1) were plotted. Curves for the CAS and control groups were near saturation as the reads increased, suggesting that the sequencing depth was adequate (**Supplementary Figures 1A,B**). The Venn diagram showed overlapping and different enriched OTUs in each group (**Supplementary Figure 1C**). Next, OTUs were annotated using the Ribosomal Database Project database[3], and the relative abundance of the gut microbiota is shown (**Figure 2A** and **Supplementary Figures 1D–G**).

[3]http://rdp.cme.msu.edu/

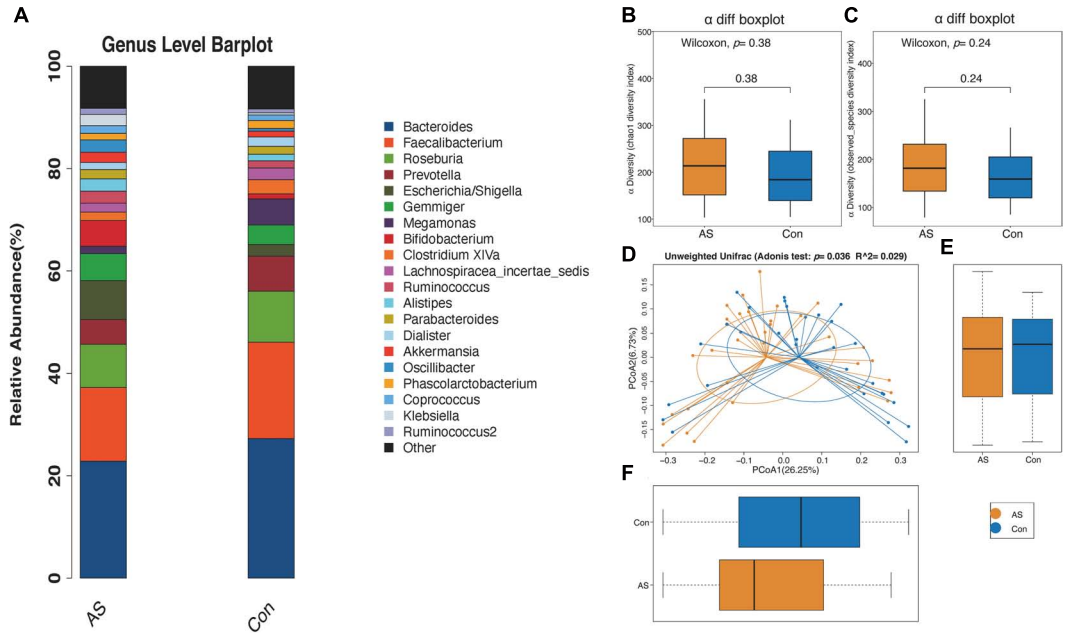



**FIGURE 2 |** Microbial diversity of CAS patients and healthy controls. **(A)** Relative abundance of gut microbiota for two groups from genus level. **(B,C)** α-Diversity (chao1 and observed_species) for two groups showed no significant difference. **(D–F)** β-Diversity is significantly increased for CAS patients.

The microbial α-diversity is shown in **Supplementary Table 1**. The Wilcoxon rank-sum test compared the α-diversity between the two groups, and no significant difference was found, which was also consistent with the rarefaction curve (**Figures 2B,C**). However, the gut microbiota communities between the CAS group and healthy control group were significantly different, as shown by β-diversity (**Figures 2D–F**).

## Differential Gut Microbiota Enriched in Carotid Atherosclerosis Patients and Healthy Controls

Using LEfSe analysis, we screened 29 different features at the phylum ($n$ = 1), class ($n$ = 3), order ($n$ = 5), family ($n$ = 9), and genus ($n$ = 11) levels with a threshold of LDA > 2 (**Figure 3A**). The Wilcoxon rank-sum test was also used to explore changes in microbiota, and 30 differentially enriched taxa were identified (**Figures 3B–D**, **Supplementary Figure 2**, and **Supplementary Table 2**). The differential microbiota at the genus level from the Wilcoxon test were the same as those that we had screened using LEfSe analysis (**Table 2**). *Acidaminococcus*, *Christensenella*, and *Lactobacillus* were enriched in CAS patients; *Anaerostipes*, *Fusobacterium*, *Gemella*, *Parvimonas*, *Romboutsia*, and *Clostridium* XVIII/XlVa/XlVb were enriched in healthy controls. The correlation between different genera was shown by Spearman's correlation test (**Figure 3E**), and these microbiota genera were further utilized in correlation analysis with differential metabolites and DEGs.

**TABLE 2 |** Differentially enriched gut microbiota from genera level.

| Gut microbiota | Mean (AS) | Mean (Con) | *p*-value | Median (AS) | Median (Con) |
|---|---|---|---|---|---|
| g__Acidaminococcus | 0.000880977 | 0.00010395 | 0.004853071 | −10.30582176 | −9.853309555 |
| g__Anaerostipes | 0.001823025 | 0.00227001 | 0.033041905 | −11.23182118 | −10.38382427 |
| g__Christensenella | 4.28794E-05 | 3.89813E-06 | 0.027190901 | −13.55374927 | −12.96878677 |
| g__Clostridium XVIII | 0.000165021 | 0.000632796 | 0.008816879 | −12.96878677 | −12.23182118 |
| g__Clostridium XlVa | 0.014102131 | 0.024426975 | 0.011581483 | −6.754083402 | −5.596422402 |
| g__Clostridium XlVb | 0.001576143 | 0.00431263 | 0.000456455 | −9.352776212 | −8.268347055 |
| g__Fusobacterium | 0.000284563 | 0.002597453 | 0.043528042 | −13.55374927 | −12.39278523 |
| g__Gemella | 3.89813E-06 | 2.079E-05 | 0.014781178 | −14.55374927 | −14.55374927 |
| g__Lactobacillus | 0.004282744 | 0.000174116 | 0.001399704 | −10.74639435 | −11.96878677 |
| g__Parvimonas | 1.29938E-06 | 1.68919E-05 | 0.012301771 | −14.55374927 | −14.55374927 |
| g__Romboutsia | 0.000267672 | 0.001351351 | 0.000467579 | −11.74639435 | −10.55657256 |

*Acidaminococcus, Christensenella, and Lactobacillus were enriched in CAS patients, whereas other genera were enriched in healthy controls.*

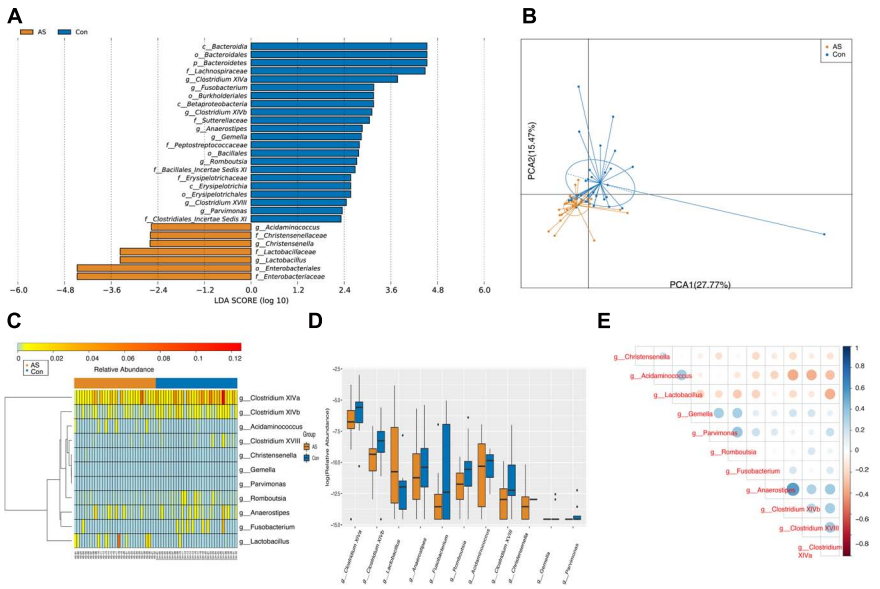



**FIGURE 3 |** Differentially enriched gut microbiota for CAS patients and healthy controls. **(A)** A total of 29 different features screened by LEfSe at phylum ($n$ = 1), class ($n$ = 3), order ($n$ = 5), family ($n$ = 9), and genus ($n$ = 11) level with a threshold of LDA > 2. **(B)** PCA plot demonstrated that CAS group is significantly different from control group based on differential genera. **(C,D)** Differentially enriched gut microbiotas were visualized in heatmap and box plot. **(E)** Associations between differential microbiotas were by correlation heatmap. Bluer color indicates a more positive correlation, whereas redder color indicates a more negative correlation.

## Different Functional Composition Profiles of Gut Microbiota Between Carotid Atherosclerosis Patients and Healthy Controls

Through PICRUSt, functional composition profiles of gut microbiota were predicted based on relative abundance and compared between CAS patients and healthy controls. In total, 65 of 265 differentially enriched KEGG pathways (level 3) were identified (**Supplementary Table 3**) with a threshold of $p < 0.05$, of which 39 were enriched in the CAS group, whereas 26 were enriched in healthy controls. Different pathways with the highest relative abundance were visualized by heatmap and boxplot (**Figure 4**).

## Metabolic Profiling of Carotid Atherosclerosis Patients and Controls

In total, 1,425 and 1,580 peaks were detected for the POS and NEG modes of UHPLC-QTOFMS, respectively, after filtering internal standards and pseudo-positive peaks.

## Differential Metabolites Screening

Two multivariate statistical analysis methods, principal component analysis (PCA) and orthogonal projections to latent structures-discriminant analysis, were utilized to classify plasma samples. Both methods showed that plasma samples for CAS patients and controls were clearly separated (**Figures 5A,B** and **Supplementary Figures 3A,B**). In addition, the permutation test indicated that the orthogonal projections to latent structures-discriminant analysis model is valid and that no overfitting exists (**Supplementary Figures 3C,D**).

With the thresholds of VIP > 1 and $p < 0.05$, 165 and 96 significantly changed metabolites were screened in POS and NEG modes, respectively (**Supplementary Figure 3E** and **Supplementary Table 4**). The patterns of differential metabolism were visualized by heatmaps and volcano plots (**Figures 5C,D** and **Supplementary Figures 3F,G**). Next, we added $|\log_2(\text{fold-change})| > 1$ as another threshold and combined POS and NEG modes to select differential metabolites (**Table 3**) for correlation analysis with different omics data. With the addition of this threshold, 11 and 12 metabolites were screened in POS and NEG modes, respectively. PAGln, upregulated in CAS patients, was the only metabolite detected in both modes and had the highest $|\log_2(\text{fold-change})|$ in POS mode.

## Metabolic Pathway Analysis for Differential Plasma Metabolites

Differential metabolites were subjected to the KEGG database to analyze the pathways in which these metabolites were involved. The bubble plot and tree plot demonstrated the $p$-value and topological impact of each enriched pathway (**Figure 6**, **Supplementary Figure 4**, and **Supplementary Table 5**).

## Adjustment for Covariates in Differential Genera and Metabolites

Most of the identified features still remain after performing adjustments for the covariates, including age and sex. First, in the differential genera after adjustments using the generalized linear model. *Anaerostipes*, *Clostridium*_XVIII, *Gemella*, and *Lactobacillus* were found to be significantly associated with sex. *Clostridium*_XlVa and *Parvimonas* were found to be significantly

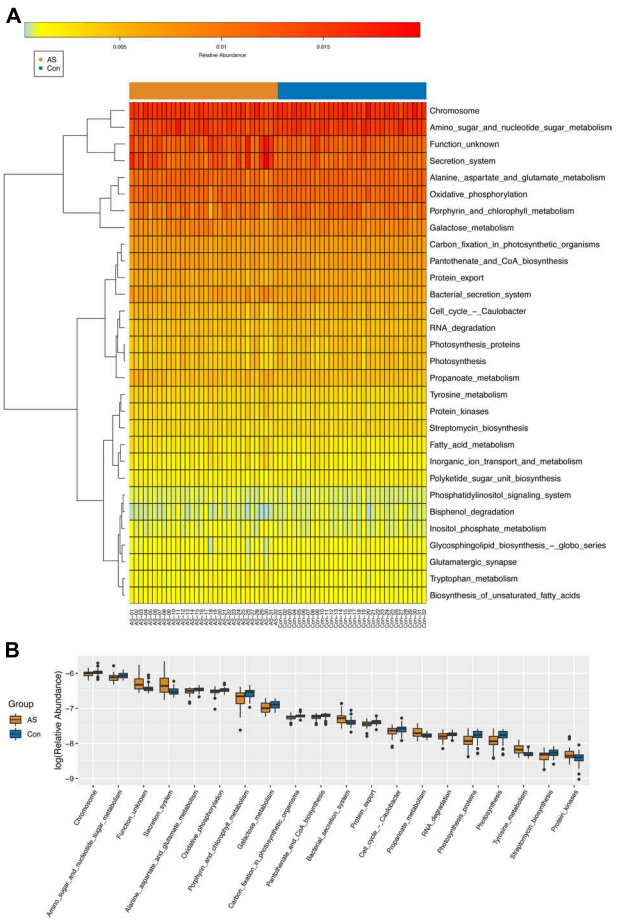



**FIGURE 4 |** KEGG pathways (level 3) predicted by PICRUSt. **(A,B)** Differential KEGG pathways with top relative abundance were visualized as heatmap and box plot. Warmer color indicates higher relative abundance, whereas colder color indicates lower relative abundance.

associated with age and sex. In the differential metabolites after adjustment using PERMANOVA, no metabolites were found to be associated with covariates ($p > 0.05$). The details of the adjustments for differential genera and metabolites could be, respectively, seen in **Supplementary Tables 6**, **7**.

## Transcriptomic Profiling of Carotid Atherosclerosis Patients and Controls

### Differentially Expressed Gene Screening

There were 32 CAS patients in the GEO datasets we selected (GSE43292). The gene expression profiles of carotid atheroma and paired macroscopically intact tissue adjacent to the atheroma plaque of each patient were shown by mRNA microarray. To reduce the effect of confounding factors, we performed a paired DEG analysis, and a total of 132 DEGs were screened, of which 76 were upregulated and 56 were downregulated with the thresholds of $|\log_2(\text{fold-change})| > 1$ and adjusted $p < 0.01$ (**Figures 7A,B**). DEGs with the top-20 $|\log_2(\text{fold-change})|$ were selected for correlation analysis (**Table 4**).

### Functional Annotation Analysis for Differentially Expressed Genes

To obtain the biological functions of the DEGs, GO-BP and KEGG pathway analyses were performed. Count number > 2 and $p$-value < 0.05 were selected as the thresholds for significantly enriched GO-BP terms and KEGG pathways. We have also performed enrichment analysis on the DEGs using the Reactome database. These DEGs were mainly associated with inflammatory and immune responses, as both KEGG and Reactome pathways enrichment analyses have shown (**Figures 7C,D** and **Supplementary Tables 8**, **9**).

## Integration of Multi-Omics Data

Spearman's correlation test was conducted between differentially enriched genera, differential metabolites, and DEGs to investigate the associations among multi-omics results (**Supplementary Table 10**). The results were visualized as correlation heatmaps (**Figures 7E–G**). The correlation analysis was also adjusted by using FDR (**Supplementary Figure 5** and **Supplementary Table 10**).

Finally, to show the potential diagnostic value of multi-omics data to discriminate CAS patients and healthy controls, we performed ROC and RF analyses for differentially enriched genera, differential metabolites, and DEGs (**Figure 8** and **Supplementary Table 11**).

# DISCUSSION

In this multi-omics study, gut microbiota and metabolite data were obtained from samples of CAS patients and healthy controls from PUMCH, and mRNA microarray data were obtained from GSE43292, which includes 32 atheromas and 32 paired control samples. The microbiome study showed significantly different β-diversity between CAS patients and healthy controls, although the α-diversity between the two groups was not significantly different, suggesting a significant difference in microbial composition, although there were similarities in microbial richness. At the genera level, 11 differentially enriched microbiota were identified (**Table 2**). In these differentially enriched microbiota, *Acidaminococcus*, *Christensenella*, and *Lactobacillus* genera were enriched in CAS patients. The metabolome analysis screened 165 and 96 differentially expressed metabolites in the POS and NEG modes, respectively. Next, 22 differential metabolites were further selected for correlation analysis by adding $|\log2(\text{fold-change})| > 1$ as an additional threshold (**Table 3**). In transcriptomic analysis, 76 upregulated and 56 downregulated genes were screened, and DEGs with top-20 $|\log_2(\text{fold-change})|$ were included for correlation analysis (**Table 4**). Spearman's correlation indices showed the association among different omics results.

In the genus-level analysis of differential gut microbiota, *Acidaminococcus*, *Christensenella*, and *Lactobacillus* were more abundant in the CAS group, whereas *Anaerostipes*, *Clostridium* XVIII/XlVa/XlVb, *Fusobacterium*, *Gemella*, *Parvimonas*, and *Romboutsia* were enriched in the healthy controls. *Acidaminococcus* is known to be a normal commensal

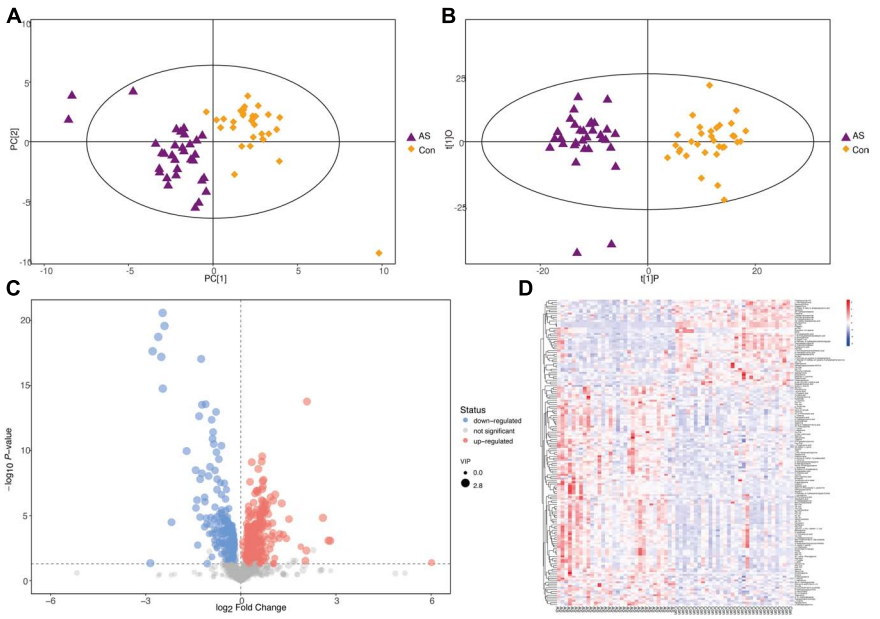



**FIGURE 5 |** Differential metabolites screening for CAS patients and healthy controls (POS mode). **(A,B)** PCA and orthogonal projections to latent structures-discriminant analysis showed that CAS patients and healthy controls could be clearly separated. **(C,D)** Patterns of differential metabolites in POS mode were demonstrated by volcano plot and heatmap.

**TABLE 3 |** Differential metabolites for correlation analysis.

| Metabolites | VIP | *p*-value | *q*-value | $Log_2$ (fold-change) |
|---|---|---|---|---|
| **POS mode** | | | | – |
| Ethanolamine | 2.816546298 | 6.3964E-18 | 8.96521E-16 | −2.509502083 |
| Gly-Pro | 2.796795801 | 1.88336E-19 | 4.55375E-17 | −2.603752686 |
| Propoxur | 2.785113538 | 2.69771E-20 | 9.78414E-18 | −2.406871344 |
| Homocitrate | 1.978541617 | 0.000652259 | 0.001823948 | 1.047764919 |
| α-N-Phenylacetyl-L-glutamine | 1.762051456 | 0.000569636 | 0.001635096 | 1.436216381 |
| Diethylcarbamazine | 1.642884521 | 0.000557875 | 0.001607395 | 1.208993076 |
| Dimethylbenzimidazole | 2.347024371 | 7.03818E-07 | 9.59896E-06 | 1.016050071 |
| Eicosapentaenoic acid | 1.807845477 | 4.49425E-05 | 0.00025414 | −1.07267148 |
| Decanoyl-L-carnitine | 1.911522868 | 1.30652E-05 | 9.80438E-05 | −1.289227648 |
| 3-Methoxy-4-Hydroxyphenylglycol Sulfate | 1.673158751 | 0.000206213 | 0.000749297 | 1.007090541 |
| O-Desmethylnaproxen | 2.140136711 | 9.1044E-09 | 2.26664E-07 | −1.017466992 |
| **NEG mode** | | | | |
| Salicylic acid | 1.734352898 | 0.01921337 | 0.0277024 | 4.145007415 |
| 3-Aminopropanesulphonic Acid | 1.039813313 | 0.040379565 | 0.047806112 | 1.796022962 |
| 6-Hydroxynicotinic acid | 1.206818152 | 0.029194166 | 0.037505435 | 2.079415637 |
| Formylanthranilic acid | 1.949132714 | 0.000220127 | 0.000651909 | 1.426589827 |
| Xanthopterin | 2.153639262 | 5.0361E-11 | 1.65249E-09 | −1.016011938 |
| N1-Methyl-4-pyridone-3-carboxamide | 2.443377406 | 8.40619E-14 | 3.7485E-12 | −1.180267083 |
| 3-Hydroxydodecanoic acid | 1.872339097 | 5.84638E-05 | 0.000209662 | −1.103961553 |
| Salicyluric acid | 1.428929548 | 0.027925205 | 0.036246094 | 1.744879658 |
| Phenylacetylglycine | 1.609315395 | 1.20649E-07 | 1.48763E-06 | 1.576664728 |
| D-Biotin | 2.385066256 | 4.85697E-11 | 1.60101E-09 | −2.067524087 |
| α-N-Phenylacetyl-L-glutamine | 1.701961669 | 0.000390232 | 0.001051251 | 1.214340063 |
| 5,10-methylene-THF | 1.922624936 | 8.93527E-07 | 8.64096E-06 | −1.040847640 |

*A threshold of | $log_2$(fold-change)| > 1 was added based on basic threshold of VIP > 1 and p-value < 0.05 to further screen metabolites for correlation analysis.*

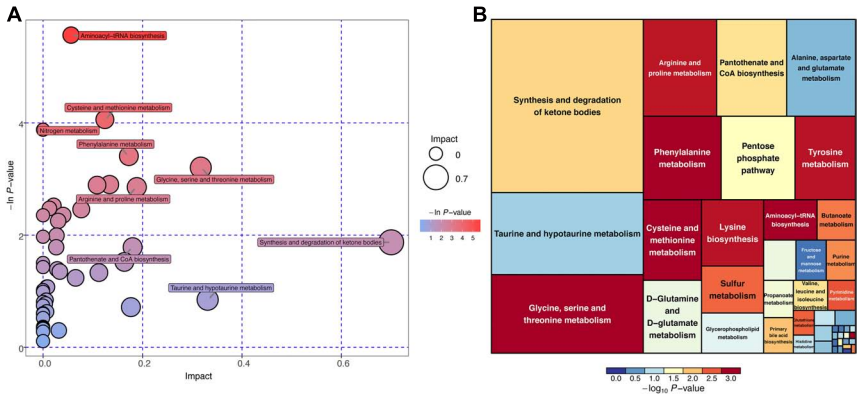



**FIGURE 6 |** KEGG pathway enrichment analysis for differential metabolites (POS mode). **(A)** Horizontal axis and size of bubble showed topological impact of pathways. Vertical axis and color of bubble showed *p*-value of pathways. **(B)** Size of square showed topological impact of pathways, whereas color of square showed *p*-value of pathways.

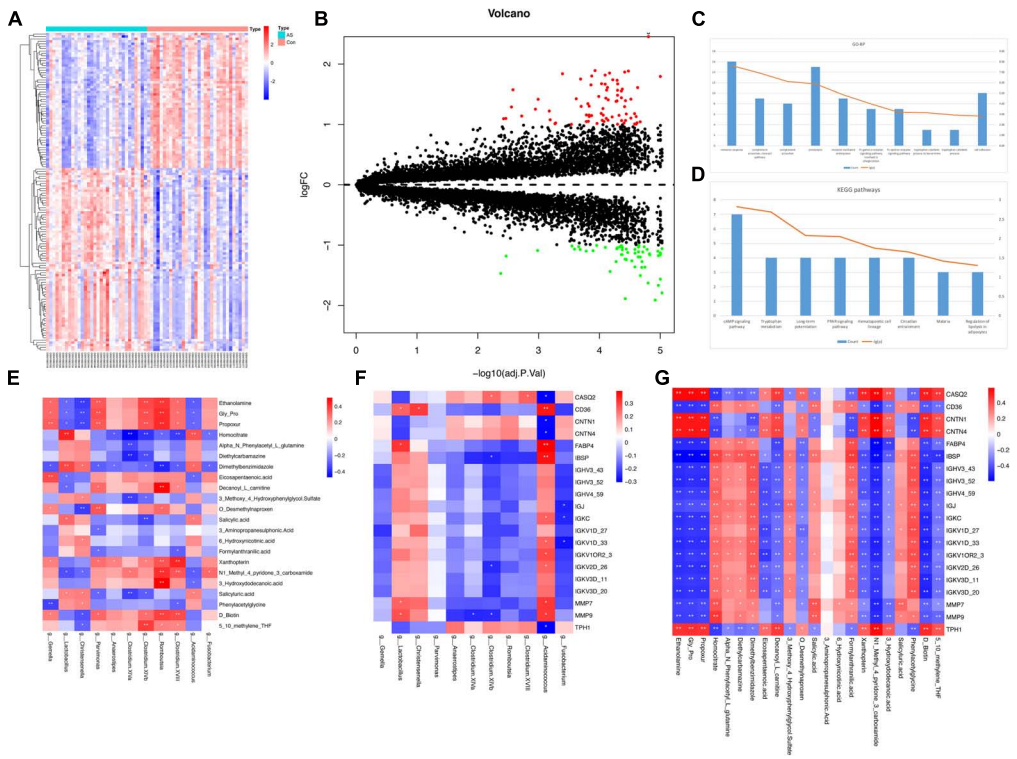



**FIGURE 7 |** DEG analysis and correlation between different types of omics data. **(A,B)** Expression patterns of DEGs were shown by heatmap and volcano plot. **(C,D)** GO-BP terms and KEGG pathways with top-10 count number were visualized. Functional enrichment analysis for DEGs showed that these DEGs were mainly associated with inflammatory and immune response. **(E–G)** Pairwise correlation between microbiome, metabolome, and transcriptome data. Reder color indicates a stronger correlation, whereas bluer color indicates a weaker correlation.

of the human gut and has been occasionally related to infective processes but always associated with polymicrobial infections (D'Auria et al., 2011). *Acidaminococcus* was reported to be enriched in the stool of patients with several inflammatory diseases, such as rheumatoid arthritis, ankylosing spondylitis, and ulcerative colitis (Altomare et al., 2019; Lee J. Y. et al., 2019; Zhou et al., 2020). Moreover, according to a recent study by Zheng et al. (2020), the abundance of *Acidaminococcus* is positively correlated with a pro-inflammatory diet, indicating that *Acidaminococcus* may be a pro-inflammatory microbiota and represent inflammatory status in the development of AS. *Christensenella* is a gram-negative, strictly anaerobic short rod associated with weight loss (Morotomi et al., 2012). Several studies have indicated that *Christensenella* was enriched in

**TABLE 4 |** DEGs with top-20 | $log_2$(fold-change)|.

| Gene symbol | Gene name | $Log_2$(fold-change) | Adjusted *p*-value |
|---|---|---|---|
| FABP4 | Fatty acid binding protein 4 | 2.454461 | 1.56E-05 |
| CNTN1 | Contactin 1 | −1.911032 | 1.21E-05 |
| IGJ | Joining chain of multimeric IgA and IgM | 1.893149 | 0.000118 |
| TPH1 | Tryptophan hydroxylase 1 | −1.886626 | 3.76E-05 |
| IGKV1D-33 | Immunoglobulin κ variable 1D-33 | 1.884738 | 3.34E-05 |
| IGHV3-52 | Immunoglobulin heavy variable 3-52 | 1.8676 | 3.85E-05 |
| IGKV3D-11 | Immunoglobulin κ variable 3D-11 | 1.851044 | 6.82E-05 |
| MMP7 | Matrix metallopeptidase 7 | 1.840231 | 0.000405 |
| MMP9 | Matrix metallopeptidase 9 | 1.817804 | 8.07E-05 |
| CD36 | CD36 molecule | 1.802205 | 0.000134 |
| IBSP | Integrin binding sialoprotein | 1.794982 | 9.96E-06 |
| CNTN4 | Contactin 4 | −1.792332 | 9.36E-06 |
| IGKV1D-27 | Immunoglobulin κ variable 1D-27 | 1.751979 | 0.000204 |
| IGHV4-59 | Immunoglobulin heavy variable 4-59 | 1.739775 | 6.03E-05 |
| IGHV3-43 | Immunoglobulin heavy variable 3-43 | 1.73119 | 5.12E-05 |
| IGKV1OR2-3 | Immunoglobulin κ variable 1/OR2-3 | 1.674665 | 9.21E-05 |
| IGKC | Immunoglobulin κ constant | 1.672012 | 5.01E-05 |
| CASQ2 | Calsequestrin 2 | −1.667664 | 1.07E-05 |
| IGKV3D-20 | Immunoglobulin κ variable 3D-20 | 1.667174 | 8.07E-05 |
| IGKV2D-26 | Immunoglobulin κ variable 2D-26 | 1.654989 | 0.000138 |

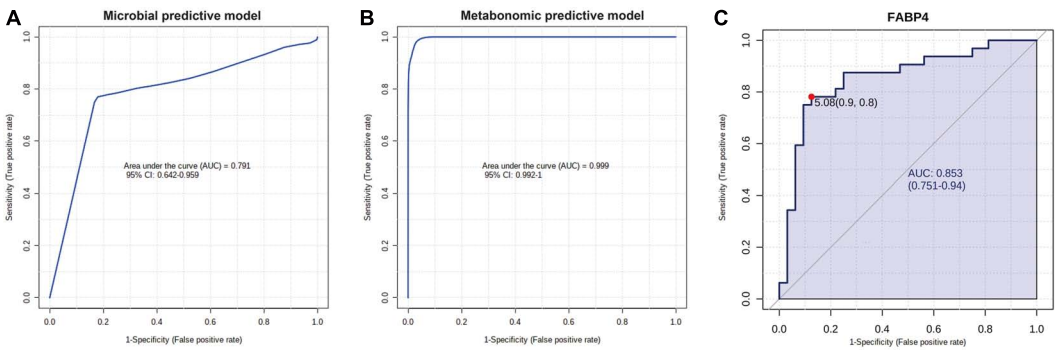



**FIGURE 8 |** Potential diagnostic value of differential gut microbiota, metabolites, and DEG with importance. **(A)** Microbial predictive model. Gut microbiota significantly enriched in CAS group were included, and RF algorithm was used to construct this model. **(B)** Metabolomic predictive model. Phenylacetylglutamine, phenylacetylglycine, ethanolamine, and eicosapentaenoic acid were included, and RF algorithm was used to construct this model. **(C)** FABP4 is the DEG with the highest $log_2$(fold-change), and potential diagnostic value of this gene was shown by ROC analysis.

type 1 diabetes patients with the decreased abundance of the SCFA-producing microbiota, Roseburia. In addition, whole-genome sequencing indicated that some genes of *Christensenella* were related to lipopolysaccharide biosynthesis, and the lipopolysaccharide from *Christensenella* can trigger a weak inflammatory response through the nuclear factor kappa-B (NF-κB) signaling pathway (Yang et al., 2018). Although *Lactobacillus* is often described as an anti-inflammatory probiotic in many studies of AS, the role of *Lactobacillus* in the pathogenesis of AS remains controversial (Ding et al., 2017). Several *Lactobacillus* species significantly reduce the inflammatory response *via* T regulatory cells and alleviate arteriosclerotic level, but some other species of *Lactobacillus* could promote the inflammatory response, which may aggravate AS (Smits et al., 2005; Bhathena et al., 2009; Karimi et al., 2009; Pan et al., 2011; Won et al., 2011; Shah et al., 2012; Dimitrijevic et al., 2014). The increased abundance of *Lactobacillus* in the CAS group may fall into different species; therefore, additional research is needed to address this question.

For the genera enriched in healthy controls, *Anaerostipes*, *Gemella*, and *Parvimonas* were reported to be scarce and primarily enriched in the healthy gut (Bodkhe et al., 2019; Hong et al., 2019; Magruder et al., 2020). *Clostridium* XVIII/XlVa/XlVb, *Fusobacterium*, and *Romboutsia* are all major SCFAs, particularly butyrate producers in the process of human metabolism (Duncan et al., 2002; Bui et al., 2014; Neijat et al., 2019). In humans, SCFAs are produced from dietary fibers and resistant starches that cannot be decomposed by digestive enzymes through fermentation by the microbiota in the cecum and colon (Cummings et al., 1987). SCFAs may suppress inflammation by

reducing migration and proliferation of immune cells, thereby reducing many types of cytokines and inducing apoptosis (Ohira et al., 2017). Furthermore, data from animal experiments found that compared with the sterile mice, the atherosclerotic plaque of mice carrying *Roseburia* was significantly reduced after they were fed with a high fiber diet. The mechanism was that SCFAs could inhibit the activation of histone deacetylase, NF-κB, and tumor necrosis factor-α signaling pathways, reduce the expression of vascular cell adhesion molecule-1, and protect endothelial function. On the other hand, SCFAs could promote the conversion of cholesterol to bile acid, thereby alleviating AS (Kasahara et al., 2018).

Based on the metabolomic analysis, the upregulated PAGln was detected in both POS and NEG modes and had the highest $|\log_2(\text{fold-change})|$ in POS mode. PAGln is a phenylalanine-derived metabolite formed from the conjugation of glutamine and phenylacetate (Aronov et al., 2011). Phenylalanine is one of the essential amino acids for human metabolism. After phenylalanine is ingested by the human body, most of this amino acid is absorbed by the small intestine. Excessive phenylalanine reaches the colon and can be metabolized into phenylpyruvic acid further into phenylacetic acid by gut microbiota. Next, glutamine and this microbial-derived phenylacetic acid are conjugated in the human liver and kidney, and PAGln is produced (Li et al., 2008; Witkowski et al., 2020). Increased level of plasma PAGln was shown to be associated with increased major adverse cardiac events (myocardial infarction, stroke, or death) by untargeted metabolomics using a large cohort ($n$ = 1,162) and a validation cohort ($n$ = 4,000) (Nemet et al., 2020). Bogiatzi et al. (2018) also discovered that PAGln was elevated in AS patients. Our results were consistent with the findings of these previous studies. In addition, the KEGG pathway analysis in our study found that differential metabolites were significantly enriched in phenylalanine metabolic pathway for both POS and NEG modes, suggesting that phenylalanine metabolism and subsequently generated PAGln play vital roles in CAS pathogenesis. A recent mechanistic study has shown that PAGln increases thrombosis potential by activating platelet functions through multiple approaches such as interacting with α2A, α2B, and β2 adrenergic receptors (Nemet et al., 2020). Another upregulated metabolite in the NEG mode, phenylacetylglycine (PAGly), has a similar function as PAGln and can enhance platelet function *via* adrenergic receptors. However, compared with PAGln, PAGly was a major product in mice found in the study of Nemet et al. (2020). Our study and previous findings indicated that PAGln and phenylalanine metabolism are crucial mediators in CAS pathogenesis and might serve as promising pharmacotherapeutic targets to slow CAS progression.

Ethanolamine was the differential metabolite with the highest VIP value in POS mode and downregulated in CAS patients. The level of ethanolamine in HDL is positively correlated with cholesterol efflux capacity and negatively associated with plaque scores in chronic kidney disease (CKD) patients (Maeba et al., 2018). The finding of downregulated ethanolamine in CAS patients in our study was consistent with this previous study and suggested that downregulation of ethanolamine might promote AS progression. In contrast to this previous study, the patients in our study were not CKD patients and might be more representative. Another downregulated metabolite, eicosapentaenoic acid (EPA), is an omega-3 fatty acid found in fish oil. EPA and its derivatives were found to have protective roles against cardiovascular disease in clinical trials (Leaf et al., 1994; Sacks et al., 1995; Bhatt et al., 2020). EPA can be enzymatically converted to resolvin E1 (RvE1) *in vivo* and affect atherosclerotic inflammation and mediate the immune response through the EPA/RvE1/ChemR23 pathway, thereby improving the outcomes of atherosclerosis-related cardiovascular disease (Carracedo et al., 2019). Our results indicated a deficiency of these beneficial metabolites in CAS patients, and supplementation with fish oil might benefit these patients.

In transcriptomic analysis, we observed that DEGs were mainly associated with inflammatory and immune response through GO-BP and KEGG pathway enrichment analysis. Our findings at the transcriptome level agree with the consensus that atherosclerosis is characterized by low-grade, chronic inflammation of the arteries and infiltration of immune cells such as macrophages, mast cells, and T lymphocytes (Hansson, 2005; Galkina and Ley, 2009; Bäck et al., 2019). Furthermore, *FABP4*, fatty acid-binding protein 4, is the upregulated DEG in CAS patients with the highest $|\log_2(\text{fold-change})|$. FABP4 is mainly expressed in adipocytes and macrophages. This protein can serve as an adipokine for the development of atherosclerosis and insulin resistance (Hotamisligil and Bernlohr, 2015). In macrophages, FABP4 can induce an inflammatory response through such pathways as NF-κB and JKN/AP-1 (Furuhashi, 2019).

In the correlation analysis between gut microbiota and plasma metabolites, PAGln was negatively associated with *Clostridium* XIVa, which belongs to the Lachnospiraceae family. In early renal function decline patients, Barrios et al. (2015) found that PAGln was negatively correlated with several genera in the Lachnospiraceae family, and in patients with coronary artery disease, Ottosson et al. (2020) identified one unknown genus in the Lachnospiraceae family that was also negatively correlated with PAGln. The results of our study were consistent with the findings of previous studies and further identified a new genus in this family that was negatively correlated with PAGln in CAS patients. Although few studies on this topic have been conducted, the association between Lachnospiraceae and PAGln might be one of the microbiota–metabolite axes mediating AS pathogenesis. EPA, the metabolite with anti-inflammatory roles in AS patients, was negatively associated with *Acidaminococcus*, a potentially pro-inflammatory microbiota genus. Because a Mediterranean diet, which mainly includes foods rich in unsaturated fatty acid (such as EPA), can have an anti-inflammatory effect (Zheng et al., 2020), we could infer that EPA might reduce atherosclerotic inflammation by targeting *Acidaminococcus*.

In the correlation analysis between transcriptomic profiles and the other two omics datasets, we also obtained some findings that might deepen the current understanding of AS pathogenesis. We found that *Acidaminococcus* was positively associated with *FABP4* and had the highest Spearman's correlation coefficient and the most significant $p$-value among

all microbiota–DEG pairs ($\rho = 0.39$, $p = 0.0014$). Although the pro-inflammatory roles of *Acidaminococcus* and *FABP4* have been widely studied (Hotamisligil and Bernlohr, 2015; Altomare et al., 2019; Butera et al., 2020), our study was the first to identify the association between them and might provide a new perspective to explore CAS pathogenesis. Furthermore, in the correlation analysis for DEGs and metabolites, *FABP4* was also positively associated with pro-atherosclerotic metabolites (PAGln and PAGly) and negatively associated with anti-atherosclerotic metabolite (ethanolamine), which implied that this adipokine was not only associated with crucial gut microbiota but also might interact with crucial metabolites in CAS pathogenesis.

In this study, we performed microbial and metabolomic analyses using fecal and plasma samples from CAS patients in PUMCH. Transcriptomic analysis was conducted based on one GEO dataset (GSE43292) containing 32 CAS carotid atheromas and paired controls. Differential gut microbiota, metabolites, DEGs, and related pathways were identified. Finally, the associations among various omics data were investigated by correlation analysis. However, our study has some limitations. First, the body mass index was marginally higher ($24.7 \pm 2.7$ for CAS patients and $23.2 \pm 2.2$ for healthy controls, $p = 0.047$) in CAS patients; the risk factor role of obesity might account for this difference (Rocha and Libby, 2009). Second, patients were only recruited from PUMCH, and the sample size was small. Future multicentric studies with large samples are needed to generalize these findings. Third, transcriptomic data obtained from the GEO database were not obtained from the same patients as the microbiome and metabolome data. This difference would result in batch effects, which need to be verified by the same cohorts. Furthermore, *in vitro* and *in vivo* experiments are warranted to elucidate further the mechanisms governing how gut microbiota, plasma metabolites, and DEGs interact with one another.

## CONCLUSION

Despite extensive researches investigating AS, in the past decade, relatively little is known regarding the mechanisms underlying the pathogenesis of CAS. Accumulating evidence has shown that the gut microbiota serve as a pivotal risk factor in cardiovascular diseases by influencing host metabolism and immune homeostasis (Battson et al., 2018). However, no direct evidence has established a direct and causal relationship between altered gut microbiota and CAS. Through an integrated analysis of multi-omics, we explored the possible "microbiota–metabolite–gene" regulatory axis that may act on CAS, thereby helping to establish a theoretical basis for the further specialized study of CAS.

## DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO), https://www.ncbi.nlm.nih.gov/geo/, GSE28829, GSE104140, and GSE43292. The study contains 16S rDNA sequencing data and we have uploaded the original fastq data in the Sequence Read Archive (SRA) database (BioProject: PRJNA674452, https://www.ncbi.nlm.nih.gov/sra/PRJNA674452).

## ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Ethics Committee of the Peking Union Medical College Hospital. The patients/participants provided their written informed consent to participate in this study.

## AUTHOR CONTRIBUTIONS

LJ and SC conceived and designed the experiments, performed the experiments, analyzed the data, contributed reagents, materials, and analysis tools, prepared figures and/or tables, authored or reviewed drafts of the manuscript, and approved the final draft. GG conceived and designed the experiments, performed the experiments, analyzed the data, authored or reviewed drafts of the manuscript, and approved the final draft. JZ, JR, and JW analyzed the data, prepared figures and/or tables, and approved the final draft. WW and DY analyzed the data, authored or reviewed drafts of the manuscript, and approved the final draft. YZ conceived and designed the experiments, authored or reviewed drafts of the manuscript, uploaded the original fastq data of 16S rDNA sequencing in the Sequence Read Archive (SRA) database and approved the final draft. All authors contributed to the article and approved the submitted version.

## FUNDING

This research was supported in part by the Major Research Program of National Natural Science Foundation of China (51890894), National Natural Science Foundation of China (81770481 and 82070492), and the Chinese Academy of Medical Sciences, innovation Fund for Medical Sciences (CIFMS 2017-I2M-1-008).

## ACKNOWLEDGMENTS

We thank the assistance from Shanghai BIOTREE Biological Technology Co., Ltd. (Shanghai, China) for data analysis of Metabolomics.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fphys.2021.645212/full#supplementary-material

**Supplementary Figure 1 |** Quality control for 16S rDNA sequencing and microbial diversity of CAS and healthy control samples from different level. **(A,B)** Rarefaction curves of richness (observed_species and chao 1) suggested the sequencing depth is adequate. **(C)** The Venn diagram showed the overlapping OTUs in microbiota among CAS patients and healthy controls. **(D–G)** The relative abundance of gut microbiota for two groups from phylum, class, order, and family levels, respectively.

**Supplementary Figure 2 |** Differentially enriched microbiota for CAS patients and healthy controls from all levels. **(A)** The differentially enriched gut microbiotas from all levels were visualized in heatmap. **(B)** PCA plot demonstrated that CAS group is significantly different from control group based on differential microbiotas from all levels. **(C)** The differentially enriched gut microbiotas were visualized in box plot.

**Supplementary Figure 3 |** Differential metabolites screening for CAS patients and healthy controls (NEG mode) and permutation test for OPLS-DA model. **(A,B)** The PCA and OPLS-DA showed that CAS patients and healthy controls can be clearly separated. **(C,D)** The permutation test showed that the OPLS-DA model is valid and no overfitting exist for both POS and NEG mode. **(E)** The Venn diagram showed 17 overlapping different metabolites between POS and NEG mode. **(F,G)** Patterns of differential metabolites in NEG mode were demonstrated by the volcano plot and heatmap.

**Supplementary Figure 4 |** KEGG pathway enrichment analysis for differential metabolites (NEG mode). **(A)** The horizontal axis and sized of the bubble showed the topological impact of pathways. The vertical axis and color of bubble showed the *p*-value of pathways. **(B)** The size of the square showed the topological impact of pathways while the color of the square showed the *p*-value of the pathways.

**Supplementary Figure 5 |** FDR adjustment for correlation between different types of omics data. **(A–C)** The FDR adjustment for correlation between microbiome, metabolome, and transcriptome data. Redder color indicates stronger correlation while bluer color indicates weaker correlation.

**Supplementary Table 1 |** The α-diversity for CAS and healthy controls.

**Supplementary Table 2 |** Differentially enriched microbiota from all levels.

**Supplementary Table 3 |** Differential KEGG pathways predicted by PICRUSt.

**Supplementary Table 4 |** All differential metabolites between CAS patients and healthy controls.

**Supplementary Table 5 |** KEGG pathways for differential metabolites.

**Supplementary Table 6 |** Covariates adjustment for differentially enriched genera between CAS patients and healthy controls using GLM analysis.

**Supplementary Table 7 |** Covariates adjustment for differential metabolites between CAS patients and healthy controls using PERMANOVA analysis.

**Supplementary Table 8 |** Functional enrichment analysis for DEGs.

**Supplementary Table 9 |** Enrichment analysis on the DEGs using the Reactome database.

**Supplementary Table 10 |** Details for correlation analysis.

**Supplementary Table 11 |** AUCs for differentially enriched gut microbiota, metabolites, and DEGs.

# REFERENCES


Aboyans, V., Ricco, J. B., Bartelink, M. E. L., Bjorck, M., Brodmann, M., Cohnert, T., et al. (2018). 2017 ESC Guidelines on the Diagnosis and Treatment of Peripheral Arterial Diseases, in collaboration with the European Society for Vascular Surgery (ESVS): Document covering atherosclerotic disease of extracranial carotid and vertebral, mesenteric, renal, upper and lower extremity arteriesEndorsed by: the European Stroke Organization (ESO)The Task Force for the Diagnosis and Treatment of Peripheral Arterial Diseases of the European Society of Cardiology (ESC) and of the European Society for Vascular Surgery (ESVS). *Eur Heart J* 39, 763–816. doi: 10.1093/eurheartj/ehx095

Altomare, A., Putignani, L., Del Chierico, F., Cocca, S., Angeletti, S., Ciccozzi, M., et al. (2019). Gut mucosal-associated microbiota better discloses inflammatory bowel disease differential patterns than faecal microbiota. *Dig Liver Dis* 51, 648–656. doi: 10.1016/j.dld.2018.11.021

Aronov, P. A., Luo, F. J., Plummer, N. S., Quan, Z., Holmes, S., Hostetter, T. H., et al. (2011). Colonic contribution to uremic solutes. *J Am Soc Nephrol* 22, 1769–1776. doi: 10.1681/asn.2010121220

Artom, N., Montecucco, F., Dallegri, F., and Pende, A. (2014). Carotid atherosclerotic plaque stenosis: the stabilizing role of statins. *Eur J Clin Invest* 44, 1122–1134. doi: 10.1111/eci.12340

Bäck, M., Yurdagul, A. Jr., Tabas, I., Öörni, K., and Kovanen, P. T. (2019). Inflammation and its resolution in atherosclerosis: mediators and therapeutic opportunities. *Nat Rev Cardiol* 16, 389–406. doi: 10.1038/s41569-019-0169-2

Barrios, C., Beaumont, M., Pallister, T., Villar, J., Goodrich, J. K., Clark, A., et al. (2015). Gut-Microbiota-Metabolite Axis in Early Renal Function Decline. *PLoS One* 10:e0134311. doi: 10.1371/journal.pone.0134311

Battson, M. L., Lee, D. M., Weir, T. L., and Gentile, C. L. (2018). The gut microbiota as a novel regulator of cardiovascular function and disease. *J Nutr Biochem* 56, 1–15. doi: 10.1016/j.jnutbio.2017.12.010

Bhathena, J., Martoni, C., Kulamarva, A., Urbanska, A. M., Malhotra, M., and Prakash, S. (2009). Orally delivered microencapsulated live probiotic formulation lowers serum lipids in hypercholesterolemic hamsters. *J Med Food* 12, 310–319. doi: 10.1089/jmf.2008.0166

Bhatt, D. L., Miller, M., Brinton, E. A., Jacobson, T. A., Steg, P. G., Ketchum, S. B., et al. (2020). REDUCE-IT USA: Results From the 3146 Patients Randomized in the United States. *Circulation* 141, 367–375. doi: 10.1161/circulationaha.119.044440

Bodkhe, R., Shetty, S. A., Dhotre, D. P., Verma, A. K., Bhatia, K., Mishra, A., et al. (2019). Comparison of Small Gut and Whole Gut Microbiota of First-Degree Relatives With Adult Celiac Disease Patients and Controls. *Front Microbiol* 10:164. doi: 10.3389/fmicb.2019.00164

Bogiatzi, C., Gloor, G., Allen-Vercoe, E., Reid, G., Wong, R. G., Urquhart, B. L., et al. (2018). Metabolic products of the intestinal microbiome and extremes of atherosclerosis. *Atherosclerosis* 273, 91–97. doi: 10.1016/j.atherosclerosis.2018.04.015

Bui, T. P. N., de Vos, W. M., and Plugge, C. M. (2014). Anaerostipes rhamnosivorans sp. nov., a human intestinal, butyrate-forming bacterium. *Int J Syst Evol Microbiol* 64(Pt 3), 787–793. doi: 10.1099/ijs.0.055061-0

Butera, A., Di Paola, M., Vitali, F., De Nitto, D., Covotta, F., Borrini, F., et al. (2020). IL-13 mRNA Tissue Content Identifies Two Subsets of Adult Ulcerative Colitis Patients With Different Clinical and Mucosa-Associated Microbiota Profiles. *J Crohns Colitis* 14, 369–380. doi: 10.1093/ecco-jcc/jjz154

Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., et al. (2010). QIIME allows analysis of high-throughput community sequencing data. *Nat Methods* 7, 335–336. doi: 10.1038/nmeth.f.303

Carracedo, M., Artiach, G., Arnardottir, H., and Bäck, M. (2019). The resolution of inflammation through omega-3 fatty acids in atherosclerosis, intimal hyperplasia, and vascular calcification. *Semin Immunopathol* 41, 757–766. doi: 10.1007/s00281-019-00767-y

Chen, Y., Xu, C., Huang, R., Song, J., Li, D., and Xia, M. (2018). Butyrate from pectin fermentation inhibits intestinal cholesterol absorption and attenuates atherosclerosis in apolipoprotein E-deficient mice. *J Nutr Biochem* 56, 175–182. doi: 10.1016/j.jnutbio.2018.02.011

Cummings, J. H., Pomare, E. W., Branch, W. J., Naylor, C. P., and Macfarlane, G. T. (1987). Short chain fatty acids in human large intestine, portal, hepatic and venous blood. *Gut* 28, 1221–1227. doi: 10.1136/gut.28.10.1221

D'Auria, G., Galan, J. C., Rodriguez-Alcayna, M., Moya, A., Baquero, F., and Latorre, A. (2011). Complete genome sequence of Acidaminococcus intestini RYC-MR95, a Gram-negative bacterium from the phylum Firmicutes. *J Bacteriol* 193, 7008–7009. doi: 10.1128/JB.06301-11

Dimitrijevic, R., Ivanovic, N., Mathiesen, G., Petrusic, V., Zivkovic, I., Djordjevic, B., et al. (2014). Effects of Lactobacillus rhamnosus LA68 on the immune system of C57BL/6 mice upon oral administration. *J Dairy Res* 81, 202–207. doi: 10.1017/S0022029914000028

Ding, Y. H., Qian, L. Y., Pang, J., Lin, J. Y., Xu, Q., Wang, L. H., et al. (2017). The regulation of immune cells by Lactobacilli: a potential therapeutic target for anti-atherosclerosis therapy. *Oncotarget* 8, 59915–59928. doi: 10.18632/oncotarget.18346

Duncan, S. H., Hold, G. L., Harmsen, H. J. M., Stewart, C. S., and Flint, H. J. (2002). Growth requirements and fermentation products of Fusobacterium prausnitzii, and a proposal to reclassify it as Faecalibacterium prausnitzii gen. nov., comb. nov. *Int J Syst Evol Microbiol* 52(Pt 6), 2141–2146. doi: 10.1099/00207713-52-6-2141

Dunn, W. B., Broadhurst, D., Begley, P., Zelena, E., Francis-McIntyre, S., Anderson, N., et al. (2011). Procedures for large-scale metabolic profiling of serum and plasma using gas chromatography and liquid chromatography coupled to mass spectrometry. *Nat Protoc* 6, 1060–1083. doi: 10.1038/nprot.2011.335

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

Edgar, R. C. (2013). UPARSE: highly accurate OTU sequences from microbial amplicon reads. *Nat Methods* 10, 996–998. doi: 10.1038/nmeth.2604

Faxon, D. P., Fuster, V., Libby, P., Beckman, J. A., Hiatt, W. R., Thompson, R. W., et al. (2004). Atherosclerotic Vascular Disease Conference: Writing Group III: pathophysiology. *Circulation* 109, 2617–2625. doi: 10.1161/01.CIR.0000128520.37674.EF

Furuhashi, M. (2019). Fatty Acid-Binding Protein 4 in Cardiovascular and Metabolic Diseases. *J Atheroscler Thromb* 26, 216–232. doi: 10.5551/jat.48710

Galkina, E., and Ley, K. (2009). Immune and inflammatory mechanisms of atherosclerosis (*). *Annu Rev Immunol* 27, 165–197. doi: 10.1146/annurev.immunol.021908.132620

Hansson, G. K. (2005). Inflammation, atherosclerosis, and coronary artery disease. *N Engl J Med* 352, 1685–1695. doi: 10.1056/NEJMra043430

Hong, B. Y., Sobue, T., Choquette, L., Dupuy, A. K., Thompson, A., Burleson, J. A., et al. (2019). Chemotherapy-induced oral mucositis is associated with detrimental bacterial dysbiosis. *Microbiome* 7, 66. doi: 10.1186/s40168-019-0679-5

Hotamisligil, G. S., and Bernlohr, D. A. (2015). Metabolic functions of FABPs–mechanisms and therapeutic implications. *Nat Rev Endocrinol* 11, 592–605. doi: 10.1038/nrendo.2015.122

Jonsson, A. L., and Backhed, F. (2017). Role of gut microbiota in atherosclerosis. *Nat Rev Cardiol* 14, 79–87. doi: 10.1038/nrcardio.2016.183

Karimi, K., Inman, M. D., Bienenstock, J., and Forsythe, P. (2009). Lactobacillus reuteri-induced regulatory T cells protect against an allergic airway response in mice. *Am J Respir Crit Care Med* 179, 186–193. doi: 10.1164/rccm.200806-951OC

Kasahara, K., Krautkramer, K. A., Org, E., Romano, K. A., Kerby, R. L., Vivas, E. I., et al. (2018). Interactions between Roseburia intestinalis and diet modulate atherogenesis in a murine model. *Nat Microbiol* 3, 1461–1471. doi: 10.1038/s41564-018-0272-x

Koren, O., Spor, A., Felin, J., Fak, F., Stombaugh, J., Tremaroli, V., et al. (2011). Human oral, gut, and plaque microbiota in patients with atherosclerosis. *Proc Natl Acad Sci U S A* 108(Suppl. 1), 4592–4598. doi: 10.1073/pnas.1011383107

Langille, M. G., Zaneveld, J., Caporaso, J. G., McDonald, D., Knights, D., Reyes, J. A., et al. (2013). Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nat Biotechnol* 31, 814–821. doi: 10.1038/nbt.2676

Leaf, A., Jorgensen, M. B., Jacobs, A. K., Cote, G., Schoenfeld, D. A., Scheer, J., et al. (1994). Do fish oils prevent restenosis after coronary angioplasty? *Circulation* 90, 2248–2257. doi: 10.1161/01.cir.90.5.2248

Lee, J. Y., Mannaa, M., Kim, Y., Kim, J., Kim, G. T., and Seo, Y. S. (2019). Comparative Analysis of Fecal Microbiota Composition Between Rheumatoid Arthritis and Osteoarthritis Patients. *Genes (Basel)* 10, doi: 10.3390/genes10100748

Lee, T. H., Cheng, M. L., Shiao, M. S., and Lin, C. N. (2019). Metabolomics study in severe extracranial carotid artery stenosis. *BMC Neurol* 19:138. doi: 10.1186/s12883-019-1371-x

Li, M., Wang, B., Zhang, M., Rantalainen, M., Wang, S., Zhou, H., et al. (2008). Symbiotic gut microbes modulate human metabolic phenotypes. *Proc Natl Acad Sci U S A* 105, 2117–2122. doi: 10.1073/pnas.0712038105

Libby, P., Buring, J. E., Badimon, L., Hansson, G. K., Deanfield, J., Bittencourt, M. S., et al. (2019). Atherosclerosis. *Nat Rev Dis Primers* 5, 56. doi: 10.1038/s41572-019-0106-z

Lindskog Jonsson, A., Hallenius, F. F., Akrami, R., Johansson, E., Wester, P., Arnerlov, C., et al. (2017). Bacterial profile in human atherosclerotic plaques. *Atherosclerosis* 263, 177–183. doi: 10.1016/j.atherosclerosis.2017.06.016

Maeba, R., Kojima, K. I., Nagura, M., Komori, A., Nishimukai, M., Okazaki, T., et al. (2018). Association of cholesterol efflux capacity with plasmalogen levels of high-density lipoprotein: A cross-sectional study in chronic kidney disease patients. *Atherosclerosis* 270, 102–109. doi: 10.1016/j.atherosclerosis.2018.01.037

Magruder, M., Edusei, E., Zhang, L., Albakry, S., Satlin, M. J., Westblade, L. F., et al. (2020). Gut commensal microbiota and decreased risk for *Enterobacteriaceae* bacteriuria and urinary tract infection. *Gut Microbes* 12, 1805281. doi: 10.1080/19490976.2020.1805281

Morotomi, M., Nagai, F., and Watanabe, Y. (2012). Description of Christensenella minuta gen. nov., sp. nov., isolated from human faeces, which forms a distinct branch in the order Clostridiales, and proposal of Christensenellaceae fam. nov. *Int J Syst Evol Microbiol* 62(Pt 1), 144–149. doi: 10.1099/ijs.0.026989-0

Neijat, M., Habtewold, J., Shirley, R. B., Welsher, A., Barton, J., Thiery, P., et al. (2019). Bacillus subtilis Strain DSM 29784 Modulates the Cecal Microbiome, Concentration of Short-Chain Fatty Acids, and Apparent Retention of Dietary Components in Shaver White Chickens during Grower, Developer, and Laying Phases. *Appl Environ Microbiol* 85, doi: 10.1128/AEM.00402-19

Nemet, I., Saha, P. P., Gupta, N., Zhu, W., Romano, K. A., Skye, S. M., et al. (2020). A Cardiovascular Disease-Linked Gut Microbial Metabolite Acts via Adrenergic Receptors. *Cell* 180, 862.e–877.e. doi: 10.1016/j.cell.2020.02.016

Ohira, H., Tsutsui, W., and Fujioka, Y. (2017). Are Short Chain Fatty Acids in Gut Microbiota Defensive Players for Inflammation and Atherosclerosis? *J Atheroscler Thromb* 24, 660–672. doi: 10.5551/jat.RV17006

Ottosson, F., Brunkwall, L., Smith, E., Orho-Melander, M., Nilsson, P. M., Fernandez, C., et al. (2020). The gut microbiota-related metabolite phenylacetylglutamine associates with increased risk of incident coronary artery disease. *J Hypertens* doi: 10.1097/hjh.0000000000002569

Pan, D. D., Zeng, X. Q., and Yan, Y. T. (2011). Characterisation of Lactobacillus fermentum SM-7 isolated from koumiss, a potential probiotic bacterium with cholesterol-lowering effects. *J Sci Food Agric* 91, 512–518. doi: 10.1002/jsfa.4214

Petty, G. W., Brown, R. D. Jr., Whisnant, J. P., Sicks, J. D., O'Fallon, W. M., and Wiebers, D. O. (2000). Ischemic stroke subtypes : a population-based study of functional outcome, survival, and recurrence. *Stroke* 31, 1062–1068. doi: 10.1161/01.str.31.5.1062

Qian, Y., Yang, X., Xu, S., Wu, C., Song, Y., Qin, N., et al. (2018). Alteration of the fecal microbiota in chinese patients with parkinson's disease. *Brain Behavior and Immunity* 70, 194–202. doi: 10.1016/j.bbi.2018.02.016

Rocha, V. Z., and Libby, P. (2009). Obesity, inflammation, and atherosclerosis. *Nat Rev Cardiol* 6, 399–409. doi: 10.1038/nrcardio.2009.55

Sacks, F. M., Stone, P. H., Gibson, C. M., Silverman, D. I., Rosner, B., and Pasternak, R. C. (1995). Controlled trial of fish oil for regression of human coronary atherosclerosis. HARP Research Group. *J Am Coll Cardiol* 25, 1492–1498. doi: 10.1016/0735-1097(95)00095-l

Segata, N., Izard, J., Waldron, L., Gevers, D., Miropolsky, L., Garrett, W. S., et al. (2011). Metagenomic biomarker discovery and explanation. *Genome Biol* 12, R60. doi: 10.1186/gb-2011-12-6-r60

Shah, M. M., Saio, M., Yamashita, H., Tanaka, H., Takami, T., Ezaki, T., et al. (2012). Lactobacillus acidophilus strain L-92 induces CD4(+)CD25(+)Foxp3(+) regulatory T cells and suppresses allergic contact dermatitis. *Biol Pharm Bull* 35, 612–616. doi: 10.1248/bpb.35.612

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

Smits, H. H., Engering, A., van der Kleij, D., de Jong, E. C., Schipper, K., van Capel, T. M., et al. (2005). Selective probiotic bacteria induce IL-10-producing regulatory T cells in vitro by modulating dendritic cell function through dendritic cell-specific intercellular adhesion molecule 3-grabbing nonintegrin. *J Allergy Clin Immunol* 115, 1260–1267. doi: 10.1016/j.jaci.2005.03.036

Song, P., Fang, Z., Wang, H., Cai, Y., Rahimi, K., Zhu, Y., et al. (2020). Global and regional prevalence, burden, and risk factors for carotid atherosclerosis: a

systematic review, meta-analysis, and modelling study. *Lancet Glob Health* 8, e721–e729. doi: 10.1016/S2214-109X(20)30117-0

Vojinovic, D., van der Lee, S. J., van Duijn, C. M., Vernooij, M. W., Kavousi, M., Amin, N., et al. (2018). Metabolic profiling of intra- and extracranial carotid artery atherosclerosis. *Atherosclerosis* 272, 60–65. doi: 10.1016/j.atherosclerosis.2018.03.015

Wahlstrom, A., Sayin, S. I., Marschall, H. U., and Backhed, F. (2016). Intestinal Crosstalk between Bile Acids and Microbiota and Its Impact on Host Metabolism. *Cell Metab* 24, 41–50. doi: 10.1016/j.cmet.2016.05.005

Wang, Z., Klipfell, E., Bennett, B. J., Koeth, R., Levison, B. S., Dugar, B., et al. (2011). Gut flora metabolism of phosphatidylcholine promotes cardiovascular disease. *Nature* 472, 57–63. doi: 10.1038/nature09922

Weber, C., and Noels, H. (2011). Atherosclerosis: current pathogenesis and therapeutic options. *Nat Med* 17, 1410–1422. doi: 10.1038/nm.2538

Witkowski, M., Weeks, T. L., and Hazen, S. L. (2020). Gut Microbiota and Cardiovascular Disease. *Circ Res* 127, 553–570. doi: 10.1161/circresaha.120.316242

Won, T. J., Kim, B., Song, D. S., Lim, Y. T., Oh, E. S., Lee, D. I., et al. (2011). Modulation of Th1/Th2 balance by Lactobacillus strains isolated from Kimchi via stimulation of macrophage cell line J774A.1 in vitro. *J Food Sci* 76, H55–H61. doi: 10.1111/j.1750-3841.2010.02031.x

Yang, Y., Gu, H., Sun, Q., and Wang, J. (2018). Effects of Christensenella minuta lipopolysaccharide on RAW 264.7 macrophages activation. *Microb Pathog* 125, 411–417. doi: 10.1016/j.micpath.2018.10.005

Zheng, J., Hoffman, K. L., Chen, J. S., Shivappa, N., Sood, A., Browman, G. J., et al. (2020). Dietary inflammatory potential in relation to the gut microbiome: results from a cross-sectional study. *Br J Nutr* 124, 931–942. doi: 10.1017/s0007114520001853

Zhou, C., Zhao, H., Xiao, X. Y., Chen, B. D., Guo, R. J., Wang, Q., et al. (2020). Metagenomic profiling of the pro-inflammatory gut microbiota in ankylosing spondylitis. *J Autoimmun* 107, 102360. doi: 10.1016/j.jaut.2019.102360

Ziganshina, E. E., Sharifullina, D. M., Lozhkin, A. P., Khayrullin, R. N., Ignatyev, I. M., and Ziganshin, A. M. (2016). Bacterial Communities Associated with Atherosclerotic Plaques from Russian Individuals with Atherosclerosis. *PLoS One* 11:e0164836. doi: 10.1371/journal.pone.0164836

**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 Physiology

ORIGINAL RESEARCH
published: 14 October 2021
doi: 10.3389/fphys.2021.607089

# Deciphering the Protein, Modular Connections and Precision Medicine for Heart Failure With Preserved Ejection Fraction and Hypertension Based on TMT Quantitative Proteomics and Molecular Docking

*Guofeng Zhou[1†], Jiye Chen[1†], Chuanhong Wu[2†], Ping Jiang[1], Yongcheng Wang[3], Yongjian Zhang[1], Yuehua Jiang[3*] and Xiao Li[3*]*

[1] First Clinical Medical College, Shandong University of Traditional Chinese Medicine, Jinan, China, [2] The Biomedical Sciences Institute of Qingdao University (Qingdao Branch of SJTU Bio-X Institutes), Qingdao University, Qingdao, China, [3] Affiliated Hospital of Shandong University of Traditional Chinese Medicine, Jinan, China

**Background:** Exploring the potential biological relationships between heart failure with preserved ejection fraction (HFpEF) and concomitant diseases has been the focus of many studies for the establishment of personalized therapies. Hypertension (HTN) is the most common concomitant disease in HFpEF patients, but the functional connections between HFpEF and HTN are still not fully understood and effective treatment strategies are still lacking.

**Methods:** In this study, tandem mass tag (TMT) quantitative proteomics was used to identify disease-related proteins and construct disease-related networks. Furthermore, functional enrichment analysis of overlapping network modules was used to determine the functional similarities between HFpEF and HTN. Molecular docking and module analyses were combined to identify therapeutic targets for HFpEF and HTN.

**Results:** Seven common differentially expressed proteins (co-DEPs) and eight overlapping modules were identified in HFpEF and HTN. The common biological processes between HFpEF and HTN were mainly related to energy metabolism. Myocardial contraction, energy metabolism, apoptosis, oxidative stress, immune response, and cardiac hypertrophy were all closely associated with HFpEF and HTN. Epinephrine, sulfadimethoxine, chloroform, and prednisolone acetate were best matched with the co-DEPs by molecular docking analyses.

**Conclusion:** Myocardial contraction, energy metabolism, apoptosis, oxidative stress, immune response, and cardiac hypertrophy were the main functional connections between HFpEF and HTN. Epinephrine, sulfadimethoxine, chloroform, and prednisolone acetate could potentially be effective for the treatment of HTN and HFpEF.

**Keywords: hypertension, heart failure with preserved ejection fraction, molecular docking, modular, therapeutic prediction**

OPEN ACCESS

***Edited by:***
*Sanjay Kumar Banerjee, National Institute of Pharmaceutical Education and Research, Guwahati, India*

***Reviewed by:***
*Vikas Kumar, University of Nebraska Medical Center, United States*
*Caroline Evans, The University of Sheffield, United Kingdom*

***Correspondence:**
*Yuehua Jiang*
*jiang_yuehua@hotmail.com*
*Xiao Li*
*lixiao617@hotmail.com*

*[†] These authors have contributed equally to this work*

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

***Received:*** *17 September 2020*
***Accepted:*** *23 September 2021*
***Published:*** *14 October 2021*

***Citation:***
*Zhou G, Chen J, Wu C, Jiang P, Wang Y, Zhang Y, Jiang Y and Li X (2021) Deciphering the Protein, Modular Connections and Precision Medicine for Heart Failure With Preserved Ejection Fraction and Hypertension Based on TMT Quantitative Proteomics and Molecular Docking. Front. Physiol. 12:607089. doi: 10.3389/fphys.2021.607089*

## INTRODUCTION

Heart failure with preserved ejection fraction (HFpEF), which is a complex syndrome characterized by a normal left ventricular ejection fraction and abnormal diastolic function, accounts for more than 50% of heart failure (HF) patients (Pieske et al., 2019). Current therapies for HFpEF include strategies to manage the coexisting conditions, reduce symptoms, and treat volume overload when necessary (Redfield, 2016). Although there has been much progress in HFpEF-related research, an effective strategy for HFpEF treatment has not yet been established (Gazewood and Turner, 2017). Compared with heart failure with reduced ejection fraction (HFrEF), HFpEF is heterogeneous, and drugs effective against HFrEF are not suitable for HFpEF (Graziani et al., 2018). Thus, a complete clinical phenotypic classification of HFpEF, including tiology, concomitant diseases, and risk factors, is required. Furthermore, exploring the underlying biological functions involved in the different types of HFpEF will help develop personalized therapies and precision medicines for HFpEF (Borlaug, 2020; Ge, 2020).

Patients with HFpEF that are diagnosed with hypertension (HTN) and coronary heart disease are regarded as having vascular-related HFpEF (Ge, 2020). Studies have shown that HTN is the most common complication in patients with HFpEF, but the biological relationship between HTN and HFpEF is still not fully understood (Tadic et al., 2018). In addition, studies have suggested that HTN is an additional risk factor for HFpEF (Dunlay et al., 2017). HFpEF and HTN share many common pathogeneses, such as dysfunction of cardiac autonomy, imbalance of the renin-angiotensin-aldosterone system, and excessive oxidative stress. In addition, some underlying biological mechanisms play an important role in the transition from HTN to HFpEF. For example, hypertension leads to diastolic dysfunction and concentric remodeled left ventricular decompensation, resulting in HFpEF (Drazner, 2011; Heinzel et al., 2015; Messerli et al., 2017; Nwabuo and Vasan, 2020). In addition, HTN also activates chronic inflammation and increases collagen deposition, further exacerbating left ventricular dysfunction (Paulus and Tschöpe, 2013). Studies have reported that myocardial contractile dysfunction, right ventricular dysfunction, arterial stiffness, ventricular-arterial coupling, and microvascular dysfunction could increase the risk of HFpEF in patients with HTN (Hicklin et al., 2020). However, in clinical trials, drugs such as angiotensin-converting enzyme inhibitors, angiotensin II receptor blockers, diuretics, and beta-blockers, which showed beneficial effects against common pathogeneses of HFpEF and HTN, did not produce significant positive effects in patients with HFpEF (Kjeldsen et al., 2020). Therefore, further research is needed to explore the potential biological relationships between HFpEF and HTN.

Disease network construction provides a solution to explore the relationships between diseases (Le and Pham, 2017), where modules of disease-related networks are responsible for various features of the diseases. Functional enrichment analysis of the overlapping modules reflects the functional links among related diseases (Dean et al., 2017). For example, using this method, a previous study showed that the negative regulation of transcription from RNA polymerase II promoter RNA and the negative regulation of apoptotic processes are overlapping biological functions among type-2 diabetes mellitus, prostate cancer, and chronic myeloid leukemia (Liu et al., 2019). Furthermore, another study showed that atherosclerosis, cholesterol homeostasis, plasma lipoprotein particle remodeling, and oxidative stress responses are common risk factors for stroke and coronary heart disease (Zhang et al., 2014).

The Dahl salt-sensitive (DS) rat model has been implemented for the study of HFpEF (Cho et al., 2017). Toward that, DS rats diagnosed with HTN or HFpEF were analyzed using proteomics. Cytoscape software and STRING platforms were used to construct a disease network. Modules of the disease network were divided using Molecular Complex Detection (MCODE). Gene Ontology (GO) enrichment analysis was performed to identify the significant functions and pathways of overlapping modules found in the Database for Annotation, Visualization, and Integrated Discovery (DAVID). Molecular docking and module analyses were combined to contribute to the development of personalized therapies and precision medicines for HFpEF and HTN treatment. A flowchart of the research design is shown in **Figure 1**.

## MATERIALS AND METHODS

### Animals and Experimental Protocols

Specific pathogen-free 6-week-old male DS rats (weight: 160–180 g; Certificate No. 2016-0006) were obtained from the Charles River Animal Laboratory (Beijing, China). Rats were housed in groups of six rats per cage under controlled conditions (12 h dark/light cycle, temperature: 20–24°C, relative humidity: 40–60%, dB $\leq$ 60) and with free access to water and food. After a week of adaptation, the rats were randomly divided into the following three groups: the HTN group (8% NaCl chow for 7 weeks, $n = 6$), the HFpEF group (8% NaCl chow for 11 weeks, $n = 6$), and the control group (0.3% NaCl chow for 7 or 11 weeks, $n = 12$). All experiments were reviewed by the Animal Ethics Committee of the Shandong University of Traditional Chinese Medicine (Ethics No. SDUTCM2018071501).

### Tissue Collection

The rats in the HTN and HFpEF groups were euthanized after 7 and 11 weeks of the high-salt diet (HSD), respectively, and the rats in the control group were randomly sacrificed after 7 or 11 weeks of the control diet. Pentobarbital (20 mg/kg, i.p.) was used for anesthesia in rats, and the left ventricle (LV) was collected from each rat and stored at −80°C.

### Tandem Mass Tag-Labeled Quantitative Proteomics

Twelve LV samples [HTN group $n = 3$, HFpEF group $n = 3$, control group (euthanized at 7 weeks) $n = 3$, control group (euthanized at 11 weeks) $n = 3$] were collected for TMT quantitative proteomics. Previous studies reported that data with three samples in each group could reliably be statistically

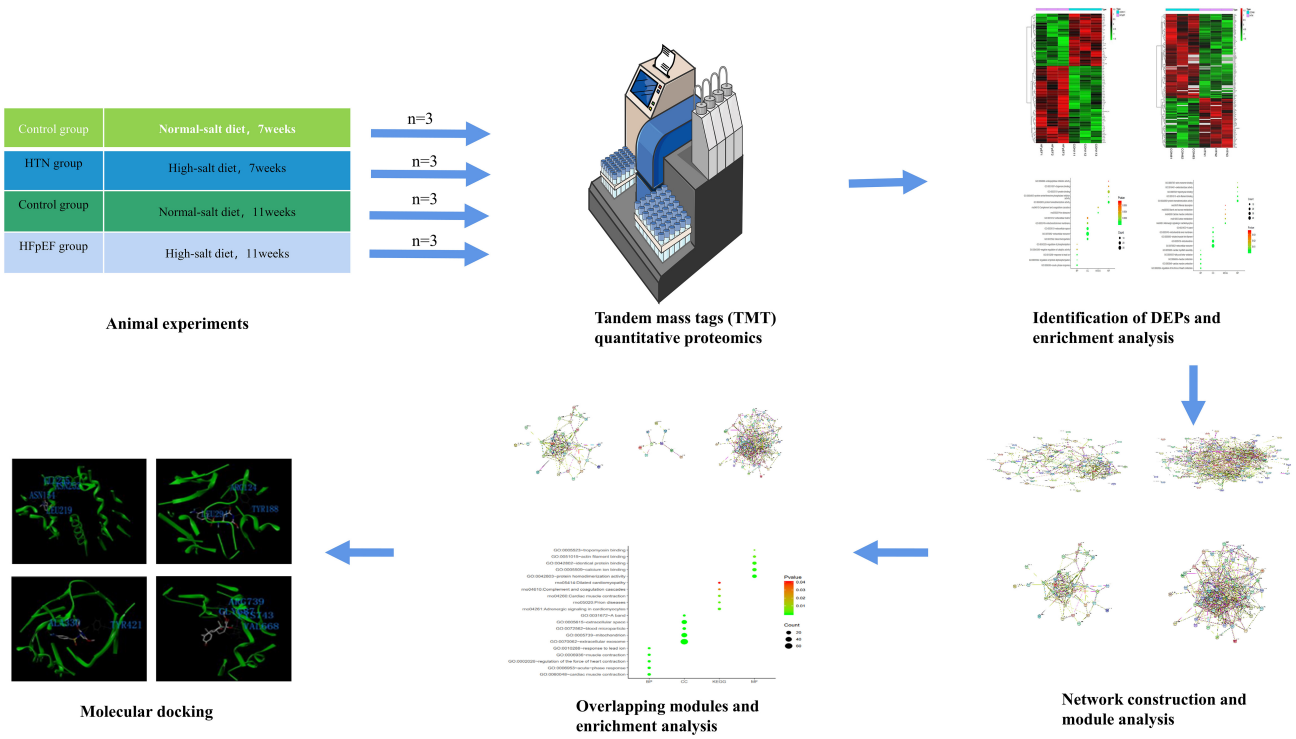



**FIGURE 1 |** Flowchart of the research design. Deciphering the protein networks and modular connections, and targeting precision medicines for HFpEF and HTN using TMT quantitative proteomics and molecular docking analyses.

analyzed (Maitiabola et al., 2020; Yan et al., 2020). The tissue was removed from the refrigerator at −80°C, ground into powder, and quickly transferred to a centrifuge tube pre-cooled with liquid nitrogen. PASP protein lysate (100 mM ammonium bicarbonate, 8 M urea, pH 8) was added to the liquid nitrogen, shaken, mixed, and ultrasonicated in an ice water bath for 5 min, followed by centrifugation at 12,000 rpm for 15 min at 4°C. Then, the supernatant was collected, 10 nM dithiothreitol was added, and the mixture was incubated at 56°C for 1 h. Then, iodoacetamide was added and the reaction was allowed to proceed for 1 h in the absence of light. Next, four volumes of pre-cooled acetone were used for precipitation, followed by centrifugation at 12,000 rpm for 15 min at 4°C, after which the precipitate was collected. The precipitate was resuspended and washed with one milliliter of −20°C pre-cooled acetone, followed by a second centrifugation at 12,000 rpm for 15 min at 4°C. Then, the precipitate was collected and air dried, and an appropriate amount of protein dissolving solution (8 M urea, 100 mM TEAB, pH 8.5) was used to dissolve the protein precipitate.

The Bradford protein quantification kit (Beyotime, China) was used to determine the protein concentration. DB protein dissolving solution (8 M urea, 100 mM TEAB, pH 8.5) was added to the protein sample to a volume of 100 μL, trypsin and 100 mM buffer were added, and mixing and digestion were performed at 37°C for 4 h. Then, pancreatin and $CaCl_2$ were used for digestion overnight. Formic acid was used to adjust the pH to less than 3, mixing was done at room temperature, and centrifugation was performed at 12,000 rpm for 5 min. The supernatant was then slowly passed through the C18 desalting column, and the cleaning solution (0.1% formic acid, 3% acetonitrile) was used for washing three times. In addition, an appropriate amount of eluent (0.1% formic acid, 70% acetonitrile) was added, and the filtrate was collected and lyophilized. One hundred microliters of 0.1 M TEAB buffer was used for reconstitution, and 41 μL of TMT labeling reagent was dissolved in acetonitrile. The mixture was mixed at room temperature for 2 h, and 8% ammonia was added to stop the reaction. An equal volume of the labeled sample was used for mixing and freeze-drying after desalting.

Mobile phase A solution (2% acetonitrile, 98% water, pH 10) and mobile phase B solution (98% acetonitrile, 2% water) were prepared. The freeze-dried powder was dissolved in solution A and centrifuged at 12,000 rpm for 10 min at room temperature. An L-3000 HPLC system and a water VEHC 18 (4.6 mm × 250 mm, 5 μm) were used for this study, and the column temperature was set to 45°C. Details of the elution gradient are shown in **Table 1**. One tube was collected every minute, divided into 10 fractions, freeze-dried, and dissolved in 0.1% formic acid.

Mobile phase A solution (100% water, 0.1% formic acid) and phase B solution (80% acetonitrile, 0.1% formic acid) were prepared. One microgram of the supernatant from each fraction was used for the test. The UHPLC system was upgraded

**TABLE 1 |** The elution gradient table of peptide fraction separation liquid chromatography.

| Time (min) | Flow rate (mL/min) | Mobile phase A (%) | Mobile phase B (%) |
|---|---|---|---|
| 0 | 1 | 97 | 3 |
| 10 | 1 | 95 | 5 |
| 30 | 1 | 80 | 20 |
| 48 | 1 | 60 | 40 |
| 50 | 1 | 50 | 50 |
| 53 | 1 | 30 | 70 |
| 54 | 1 | 0 | 100 |

with the EASY-nLC 1200 system. We prepared both the pre-column (4.5 cm × 75 μm, 3 μm) and the analytical column (15 cm × 150 μm, 1.9 μm). The elution conditions for liquid chromatography are shown in **Table 2**. A Q Exactive series mass spectrometer was used for this study, the ion source was the Nanospray Flex, the ion spray voltage was 2.3 kV, the temperature of the ion transfer tube was 320°C, and the data-dependent acquisition mode was used. The full scan range of the mass spectrum was 350–1,500 m/z. The resolution of the primary mass spectrometry was set to 60,000 (200 m/z), the maximum capacity of the C-trap was $3 \times 10^6$, and the maximum injection time of the C-trap was 20 ms. The top 40 precursor ions were selected for the full scan, and the higher-energy collision dissociation (HCD) method was used for the fragment, which contributed to the secondary mass spectrometry detection. The isolation window of MS2 spectrum is 2 m/z. HCD spectrum ranged from 120 to m/z (precursor ion) × z (charge number) + 100 m/z. The resolution was 45,000 (200 m/z), the maximum capacity of the C-trap was $5 \times 10^4$, the maximum injection time of the C-trap was 86 ms, the threshold intensity was $1.2 \times 10^5$, and the dynamic exclusion range was 20 s.

MS/MS raw files were processed using the MASCOT engine (Matrix Science, London, United Kingdom; version 2.6) embedded into Proteome Discoverer software, and searched against the UniProt database, including Uniprot_RattusNorvegicus_36080_20180123 sequences[1]. The search parameters included trypsin as the enzyme used to generate peptides with a maximum of two missed cleavages permitted. A precursor mass tolerance of 10 ppm was specified along with a 0.05 Da tolerance for MS2 fragments. Except for the TMT labels, carbamidomethyl (C) was set as a fixed modification. The variable modifications were oxidation (M) and acetyl (protein N-term). A peptide and protein false discovery rate of 1% was enforced using a reverse database search strategy. The quantitative values of proteins obtained from two pairs of samples were examined using the *t*-test, and the *p*-values were calculated. Fold change >1.1, fold change <0.91, and *P*-value < 0.05, were considered to filter differentially expressed proteins (DEPs). Proteomic data is provided as **Supplementary Material**.

[1]http://www.uniprot.org

**TABLE 2 |** Elution gradient table of liquid chromatography.

| Time (min) | Flow rate (nL/min) | Mobile phase A (%) | Mobile phase B (%) |
|---|---|---|---|
| 0 | 600 | 94 | 6 |
| 2 | 600 | 85 | 15 |
| 78.5 | 600 | 60 | 40 |
| 80.5 | 600 | 50 | 50 |
| 81.5 | 600 | 45 | 55 |
| 90 | 600 | 0 | 100 |

## Constructing the Protein-Protein Interaction Networks for Heart Failure With Preserved Ejection Fraction and Hypertension

The STRING database (version 10.5)[2] was used to predict protein interactions and functional associations. The PPI networks of HFpEF- and HTN-DEPs were obtained under controlled parameters (interaction score >0.4). PPI networks were analyzed using Cytoscape (version 3.6.1)[3].

## Functional Enrichment Analysis

HFpEF- and HTN-DEPs were submitted to the Database for annotation, visualization, and integrated discovery for functional enrichment, including GO and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analyses (version 6.8)[4]. The *P*-value was set at <0.05 for GO and KEGG pathway enrichment, as is standard in the field (Yuan et al., 2016; Liu et al., 2019).

## Division and Identification of Network Modules

Disease-related networks were analyzed using Cytoscape (version 3.6.1)[5]. Furthermore, the modules were divided by Molecular Complex Detection (MCODE, version 1.3.2)[6]. The modules were obtained under controlled parameters (degree cutoff = 2, node score cutoff = 0.1, core threshold $K = 2$, flux density cutoff = 0.1, K-core 2, max. depth = 100). The parameters for Cytoscape were set as default, as recommended by previous studies (Yuan et al., 2016; Liu et al., 2019).

## Identification of Modern Medicine Symptoms Related to Common Differentially Expressed Proteins and Links to Cardiovascular Diseases

The association between the co-DEPs and cardiovascular diseases was analyzed using the Comparative Toxicogenomics Database (CTD)[7], which integrates the relationships between gene

[2]http://string-db.org/
[3]https://cytoscape.org/
[4]https://david.ncifcrf.gov/
[5]https://www.cytoscape.org/
[6]https://baderlab.org/Software/MCODE
[7]http://ctdbase.org/

products, diseases, chemicals, and environments. Furthermore, related MM symptoms of the co-DEPs were observed in SymMap[8] and in previous publications.

## Drug Discovery and Molecular Docking

Small-molecule compounds related to the co-DEPs were observed from the DrugBank[9], through the target search. Molecular docking analysis is a common strategy for drug discovery. The structures of the co-DEPs were downloaded from the Protein Data Bank (PDB)[10], and the PDB IDs of the proteins are shown in **Table 3**. The structures of the drugs were obtained from PubChem[11], and the CIDs are shown in **Table 4**. Eutectic ligands and water molecules were removed, and residue repair, side chain fixation, and hydrogenation were used for protein preparation. The Surflex-Dock module of SYBYL 2.1 was used for molecular docking. Furthermore, AMBR7 was used for energy optimization, and the active pockets were obtained in automatic mode. The parameters for SYBYL 2.1 were set as default, as recommended by previous studies (Tan et al., 2020). The molecular docking scores reflected the binding effects between the small-molecule compounds and the co-DEPs, and the highest scoring small molecules were considered as potential drugs.

# RESULTS

## Identification of Differentially Expressed Proteins

A total of 83 DEPs, including 52 upregulated proteins, were obtained by comparing the HFpEF group with the control group sacrificed at 11 weeks (**Figure 2A**). A total of 132 DEPs, including 85 upregulated proteins, were identified by comparing the HTN group with the control group sacrificed at 7 weeks (**Figure 2B**). Among the HFpEF-DEPs and HTN-DEPs, there were 7 co-DEPs, including haptoglobin (Hp), coenzyme Q9 (COQ9), serotransferrin (Tf), major prion protein (Prnp), acetyl-CoA acetyltransferase, mitochondrial (Acat1), translocase of inner mitochondrial membrane 44 (Timm44), and ATP-binding cassette sub-family B member 6 (Abcb6; **Figure 2C**). Furthermore, the CTD database showed links between the co-DEPs and various cardiovascular diseases (**Figures 3A–G**). Finally, the related MM symptoms of the co-DEPs were determined using SymMap[12] and previous publications (**Figure 3H**).

[8]https://www.symmap.org/
[9]https://www.drugbank.ca/
[10]http://www.rcsb.org/
[11]https://pubchem.ncbi.nlm.nih.gov/

**TABLE 3 |** PDB ID of protein.

| Protein | PDB ID |
|---|---|
| Hp | 4X0lL |
| COQ9 | 6awl |
| Tf | 1ryo |
| Prnp | 2ol9 |
| Acta1 | 2ib8 |
| Timm44 | 2cw9 |
| Abcb6 | 3nh6 |

**TABLE 4 |** CIDs of molecule compounds.

| Molecule | PubChem ID |
|---|---|
| Prednisolone acetate | 5834 |
| Bismuth subsalicylate | 16682734 |
| Phenoxymethylpenicillin | 6869 |
| Polyethylene glycol | 40786 |
| Prednisolone | 5755 |
| Chloroform | 6212 |
| Salicylic acid | 338 |
| Epinephrine | 5816 |
| Triptorelin | 25074470 |
| Benzylpenicillin | 5904 |
| Propofol | 4743 |
| Sulfadimethoxine | 5323 |

## Functional Enrichment Analysis

All HFpEF- and HTN-DEPs were submitted to DAVID for GO and KEGG functional enrichment analyses. The DEPs of HFpEF were mainly involved in immune response, energy metabolism, inflammation response, and post-translational modification (**Figure 4**). For example, DnaJ heat shock protein family (Hsp40) member A1 (Dnaja1), Hp, immunoglobulin heavy chain 6 (Igh-1a), milk fat globule EGF and factor V/VIII domain containing (Mfge8), transforming growth factor beta 1 induced transcript 1 (Tgfb1i1), apolipoprotein M (Apom), paraoxonase 1 (Pon1), protein tyrosine phosphatase, and non-receptor type 6 (Ptpn6) were closely related to immune response; acetyl-CoA acetyltransferase, mitochondrial (Acta1), Hp, mitochondrial inner membrane protein (Oxa1l), glutamine fructose-6-phosphate transaminase 1 (Gfpt1), UDP-N-acetylglucosamine pyrophosphorylase 1 (Uap1), calponin 3 (Cnn3), and follistatin-like 1 (Fstl1) were involved in energy metabolism; serpin family A member 1 (Serpina1), alpha-2-HS-glycoprotein (Ahsg), serpin family A member 10 (Serpina10), murinoglobulin 1 (Mug1), SMAD family member 1 (Smad1), and kininogen 1 (Kng1) were connected with inflammation response; mannosidase, alpha, class 2A, member 2 (Man2a2), N-glycanase 1 (Ngly1), protein phosphatase 1, regulatory (inhibitor) subunit 14B (Ppp1r14b), protein phosphatase 1, regulatory (inhibitor) subunit 14C (Ppp1r14c), protein phosphatase 1, and regulatory (inhibitor) subunit 14A (Ppp1r14a) were involved in post-translational modification. For HTN, myocardial contraction, energy metabolism, apoptosis, and oxidative stress were the main biological functions (**Figure 5**). Carcass protein in high growth mice 3 (Carp3), myosin light chain 3 (Myl3), titin (Ttn), tropomodulin (Tmod1), hydroxysteroid (17-beta) dehydrogenase 4 (Hsd17b4), actinin alpha 2 (Actn2), Acta1,

[12]https://www.symmap.org

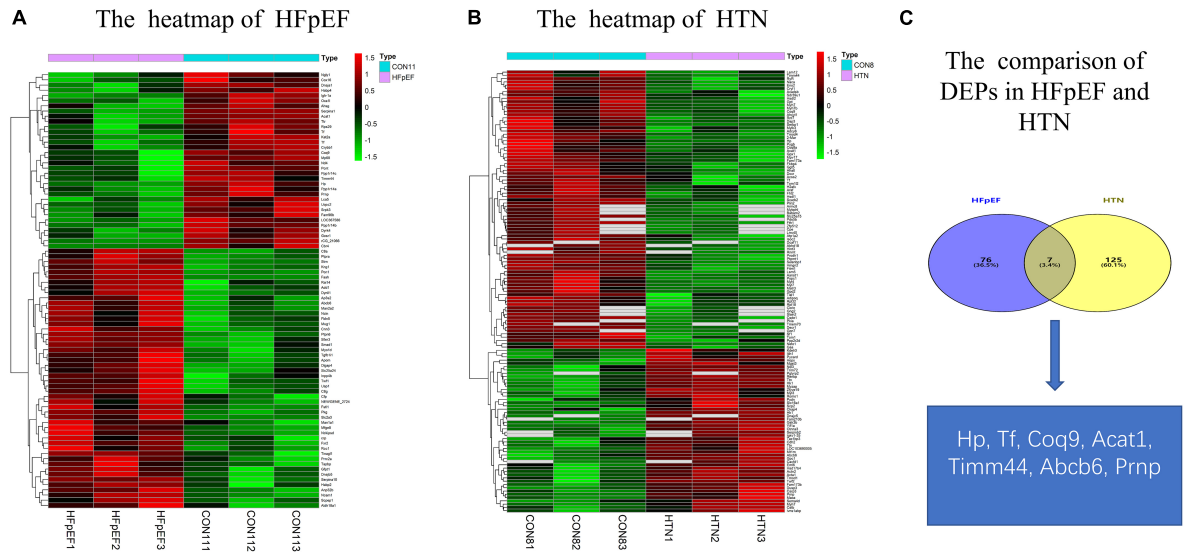



**FIGURE 2 |** Comparison of the DEPs in HFpEF and HTN. **(A)** Heatmap of HFpEF. **(B)** Heatmap of HTN. **(C)** The co-DEPs between HFpEF and HTN. Hp, haptoglobin; COQ9, coenzyme Q9; Tf, serotransferrin; Prnp, major prion protein; Acat1, acetyl-CoA acetyltransferase, mitochondrial; Timm44, translocase of inner mitochondrial membrane 44; Abcb6, ATP-binding cassette sub-family B member 6.

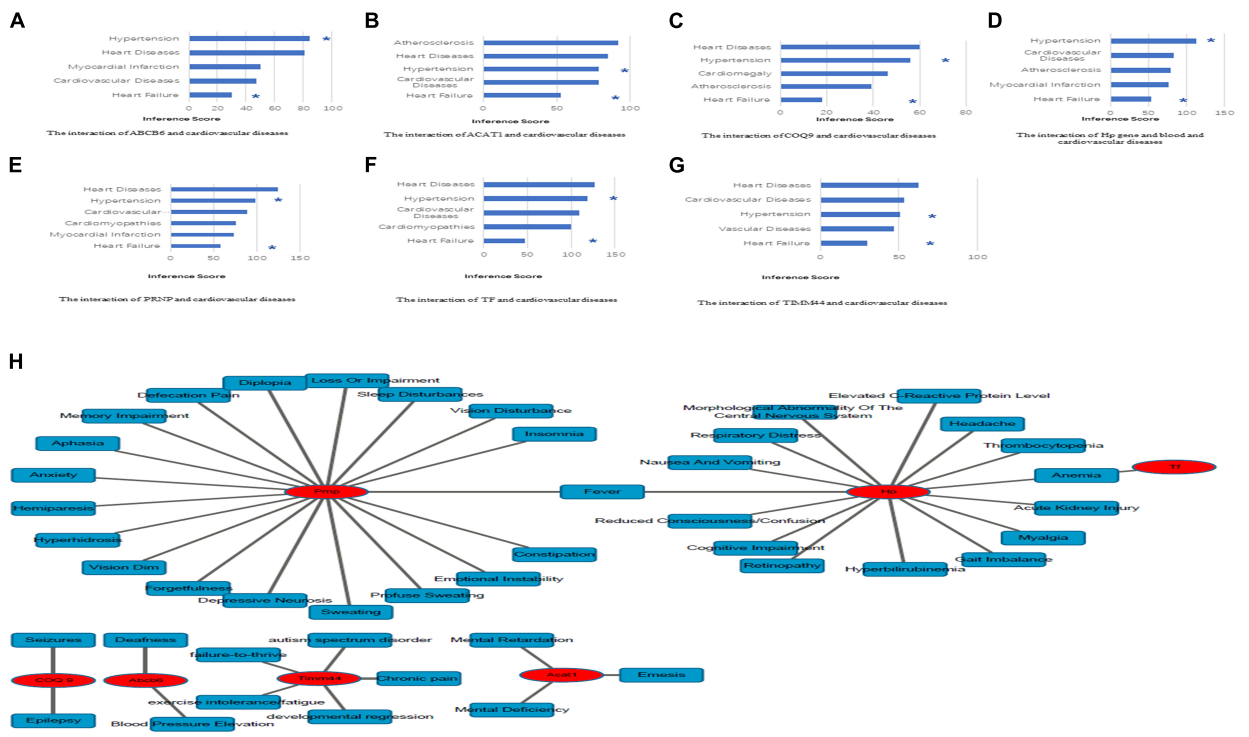



**FIGURE 3 |** Identification of co-DEP-related MM symptoms and links to cardiovascular diseases. **(A)** Abcb6, **(B)** Acat1, **(C)** COQ9, **(D)** Hp, **(E)** Prnp, **(F)** Tf, and **(G)** Timm44. *, direct evidence. **(H)** The related MM symptoms of the co-DEPs. Hp, haptoglobin; COQ9, coenzyme Q9; Tf, serotransferrin; Prnp, Major prion protein; Acat1, acetyl-CoA acetyltransferase, mitochondrial; Timm44, translocase of inner mitochondrial membrane 44; Abcb6, ATP-binding cassette sub-family B member 6.

myosin, light chain 4 (Myl4), Tf, alpha glucosidase (Gaa), perilipin 2 (Plin2), ATPase Na+/K+ transporting subunit alpha 2 (Atp1a2), enolase 2 (Eno2), myosin heavy chain 7 (Myh7), glutathione peroxidase 1 (Gpx1), tropomodulin 4 (Tmod4), Acta1, Hp, and myosin light chain kinase 3 (Mylk3) were closely related to myocardial contraction; glycogen synthase kinase

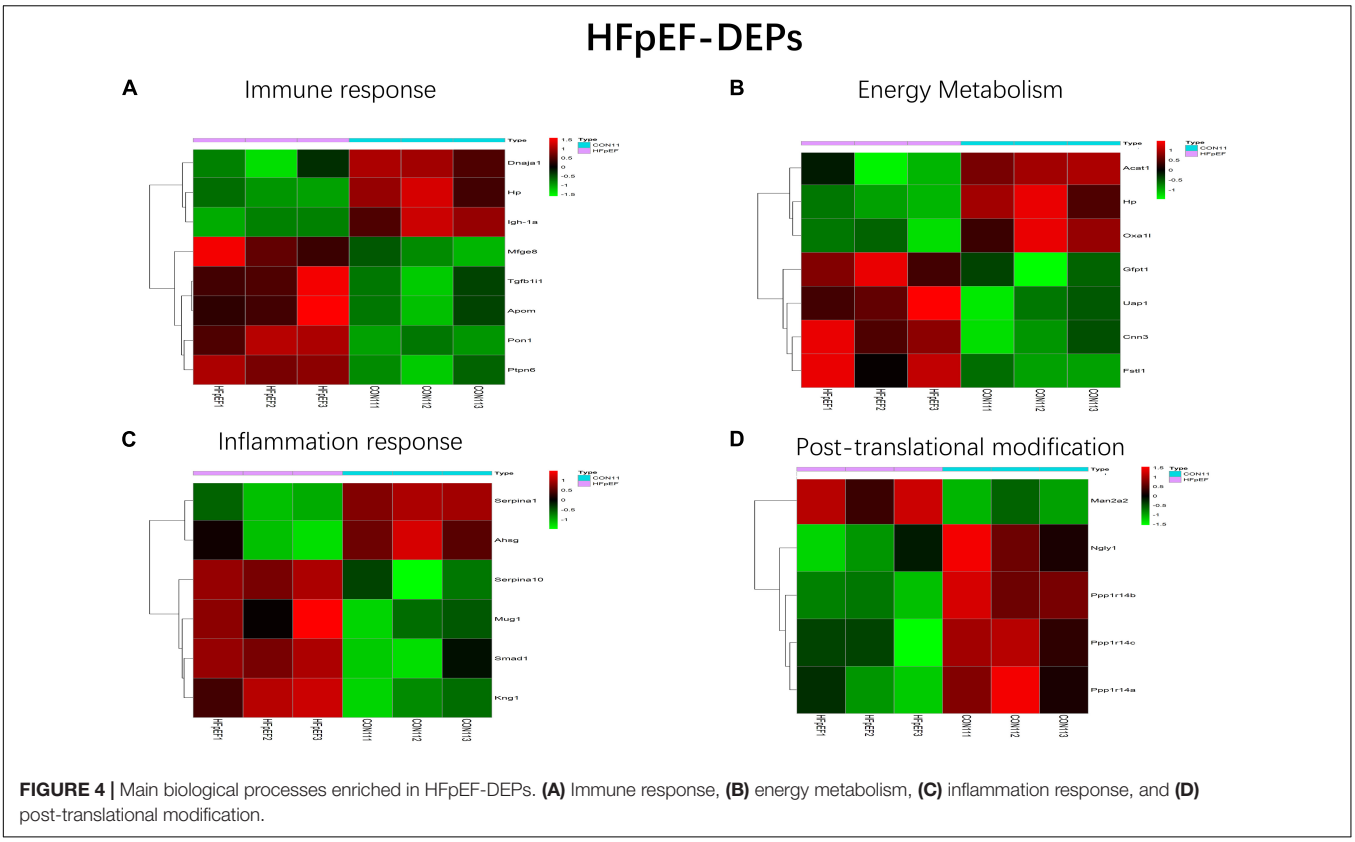



**FIGURE 4 |** Main biological processes enriched in HFpEF-DEPs. **(A)** Immune response, **(B)** energy metabolism, **(C)** inflammation response, and **(D)** post-translational modification.

3 beta (Gsk3b), hydroxysteroid (17-beta) dehydrogenase 4 (Hsd17b4), alpha glucosidase (Gaa), 2,4-dienoyl-CoA reductase 1 (Decr1), acyl-CoA dehydrogenase, short/branched chain (Acadsb), adiponectin C1Q and collagen domain containing (Adipoq), Acta1, and glycogen phosphorylase B (Pygb) were involved in energy metabolism; nucleolar protein 3 (Nol3), prion protein (Prnp), serpin family B member 2 (Serpinb2), glycogen synthase kinase 3 beta (Gsk3b), hexokinase 1 (Hk1), ferritin heavy chain 1 (Fth1), four and a half LIM domains 2 (Fhl2), A-Raf proto-oncogene, serine/threonine kinase (Araf), glutathione peroxidase 1 (Gpx1), nascent polypeptide associated complex subunit alpha (Naca), and ring finger protein 7 (Rnf7) were connected with apoptosis.

The top five biological processes among HFpEF-DEPs were acute-phase response (count 5, *P*-Value 2.15E-05), regulation of protein dephosphorylation (count 3, *P*-Value 2.75E-04), response to lead ion (count 4, *P*-Value 6.39E-04), negative regulation of catalytic activity (count 4, *P*-Value 0.003203462), and regulation of phosphorylation (count 3, *P*-Value 0.003348677). Blood microparticles (count 10, *P*-Value 1.80E-09), extracellular exosome (count 33, *P*-Value 1.01E-08), extracellular space (count 18, *P*-Value 3.18E-05), mitochondrial inner membrane (count 7, *P*-Value 0.002195356), and extracellular matrix (count 6, *P*-Value 0.00460952) related cell compositions were significantly enriched. Furthermore, the main enriched molecular functions were protein homodimerization activity (count 12, *P*-Value 5.67E-04), protein serine/threonine phosphatase inhibitor activity (count 3, *P*-Value 9.62E-04), protein binding (count 16, *P*-Value 0.002503286), chaperone binding (count 4, *P*-Value 0.004337384), and endopeptidase inhibitor activity (count 3, *P*-Value 0.007219962). The enriched KEGG pathways were prion diseases (count 4, *P*-Value 3.21E-04) and the complement and coagulation cascades (count 4, *P*-Value 0.003144821) (**Figure 6A**). With respect to HTN-DEPs, the most enriched biological processes were the regulation of heart contraction force (count 6, *P*-Value 2.73E-07), cardiac muscle contraction (count 7, *P*-Value 5.96E-07), muscle contraction (count 6, *P*-Value 1.50E-05), fatty acid beta-oxidation (count 5, *P*-Value 1.95E-04), and cardiac myofibril assembly (count 3, *P*-Value 0.002067005). The main components were extracellular exosome (count 48, *P*-Value 3.12E-11), mitochondrion (count 32, *P*-Value 3.12E-08), striated muscle thin filament (count 4, *P*-Value 3.24E-05), mitochondrial inner membrane (count 10, *P*-Value 2.30E-04), and a band (count 4, *P*-Value 4.49E-04). The molecular functions of HTN-PEGs were mainly enriched in protein homodimerization activity (count 15, *P*-Value 4.74E-04), actin filament binding (count 6, *P*-Value 0.001790743), tropomyosin binding (count 3, *P*-Value 0.003348376), oxidoreductase activity (count 5, *P*-Value 0.00825633), and actin monomer binding (count 3, *P*-Value 0.010556462). The enriched KEGG pathways were adrenergic signaling in cardiomyocytes (count 6, *P*-Value 0.004149388), carbon metabolism (count 5, *P*-Value 0.014082125), cardiac muscle contraction (count 4, *P*-Value 0.022270779), starch and sucrose metabolism (count 3, *P*-Value 0.023739862), and mineral absorption (count 3, *P*-Value 0.03637201) (**Figure 6B**).

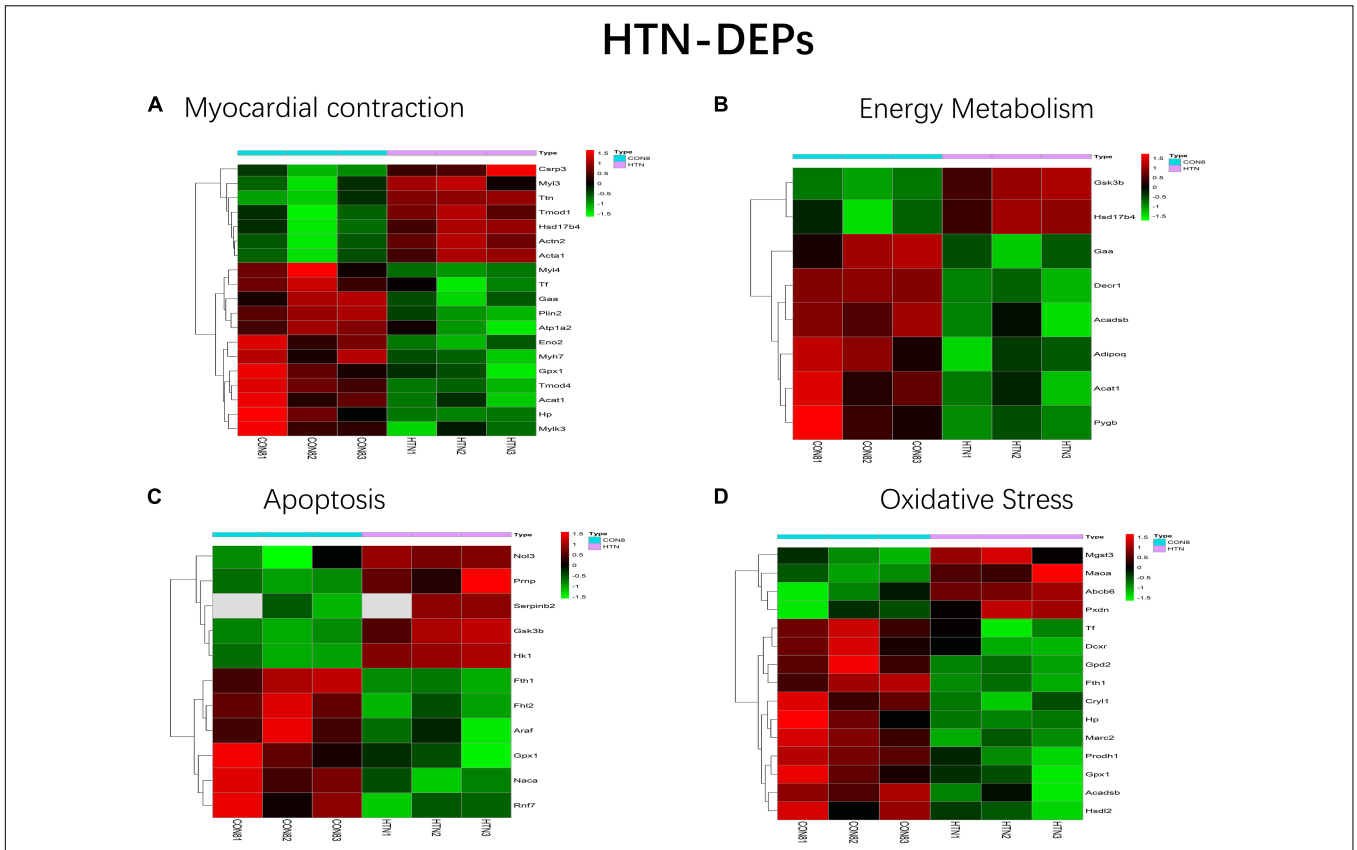



**FIGURE 5 |** Main biological processes enriched in HTN-DEPs. **(A)** Myocardial contraction, **(B)** energy metabolism, **(C)** apoptosis, and **(D)** oxidative stress.

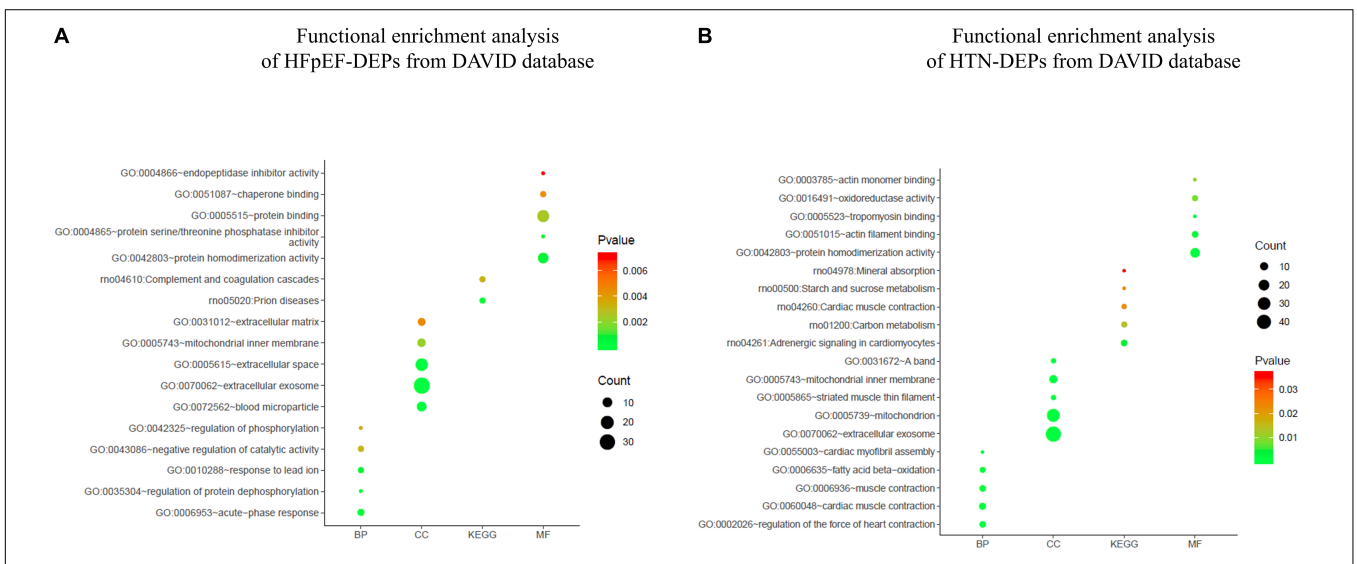



**FIGURE 6 |** Functional enrichment analysis. **(A)** Functional enrichment analysis of HFpEF-DEPs from the DAVID database. **(B)** Functional enrichment analysis of HTN-DEPs from the DAVID database.

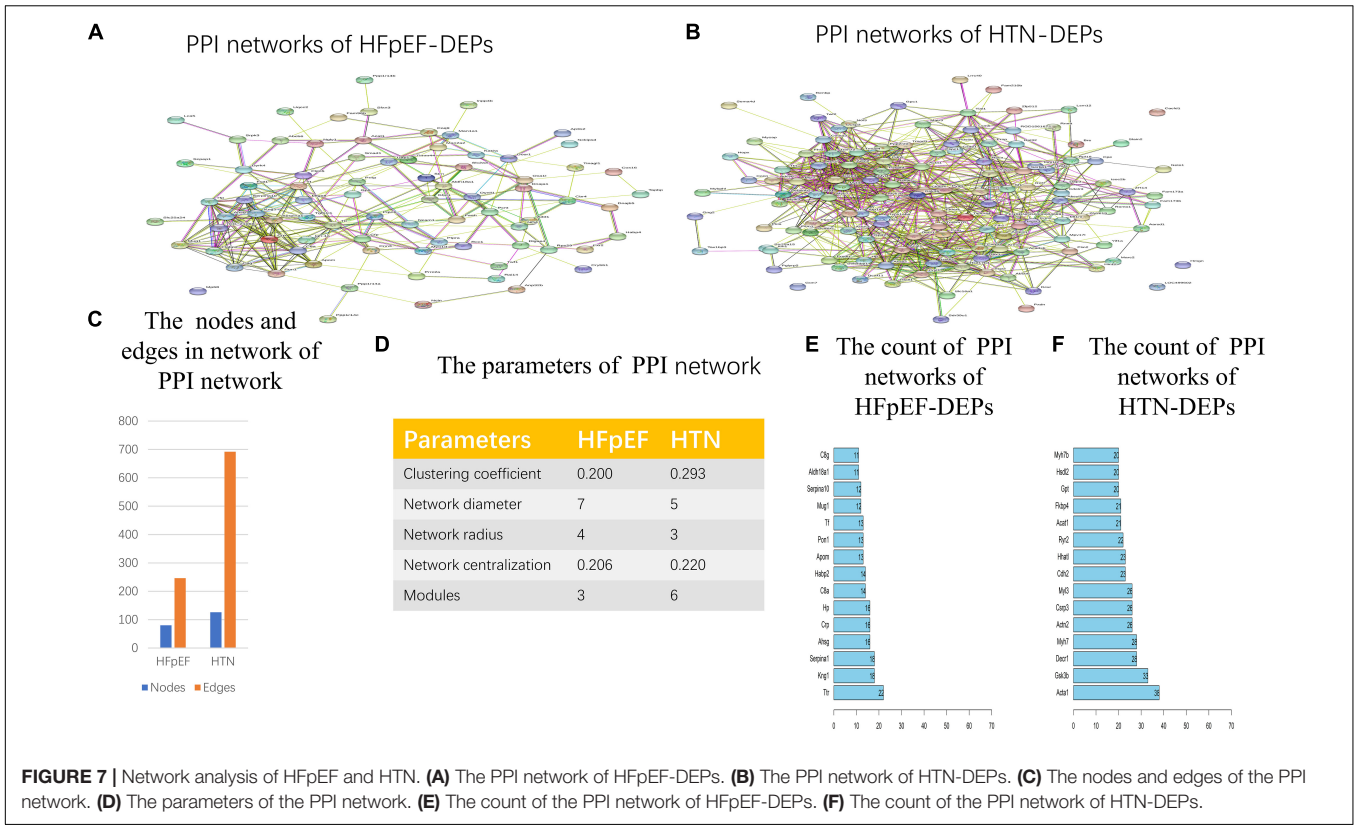



**FIGURE 7 |** Network analysis of HFpEF and HTN. **(A)** The PPI network of HFpEF-DEPs. **(B)** The PPI network of HTN-DEPs. **(C)** The nodes and edges of the PPI network. **(D)** The parameters of the PPI network. **(E)** The count of the PPI network of HFpEF-DEPs. **(F)** The count of the PPI network of HTN-DEPs.

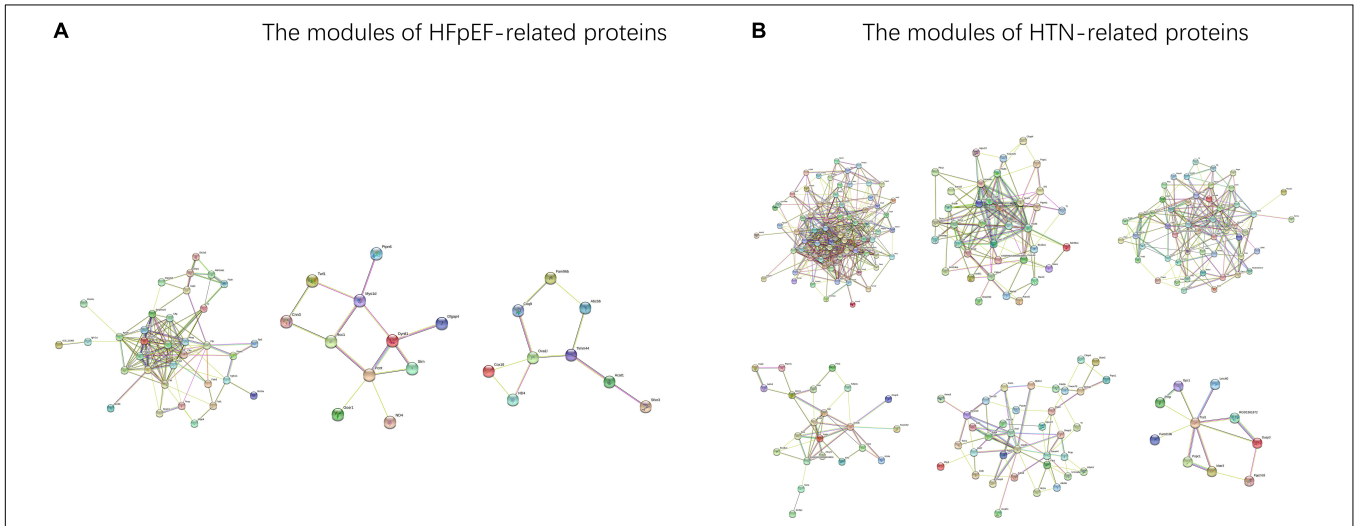



**FIGURE 8 |** The modules of the HFpEF- and HTN-DEPs. **(A)** The modules of the HFpEF-DEPs. **(B)** The modules of the HTN-DEPs.

## Protein-Protein Interaction Network Analysis and Modularity Analysis

In total, from the PPI network of HFpEF-DEPs (**Figure 7A**) and that of the HTN-DEPs (**Figure 7B**), 80 nodes and 246 edges, and 126 nodes and 692 edges were identified, respectively (**Figure 7C**). The parameters of the PPI network of the HFpEF- and HTN-DEPs are shown in **Figure 7D**. Furthermore, transthyretin (Ttr; degree = 22), kininogen-1 (Kng1; degree = 18), alpha-1-antiproteinase (Serpina1; degree = 18), alpha-2-HS-glycoprotein (Ahsg; degree = 16), and pentaxin (Crp; degree = 16) were identified as hub proteins in the HFpEF-DEP PPI network (**Figure 7E**). Acetyl-CoA acetyltransferase, mitochondrial (Acta1; degree = 38), glycogen synthase kinase 3 beta (Gsk3b; degree = 33), 2,4-dienoyl-CoA reductase 1 (Decr1; degree = 28), myosin heavy chain 7 (Myh7; degree = 28), and actinin alpha 2 (Actn2;

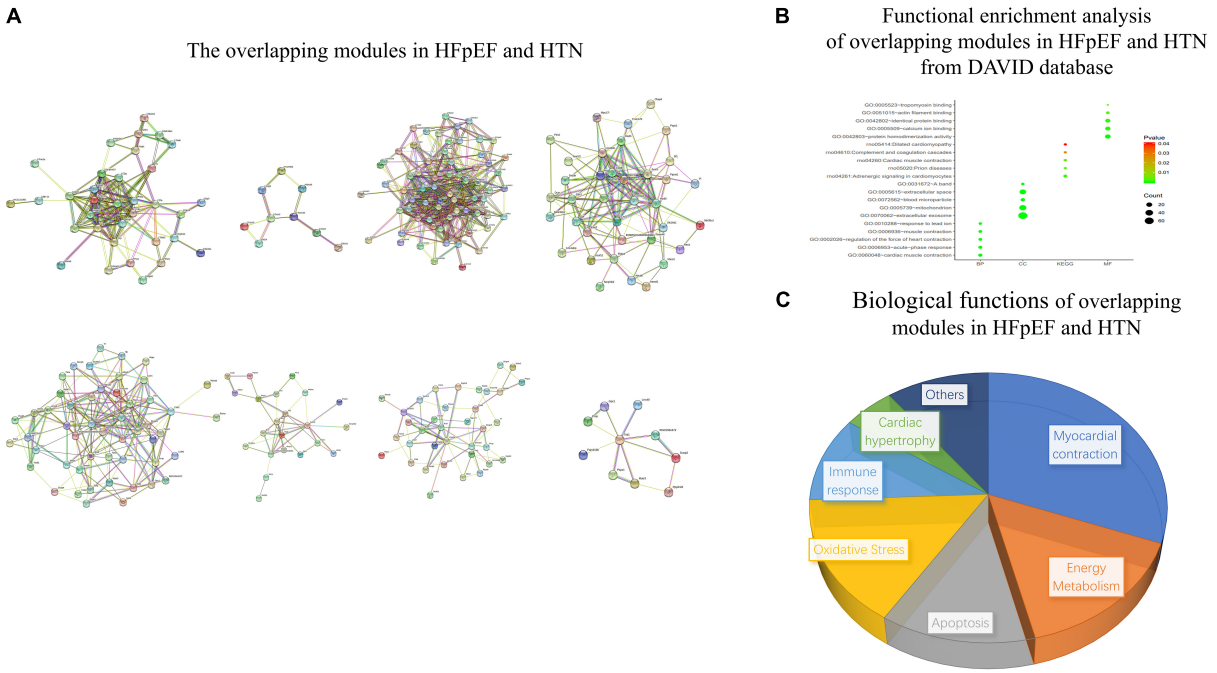



**FIGURE 9 |** Functional enrichment analysis of overlapping modules between the HFpEF and HTN networks. **(A)** The overlapping modules in HFpEF and HTN. **(B)** Functional enrichment analysis of overlapping modules in HFpEF and HTN from the DAVID database. **(C)** Biological functions of overlapping modules in HFpEF and HTN.

degree = 26) were identified as hub proteins in the HTN-DEP PPI network (**Figure 7F**). Finally, three modules were obtained from the HFpEF-DEP PPI network (**Figure 8A**), and five modules were identified from the HTN-DEP PPI network (**Figure 8B**).

## Overlapping Modules Between the Heart Failure With Preserved Ejection Fraction and Hypertension Networks

The overlapping modules between the HFpEF and HTN networks are shown in **Figure 9A**. The top five biological processes among the overlapping modules were cardiac muscle contraction (count 8, *P*-Value 5.59E-08), acute-phase response (count 7, *P*-Value 2.79E-07), regulation of the heart contraction force (count 6, *P*-Value 5.37E-07), muscle contraction (count 6, *P*-Value 2.89E-05), and response to lead ion (count 5, *P*-Value 1.72E-04). Extracellular exosome (count 62, *P*-Value 3.45E-18), mitochondrion (count 33, *P*-Value 1.10E-07), blood microparticle (count 9, *P*-Value 2.42E-06), extracellular space (count 27, *P*-Value 2.68E-06), and a band (count 5, *P*-Value 2.15E-05) related cell compositions were significantly enriched. Furthermore, the main enriched molecular functions were protein homodimerization activity (count 18, *P*-Value 6.42E-05), calcium ion binding (count 15, *P*-Value 4.86E-04), identical protein binding (count 14, *P*-Value 7.87E-04), actin filament binding (count 6, *P*-Value 0.003338211), and tropomyosin binding (count 3, *P*-Value 0.004429531). The enriched KEGG pathways included adrenergic signaling in cardiomyocytes (count 7, *P*-Value 0.001472787), prion diseases (count 4, *P*-Value 0.003172377), cardiac muscle contraction (count 5, *P*-Value 0.005270334), complement and coagulation cascades (count 4, *P*-Value 0.02702271), and adrenergic signaling in cardiomyocytes (count 4, *P*-Value 0.040014194) (**Figure 9B**). Finally, the main functional biological processes were myocardial contraction (30.77%), energy metabolism (15.38%), apoptosis (12.82%), oxidative stress (15.38%), immune response (10.26%), and cardiac hypertrophy (5.13%) (**Figure 9C**).

## Drug Discovery and Molecular Docking

Twelve small-molecule compounds were obtained from the DrugBank, including prednisolone acetate, bismuth subsalicylate, phenoxymethylpenicillin, polyethylene glycol, prednisolone, chloroform, salicylic acid, epinephrine, triptorelin, benzylpenicillin, propofol, and sulfadimethoxine. The binding affinity between the co-DEPs and the small-molecule compounds are shown in **Figure 10A**. For Hp, the docking score between epinephrine and Hp was the highest. Epinephrine generated hydrogen bonds with GLU314 and LEU334 of Hp (**Figure 10B**). For COQ9, epinephrine was again the best match. The ASN154, LEU219, ASN252, and GLU255 residues of COQ9 were suggested to be the binding sites of epinephrine (**Figure 10C**). Furthermore, Tf displayed the strongest binding affinity with sulfadimethoxine, potentially through hydrogen bonding at the LEU294, ARG124, and TYR188+ residues of Tf (**Figure 10D**). For Prnp, the docking score between chloroform and Prnp was the highest

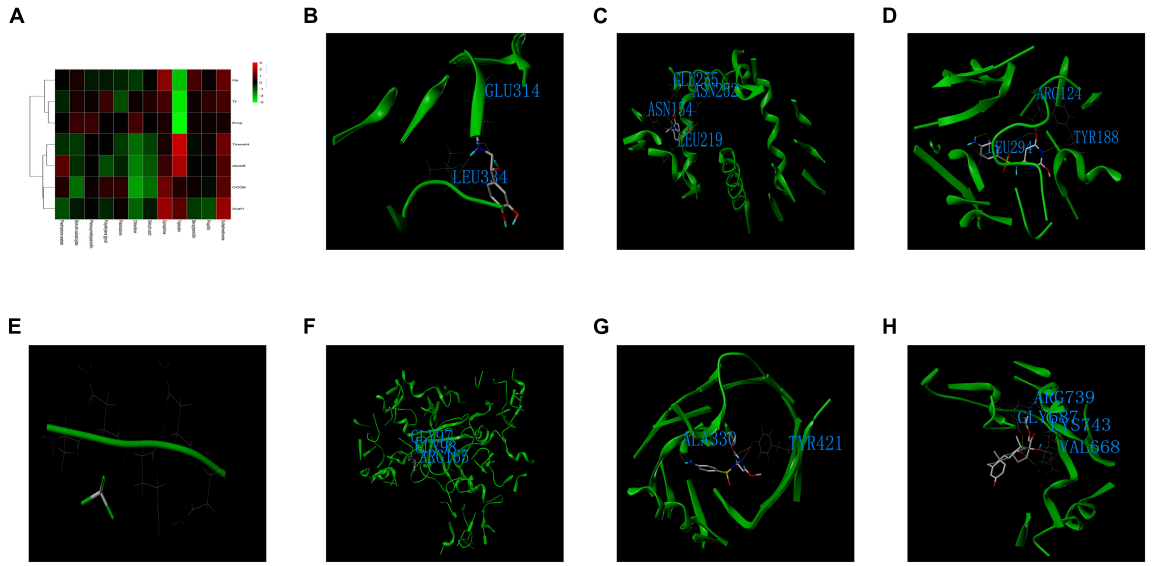



**FIGURE 10 |** Drug discovery and molecular docking. **(A)** Heatmap of the binding scores between the small-molecule compounds and co-DEPs. **(B)** The binding sites and interactions between epinephrine and Hp. **(C)** The binding sites and interactions between epinephrine and COQ9. **(D)** The binding sites and interactions between sulfadimethoxine and Tf. **(E)** The binding sites and interactions between chloroform and Prnp. **(F)** The binding sites and interactions between epinephrine and Acta1. **(G)** The binding sites and interactions between sulfadimethoxine and Timm44. **(H)** The binding sites and interactions between prednisolone acetate and Abcb6.

(**Figure 10E**). Epinephrine was the best match for Acta1, with potential binding at the GLY97, GLY98, and ARG165 residues of Acta1 (**Figure 10F**). The binding affinity between sulfadimethoxine and Timm44 was the strongest, with potential hydrogen bonding at the ALA330 and TYR421 residues of Timm44 (**Figure 10G**). Finally, prednisolone acetate was best matched with Abcb6, and ARG739, GLY687, LYS743, and VAL668 of Abcb6 were the potential targets of prednisolone acetate (**Figure 10H**).

## DISCUSSION

Heart failure with preserved ejection fraction is a complex syndrome that includes many types of clinical phenotypes. Huge pathophysiological differences exist among patients with different clinical HFpEF phenotypes, and no treatment strategy is suitable for all patients with HFpEF (Ge, 2020). Exploring the underlying pathophysiological mechanisms of different types of HFpEF will aid the discovery of personalized therapies and precision medicines for HFpEF treatment. Patients with HFpEF who are also diagnosed with HTN are considered to have vascular-related HFpEF, so exploring the functional connections between HFpEF and HTN will contribute to finding effective therapeutic targets for HFpEF and HTN treatment.

In this study, TMT-labeled quantitative proteomics was used to identify HFpEF- and HTN-related proteins. The functional links between HFpEF and HTN were analyzed at the network, module, and protein levels. Furthermore, molecular docking was used to determine precision medicine targets for HFpEF and HTN treatment. Seven co-DEPs were found among the HFpEF- and HTN-DEPs identified, including Hp, Tf, COQ9, Acat1, Timm44, Abcb6, and Prnp. Notably, Hp levels are closely related to hypertension and heart failure (Schröcksnadel, 1990; Lu et al., 2019; Rodrigues et al., 2019). Moreover, clinical studies have shown that inflammation plays an important role in the transition from HTN to HFpEF (Quaye, 2008), and that Hp is an indicator of inflammation in cardiovascular diseases (Szelényi et al., 2015). This suggests that Hp could be used to diagnose HTN and HFpEF and that Hp could be an effective therapeutic target for HTN and HFpEF. COQ9 is involved in the basic functions of mitochondria (Ferko et al., 2015), and the impairment of mitochondrial function is a common pathophysiological mechanism underlying both HTN and HFpEF (He et al., 2019; Zeng and Chen, 2019). Thus, COQ9 is also a potential biomarker for HTN and HFpEF. Several studies have shown that the expression of Tf and Prnp is altered in many cardiovascular diseases and that Tf and Prnp may be novel biomarkers for HTN and HFpEF (Gao et al., 2013; Rahim et al., 2018; Roura et al., 2018; Pang et al., 2020). Acta1 is involved in the pathological progression of myocardial remodeling (Pagano et al., 2017; Cañes et al., 2020), which is closely related to the prognosis of HTN and HFpEF (Fortuño et al., 2001; Georgiopoulou et al., 2010; Heinzel et al., 2015). Abcb6 and Timm44 are involved in mitochondrial functions and are potential biomarkers for HTN and HFpEF (Boswell-Casteel et al., 2017; Gao et al., 2020). Analysis using the CTD database indicated that there were strong connections between the co-DEPs and cardiovascular diseases in humans, including HF and HTN. Additionally, HFpEF was the most common type of HF, suggesting that

the co-DEPs may be effective therapeutic targets for HFpEF and HTN treatment.

Fever and anemia were found to be important co-DEP-related MM symptoms, which is consistent with previous findings (Bocchi et al., 2013; Tanimura et al., 2015; Burns et al., 2018). The common biological processes of HFpEF and HTN were closely related to energy metabolism. Previous studies have also indicated that mitochondrial oxidative capacity plays an important role in both HFpEF and HTN (Gueugneau et al., 2016; De Jong and Lopaschuk, 2017).

Heart failure with preserved ejection fraction and HTN shared eight overlapping modules, and the main biological functions enriched in these modules were myocardial contraction, energy metabolism, apoptosis, oxidative stress, immune response, and cardiac hypertrophy. We also found that post-translational modification and regulation of actin filaments could play an important role in HFpEF and HTN. Previous research has shown that phosphorylation of cardiac myosin-binding protein-C influences the progress of cross-bridge detachment and that deficient phosphorylation leads to diastolic dysfunction (Rosas et al., 2015). Furthermore, fibroblasts with abnormal proliferation contribute to the progression of HTN to HFpEF (Oatmen et al., 2020). Further research into the biological process of barbed-end actin filament capping may provide new insights.

The study also suggested that chloroform, epinephrine, sulfadimethoxine, and prednisolone acetate could be effective drugs for treating HTN and HFpEF. In addition, epinephrine, sulfadimethoxine, and prednisolone acetate have been widely used in many clinical diseases, but chloroform was not approved drug for human use. It suggested that chloroform maybe an candidate drugs for HTN-HFpEF. Previous studies have shown that gut microbiota dysfunction is closely related to the development of HTN and HFpEF (Hsu et al., 2020; Pakhomov and Baugh, 2020), and some antibiotics that regulate the gut microbiota have shown beneficial effects against HTN and other heart diseases (Chen et al., 2020; Du et al., 2020; Wu et al., 2020). A previous study showed that sulfadimethoxine could also regulate the gut microbiota (Mourand et al., 2014) and contribute to the normalization of blood pressure. Notably, chloroform injection can decrease the mean blood pressure (Loyke, 1971). Furthermore, prednisolone can prevent post-transplantation hypertension in rat renal allograft recipients (de Keijzer et al., 1987), indicating that prednisolone may be an effective drug for treating HTN and HFpEF. For patients with mild essential hypertension, intravenous infusion of small amounts of epinephrine has shown beneficial effects on hemodynamics, renal electrolyte excretion, and blood platelets (Kjeldsen et al., 1988). This indicates that epinephrine may be an effective drug for treating HTN and HFpEF. Most treatment strategies for HFpEF are empiric and are greatly influenced by expert consensus. In addition, some treatment strategies showed beneficial effects in patients with HFpEF, including the use of diuretics to control hypervolemia, treatment with mineralocorticoid antagonists, exercise therapies, and classical treatments for comorbidities. The results of this study, which are based on molecular docking and bioinformatics analyses, indicated that chloroform, epinephrine, sulfadimethoxine, and prednisolone acetate could be effective medicines for HTN and HFpEF. These drugs could be used to treat HTN and HFpEF, to reduce the occurrence of HFpEF in patients with HTN, or as personalized medicines for patients with HFpEF. Further animal experiments and small-scale clinical trials are needed to elucidate the functions and effects of these drugs in HTN and HFpEF. Nevertheless, it is important to note that chloroform is currently not approved for human use.

According to the enrichment results of the overlapping modules, myocardial contraction was the most important biological function shared between HFpEF and HTN. HTN influences the structure and function of the heart, suppresses myocardial contractions, and increases the prevalence of HFpEF (Chirinos et al., 2017). Previous studies have also noted the importance of myocardial contraction, as impaired diastolic function is a common phenotype of HFpEF and HTN. Here, the co-DEPs Acat1, Tf, and Hp were associated with myocardial contraction. Acta1 is involved in skeletal muscle thin filament assembly, which influences the contractile force of the heart (Winter et al., 2016). Tf and Hp are related to the response to lead ion, and result in myocardial contraction-related neurotoxic effects (Pappas et al., 2015). Previous studies have indicated that epinephrine treatment enhances myocardial contraction, but the effects of sulfadimethoxine on myocardial contraction remain unclear (Paur et al., 2012).

Here, we found that energy metabolism is closely related to HFpEF and HTN. These findings are consistent with earlier observations (Baltatu et al., 2017; De Jong and Lopaschuk, 2017). Acta1, Timm44, and Abcb6 are involved in fatty acid beta-oxidation and in the biological functions of energy metabolism. In animal experiments, fatty acid beta-oxidation is associated with the severity of myocardial fibrosis. Additionally, it is associated with a risk for HFpEF. Epinephrine, sulfadimethoxine, and prednisolone also have beneficial effects on energy metabolism (Park et al., 2001; Laskewitz et al., 2010; Wang et al., 2019).

Furthermore, apoptosis was found to be an important biological function in HTN and HFpEF. Activation of apoptosis can lead to cardiac dysfunction (Ekhterae et al., 1999), and inhibition of apoptosis can improve heart function and lead to beneficial effects in HFpEF and HTN therapies (Liu et al., 2018; Chen et al., 2019). Hp and Tf were involved in the response to hypoxia, which promotes cardiomyocyte apoptosis. Previous studies have indicated that Prnp is associated with the negative regulation of apoptosis in other diseases (Gao et al., 2019). As chloroform was best matched with Hp, Tf, and Prnp, further studies exploring the anti-apoptotic effect of chloroform in HFpEF and HTN treatment are needed.

In our study, oxidative stress was important in both HFpEF and HTN. Myocardial fibrosis, the major factor leading to myocardial remodeling, was found to be a

common pathological mechanism among HFpEF and HTN. Previous studies using an animal model of HFpEF and HTN have confirmed that the regulation of oxidative stress contributes to the inhibition of myocardial fibrosis (Wu et al., 2016; van der Pol et al., 2018). Similarly, other studies showed that Hp and COQ9 are involved in oxidative stress, including cellular oxidant detoxification and negative regulation of oxidoreductase activity (Swain et al., 2020; Yoshida et al., 2020). However, the effect of epinephrine on oxidoreductase activity in HFpEF and HTN still needs to be explored.

Immune responses are activated in both HFpEF and HTN (Carnevale and Wenzel, 2018; Michels da Silva et al., 2019). Although a treatment strategy targeting the immune response achieved some positive results in HTN, no obvious beneficial effects were observed in HFpEF (Michels da Silva et al., 2019; Zhao et al., 2019). Previous studies have shown that aging influences the immune response in HFpEF and HTN, and here, we found that Hp was associated with aging (De la Fuente et al., 2005; Forman and Goodpaster, 2018). Furthermore, Prnp was associated with the negative regulation of the T cell receptor signaling pathway, which is known to influence the immune response (Wong et al., 2017). Thus, chloroform may be an effective drug for targeting the immune response in HFpEF and HTN treatment.

The results showed that cardiac hypertrophy, which is associated with diastolic function, was significantly associated with HFpEF and HTN (Schmieder, 1990). Angiotensin II receptor blockers (ARBs) have been used in clinical trials for the treatment of HTN, as they not only reduce blood pressure but also have beneficial effects on cardiac hypertrophy, diastolic function, and renal function (Israili, 2000). ARBs also affect the blood pressure of patients with HFpEF, but do not have significant effects on echocardiographic parameters, 6-min walk test distances, or brain natriuretic peptide levels (Parthasarathy et al., 2009). Previous studies have shown that epinephrine can also suppress cardiac hypertrophy. Further research is necessary to determine whether chloroform and prednisolone can induce similar beneficial effects in HFpEF and HTN.

## CONCLUSION

Seven co-DEPs were observed between the HFpEF-DEPs and HTN-DEPs, including Hp, Tf, COQ9, Acat1, Timm44, Abcb6, and Prnp. These co-DEPs were closely related to the main functional similarities of HFpEF and HTN, including myocardial contraction, energy metabolism, apoptosis, oxidative stress, immune response, and cardiac hypertrophy. These co-DEPs may serve as biomarkers and drug targets for HFpEF and HTN. Furthermore, epinephrine, sulfadimethoxine, chloroform, and prednisolone acetate may serve as precision medicines for the treatment of HTN and HFpEF. Our study provides several targets for of the development of personalized therapies and precision medicines to treat HFpEF and other comorbidities.

## LIMITATIONS

There are some limitations to this study. Proteins with low expression levels or those showing insignificant changes could have been ignored in the analyses. Furthermore, these results need to be validated through fundamental research and clinical trials. Further animal experiments will help to explore the function of these drugs in HTN and HFpEF, and small-scale clinical trials will contribute to identifying whether these drugs have similar effects in patients with HTN and those with HFpEF.

## DATA AVAILABILITY STATEMENT

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

## ETHICS STATEMENT

The animal study was reviewed and approved by the Animal Ethics Committee of Shandong University of Traditional Chinese Medicine.

## AUTHOR CONTRIBUTIONS

GZ, JC, and CW conceived the study, acquired the data, and wrote the manuscript. PJ designed the experiments and interpreted the data. YW and YZ performed the experiments and statistical analysis. YJ and XL designed the study and revised the manuscript. All authors read and approved the final manuscript.

## FUNDING

This work has been supported by the National Natural Science Foundation of China (No. 81673970) and the Construction Project of National TCM Clinical Research Base for Hypertension [Guo Zhong Yi Yao Fa (2008) No. 23].

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fphys.2021.607089/full#supplementary-material

## REFERENCES


Baltatu, O. C., Amaral, F. G., Campos, L. A., and Cipolla-Neto, J. (2017). Melatonin, mitochondria and hypertension. *Cell. Mol. Life Sci.* 74, 3955–3964. doi: 10.1007-/s00018-017-2613-y

Bocchi, E. A., Arias, A., Verdejo, H., Diez, M., Gómez, E., and Castro, P. (2013). The reality of heart failure in Latin America. *J. Am. Coll. Cardiol.* 62, 949–958. doi: 10.1016/j.jacc.2013.06.013

Borlaug, B. A. (2020). Evaluation and management of heart failure with preserved ejection fraction. *Nat. Rev. Cardiol.* 17, 559–573. doi: 10.1038/s41569-020-0363-2

Boswell-Casteel, R. C., Fukuda, Y., and Schuetz, J. D. (2017). ABCB6, an ABC Transporter Impacting Drug Response and Disease. *Aaps J.* 20:8. doi: 10.1208/s12248-017-0165-6

Burns, J. A., Sanchez, C., Beussink, L., Daruwalla, V., Freed, B. H., Selvaraj, S., et al. (2018). Lack of Association Between Anemia and Intrinsic Left Ventricular Diastolic Function or Cardiac Mechanics in Heart Failure With Preserved Ejection Fraction. *Am. J. Cardiol.* 122, 1359–1365. doi: 10.1016/j.amjcard.2018.06.045

Cañes, L., Martí-Pàmies, I., Ballester-Servera, C., Herraiz-Martínez, A., Alonso, J., Galán, M., et al. (2020). Neuron-derived orphan receptor-1 modulates cardiac gene expression and exacerbates angiotensin II-induced cardiac hypertrophy. *Clin. Sci.* 134, 359–377. doi: 10.1042/cs20191014

Carnevale, D., and Wenzel, P. (2018). Mechanical stretch on endothelial cells interconnects innate and adaptive immune response in hypertension. *Cardiovasc. Res.* 114, 1432–1434. doi: 10.1093/cvr/cvy148

Chen, Y. P., Sivalingam, K., Shibu, M. A., Peramaiyan, R., Day, C. H., Shen, C. Y., et al. (2019). Protective effect of Fisetin against angiotensin II-induced apoptosis by activation of IGF-IR-PI3K-Akt signaling in H9c2 cells and spontaneous hypertension rats. *Phytomedicine* 57, 1–8. doi: 10.1016/j.phymed.2018.09.179

Chen, Y., Zhu, Y., Wu, C., Lu, A., Deng, M., Yu, H., et al. (2020). Gut dysbiosis contributes to high fructose-induced salt-sensitive hypertension in Sprague-Dawley rats. *Nutrition* 7:110766. doi: 10.1016/j.nut.2020.110766

Chirinos, J. A., Phan, T. S., Syed, A. A., Hashmath, Z., Oldland, H. G., Koppula, M. R., et al. (2017). Late Systolic Myocardial Loading Is Associated With Left Atrial Dysfunction in Hypertension. *Circ. Cardiovasc. Imaging* 10:e006023. doi: 10.1161/circimaging.116.006023

Cho, J. H., Zhang, R., Kilfoil, P. J., Gallet, R., de Couto, G., Bresee, C., et al. (2017). Delayed Repolarization Underlies Ventricular Arrhythmias in Rats With Heart Failure and Preserved Ejection Fraction. *Circulation* 136, 2037–2050. doi: 10.1161/circulationaha.117.028202

De Jong, K. A., and Lopaschuk, G. D. (2017). Complex Energy Metabolic Changes in Heart Failure With Preserved Ejection Fraction and Heart Failure With Reduced Ejection Fraction. *Can. J. Cardiol.* 33, 860–871. doi: 10.1016/j.cjca.2017.03.009

de Keijzer, M. H., Provoost, A. P., Van Aken, M., Wolff, E. D., and Molenaar, J. C. (1987). Prednisolone and posttransplantation hypertension in rat renal allograft recipients. *Transplantation* 43, 353–357. doi: 10.1097/00007890-198703000-00007

De la Fuente, M., Hernanz, A., and Vallejo, M. C. (2005). The immune system in the oxidative stress conditions of aging and hypertension: favorable effects of antioxidants and physical exercise. *Antioxid. Redox. Signal.* 7, 1356–1366. doi: 10.1089/ars.2005.7.1356

Dean, J. L., Zhao, Q. J., Lambert, J. C., Hawkins, B. S., Thomas, R. S., and Wesselkamper, S. C. (2017). Editor's Highlight: application of Gene Set Enrichment Analysis for Identification of Chemically Induced, Biologically Relevant Transcriptomic Networks and Potential Utilization in Human Health Risk Assessment. *Toxicol. Sci.* 157, 85–99. doi: 10.1093/toxsci/kfx021

Drazner, M. H. (2011). The progression of hypertensive heart disease. *Circulation* 123, 327–334. doi: 10.1161/circulationaha.108.845792

Du, Z., Wang, J., Lu, Y., Ma, X., Wen, R., Lin, J., et al. (2020). The cardiac protection of Baoyuan decoction via gut-heart axis metabolic pathway. *Phytomedicine* 79:153322. doi: 10.1016/j.phymed.2020.153322

Dunlay, S. M., Roger, V. L., and Redfield, M. M. (2017). Epidemiology of heart failure with preserved ejection fraction. *Nat. Rev. Cardiol.* 14, 591–602. doi: 10.1038/nrcardio.2017.65

Ekhterae, D., Lin, Z., Lundberg, M. S., Crow, M. T., Brosius, F. C. III, and Núñez, G. (1999). ARC inhibits cytochrome c release from mitochondria and protects against hypoxia-induced apoptosis in heart-derived H9c2 cells. *Circ. Res.* 85, e70–77. doi: 10.1161/01.res.85.12.e70

Ferko, M., Kancirová, I., Jašová, M., Waczulíková, I., Ďarnická, S., Kucharská, J., et al. (2015). Participation of heart mitochondria in myocardial protection against ischemia/reperfusion injury: benefit effects of short-term adaptation processes. *Physiol. Res.* 64, S617–S625. doi: 10.33549/physiolres.933218

Forman, D. E., and Goodpaster, B. H. (2018). Weighty Matters in HFpEF and Aging. *JACC Heart Fail.* 6, 650–652. doi: 10.1016/j.jchf.2018.06.016

Fortuño, M. A., Ravassa, S., Fortuño, A., Zalba, G., and Díez, J. (2001). Cardiomyocyte apoptotic cell death in arterial hypertension: mechanisms and potential management. *Hypertension* 38, 1406–1412. doi: 10.1161/hy1201.099615

Gao, G., Xuan, C., Yang, Q., Liu, X. C., Liu, Z. G., and He, G. W. (2013). Identification of altered plasma proteins by proteomic study in valvular heart diseases and the potential clinical significance. *PLoS One* 8:e72111. doi: 10.1371/journal.pone.0072111

Gao, L. P., Xiao, K., Wu, Y. Z., Chen, D. D., Yang, X. H., Shi, Q., et al. (2020). Enhanced Mitophagy Activity in Prion-Infected Cultured Cells and Prion-Infected Experimental Mice via a Pink1/Parkin-Dependent Mitophagy Pathway. *ACS Chem. Neurosci.* 11, 814–829. doi: 10.1021/acschemneuro.0c00039

Gao, Z., Peng, M., Chen, L., Yang, X., Li, H., Shi, R., et al. (2019). Prion Protein Protects Cancer Cells against Endoplasmic Reticulum Stress Induced Apoptosis. *Virol. Sin.* 34, 222–234. doi: 10.1007/s12250-019-00107-2

Gazewood, J. D., and Turner, P. L. (2017). Heart Failure with Preserved Ejection Fraction: diagnosis and Management. *Am. Fam. Phys.* 96, 582–588.

Ge, J. (2020). Coding proposal on phenotyping heart failure with preserved ejection fraction: a practical tool for facilitating etiology-oriented therapy. *Cardiol. J.* 27, 97–98. doi: 10.5603/cj.2020.0023

Georgiopoulou, V. V., Kalogeropoulos, A. P., Raggi, P., and Butler, J. (2010). Prevention, diagnosis, and treatment of hypertensive heart disease. *Cardiol. Clin.* 28, 675–691. doi: 10.1016/j.ccl.2010.07.005

Graziani, F., Varone, F., Crea, F., and Richeldi, L. (2018). Treating heart failure with preserved ejection fraction: learning from pulmonary fibrosis. *Eur. J. Heart Fail.* 20, 1385–1391. doi: 10.1002/ejhf.1286

Gueugneau, M., Coudy-Gandilhon, C., Meunier, B., Combaret, L., Taillandier, D., Polge, C., et al. (2016). Lower skeletal muscle capillarization in hypertensive elderly men. *Exp. Gerontol.* 76, 80–88. doi: 10.1016/j.exger.2016.01.013

He, J., Liu, X., Su, C., Wu, F., Sun, J., Zhang, J., et al. (2019). Inhibition of Mitochondrial Oxidative Damage Improves Reendothelialization Capacity of Endothelial Progenitor Cells via SIRT3 (Sirtuin 3)-Enhanced SOD2 (Superoxide Dismutase 2) Deacetylation in Hypertension. *Arterioscler. Thromb. Vasc. Biol.* 39, 1682–1698. doi: 10.1161/atvbaha.119.312613

Heinzel, F. R., Hohendanner, F., Jin, G., Sedej, S., and Edelmann, F. (2015). Myocardial hypertrophy and its role in heart failure with preserved ejection fraction. *J. Appl. Physiol.* 119, 1233–1242. doi: 10.1152/japplphysiol.00374.2015

Hicklin, H. E., Gilbert, O. N., Ye, F., Brooks, J. E., and Upadhya, B. (2020). Hypertension as a Road to Treatment of Heart Failure with Preserved Ejection Fraction. *Curr. Hypertens Rep.* 22:82. doi: 10.1007/s11906-020-01093-7

Hsu, C. N., Yang, H. W., Hou, C. Y., Chang-Chien, G. P., Lin, S., Tan, Y. L., et al. (2020). Maternal Adenine-Induced Chronic Kidney Disease Programs Hypertension in Adult Male Rat Offspring: implications of Nitric Oxide and Gut Microbiome Derived Metabolites. *Int. J. Mol. Sci.* 21:7237. doi: 10.3390/ijms21197237

Israili, Z. H. (2000). Clinical pharmacokinetics of angiotensin II (AT1) receptor blockers in hypertension. *J. Hum. Hypertens* 14, S73–S86. doi: 10.1038/sj.jhh.1000991

Kjeldsen, S. E., Os, I., Westheim, A., Lande, K., Gjesdal, K., Hjermann, I., et al. (1988). Hyper-responsiveness to low-dose epinephrine infusion in mild essential hypertension. *J. Hypertens Suppl.* 6, S581–S583. doi: 10.1097/00004872-198812040-00182

Kjeldsen, S. E., von Lueder, T. G., Smiseth, O. A., Wachtell, K., Mistry, N., Westheim, A. S., et al. (2020). Medical Therapies for Heart Failure With Preserved Ejection Fraction. *Hypertension* 75, 23–32. doi: 10.1161/hypertensionaha.119.14057

Laskewitz, A. J., van Dijk, T. H., Bloks, V. W., Reijngoud, D. J., van Lierop, M. J., Dokter, W. H., et al. (2010). Chronic prednisolone treatment reduces

hepatic insulin sensitivity while perturbing the fed-to-fasting transition in mice. *Endocrinology* 151, 2171–2178. doi: 10.1210/en.2009-1374

Le, D. H., and Pham, V. H. (2017). HGPEC: a Cytoscape app for prediction of novel disease-gene and disease-disease associations and evidence collection based on a random walk on heterogeneous network. *BMC Syst. Biol.* 11:61. doi: 10.1186/s12918-017-0437-x

Liu, Q., Zhang, Y., Wang, P., Liu, J., Li, B., Yu, Y., et al. (2019). Deciphering the scalene association among type-2 diabetes mellitus, prostate cancer, and chronic myeloid leukemia via enrichment analysis of disease-gene network. *Cancer Med.* 8, 2268–2277. doi: 10.1002/cam4.1845

Liu, Y., Li, L. N., Guo, S., Zhao, X. Y., Liu, Y. Z., Liang, C., et al. (2018). Melatonin improves cardiac function in a mouse model of heart failure with preserved ejection fraction. *Redox Biol.* 18, 211–221. doi: 10.1016/j.redox.2018.07.007

Loyke, H. F. (1971). The effect of injected chloroform on renal hypertension. *Anesth. Analg.* 50, 825–828. doi: 10.1213/00000539-197150050-00025

Lu, D. Y., Lin, C. P., Wu, C. H., Cheng, T. M., and Pan, J. P. (2019). Plasma haptoglobin level can augment NT-proBNP to predict poor outcome in patients with severe acute decompensated heart failure. *J. Investig. Med.* 67, 20–27. doi: 10.1136/jim-2018-000710

Maitiabola, G., Tian, F., Sun, H., Zhang, L., Gao, X., Xue, B., et al. (2020). Proteome Characteristics of Liver Tissue from Patients with Parenteral Nutrition-Associated Liver Disease. *Nutr. Metab.* 17:43. doi: 10.1186/s12986-020-00453-z

Messerli, F. H., Rimoldi, S. F., and Bangalore, S. (2017). The Transition From Hypertension to Heart Failure: contemporary Update. *JACC Heart Fail.* 5, 543–551. doi: 10.1016/j.jchf.2017.04.012

Michels da Silva, D., Langer, H., and Graf, T. (2019). Inflammatory and Molecular Pathways in Heart Failure-Ischemia, HFpEF and Transthyretin Cardiac Amyloidosis. *Int. J. Mol. Sci.* 20:2322. doi: 10.3390/ijms20092322

Mourand, G., Jouy, E., Bougeard, S., Dheilly, A., Kérouanton, A., Zeitouni, S., et al. (2014). Experimental study of the impact of antimicrobial treatments on Campylobacter, Enterococcus and PCR-capillary electrophoresis single-strand conformation polymorphism profiles of the gut microbiota of chickens. *J. Med. Microbiol.* 63, 1552–1560. doi: 10.1099/jmm.0.074476-0

Nwabuo, C. C., and Vasan, R. S. (2020). Pathophysiology of Hypertensive Heart Disease: beyond Left Ventricular Hypertrophy. *Curr. Hypertens Rep.* 22:11. doi: 10.1007/s11906-020-1017-9

Oatmen, K. E., Cull, E., and Spinale, F. G. (2020). Heart failure as interstitial cancer: emergence of a malignant fibroblast phenotype. *Nat. Rev. Cardiol.* 17, 523–531. doi: 10.1038/s41569-019-0286-y

Pagano, F., Angelini, F., Castaldo, C., Picchio, V., Messina, E., Sciarretta, S., et al. (2017). Normal versus Pathological Cardiac Fibroblast-Derived Extracellular Matrix Differentially Modulates Cardiosphere-Derived Cell Paracrine Properties and Commitment. *Stem Cell. Int.* 2017:7396462. doi: 10.1155/2017/7396462

Pakhomov, N., and Baugh, J. A. (2020). The Role of Diet-Derived Short Chain Fatty Acids in Regulating Cardiac Pressure Overload. *Am. J. Physiol. Heart Circ. Physiol.* 320, H475–H486. doi: 10.1152/ajpheart.00573.2020

Pang, B., Hu, C., Wu, G., Zhang, Y., and Lin, G. (2020). Identification of Target Genes in Hypertension and Left Ventricular Remodeling. *Medicine* 99:e21195. doi: 10.1097/md.0000000000021195

Pappas, C. T., Mayfield, R. M., Henderson, C., Jamilpour, N., Cover, C., Hernandez, Z., et al. (2015). Knockout of Lmod2 results in shorter thin filaments followed by dilated cardiomyopathy and juvenile lethality. *Proc. Natl. Acad. Sci. U. S. A.* 112, 13573–13578. doi: 10.1073/pnas.1508273112

Park, W. S., Chang, Y. S., Chung, S. H., Seo, D. W., Hong, S. H., and Lee, M. (2001). Effect of hypothermia on bilirubin-induced alterations in brain cell membrane function and energy metabolism in newborn piglets. *Brain Res.* 922, 276–281. doi: 10.1016/s0006-8993(01)03186-9

Parthasarathy, H. K., Pieske, B., Weisskopf, M., Andrews, C. D., Brunel, P., Struthers, A. D., et al. (2009). A randomized, double-blind, placebo-controlled study to determine the effects of valsartan on exercise time in patients with symptomatic heart failure with preserved ejection fraction. *Eur. J. Heart Fail.* 11, 980–989. doi: 10.1093/eurjhf/hfp120

Paulus, W. J., and Tschöpe, C. (2013). A novel paradigm for heart failure with preserved ejection fraction: comorbidities drive myocardial dysfunction and remodeling through coronary microvascular endothelial inflammation. *J. Am. Coll. Cardiol.* 62, 263–271. doi: 10.1016/j.jacc.2013.02.092

Paur, H., Wright, P. T., Sikkel, M. B., Tranter, M. H., Mansfield, C., O'Gara, P., et al. (2012). High levels of circulating epinephrine trigger apical cardiodepression in a β2-adrenergic receptor/Gi-dependent manner: a new model of Takotsubo cardiomyopathy. *Circulation* 126, 697–706. doi: 10.1161/circulationaha.112.111591

Pieske, B., Tschöpe, C., de Boer, R. A., Fraser, A. G., Anker, S. D., Donal, E., et al. (2019). How to diagnose heart failure with preserved ejection fraction: the HFA-PEFF diagnostic algorithm: a consensus recommendation from the Heart Failure Association (HFA) of the European Society of Cardiology (ESC). *Eur. Heart J.* 40, 3297–3317. doi: 10.1093/eurheartj/ehz641

Quaye, I. K. (2008). Haptoglobin, inflammation and disease. *Trans. R. Soc. Trop. Med. Hyg.* 102, 735–742. doi: 10.1016/j.trstmh.2008.04.010

Rahim, M. A. A., Rahim, Z. H. A., Ahmad, W. A. W., Bakri, M. M., Ismail, M. D., and Hashim, O. H. (2018). Inverse changes in plasma tetranectin and titin levels in patients with type 2 diabetes mellitus: a potential predictor of acute myocardial infarction? *Acta Pharmacol. Sin.* 39, 1197–1207. doi: 10.1038/aps.2017.141

Redfield, M. M. (2016). Heart Failure with Preserved Ejection Fraction. *N. Engl. J. Med.* 375, 1868–1877. doi: 10.1056/NEJMcp1511175

Rodrigues, K. F., Pietrani, N. T., Carvalho, L. M. L., Bosco, A. A., Sandrim, V. C., Ferreira, C. N., et al. (2019). Haptoglobin levels are influenced by Hp1-Hp2 polymorphism, obesity, inflammation, and hypertension in type 2 diabetes mellitus. *Endocrinol. Diabetes Nutr.* 66, 99–107. doi: 10.1016/j.endinu.2018.07.008

Rosas, P. C., Liu, Y., Abdalla, M. I., Thomas, C. M., Kidwell, D. T., Dusio, G. F., et al. (2015). Phosphorylation of cardiac Myosin-binding protein-C is a critical mediator of diastolic function. *Circ. Heart Fail.* 8, 582–594. doi: 10.1161/circheartfailure.114.001550

Roura, S., Gámez-Valero, A., Lupón, J., Gálvez-Montón, C., Borràs, F. E., and Bayes-Genis, A. (2018). Proteomic signature of circulating extracellular vesicles in dilated cardiomyopathy. *Lab. Invest.* 98, 1291–1299. doi: 10.1038/s41374-018-0044-5

Schmieder, R. E. (1990). Risk reduction following regression of cardiac hypertrophy. *Clin Exp Hypertens A* 12, 903–916. doi: 10.3109/10641969009073508

Schröcksnadel, H. (1990). Haptoglobin and free haemoglobin in pregnancy-induced hypertension. *Lancet* 336:1594. doi: 10.1016/0140-6736(90)93383-z

Swain, N., Samanta, L., Agarwal, A., Kumar, S., Dixit, A., Gopalan, B., et al. (2020). Aberrant Upregulation of Compensatory Redox Molecular Machines May Contribute to Sperm Dysfunction in Infertile Men with Unilateral Varicocele: a Proteomic Insight. *Antioxid. Redox. Signal.* 32, 504–521. doi: 10.1089/ars.2019.7828

Szelényi, Z., Fazakas, Á, Szénási, G., Kiss, M., Tegze, N., Fekete, B. C., et al. (2015). Inflammation and oxidative stress caused by nitric oxide synthase uncoupling might lead to left ventricular diastolic and systolic dysfunction in patients with hypertension. *J. Geriatr. Cardiol.* 12, 1–10. doi: 10.11909/j.issn.1671-5411.2015.01.001

Tadic, M., Cuspidi, C., Frydas, A., and Grassi, G. (2018). The role of arterial hypertension in development heart failure with preserved ejection fraction: just a risk factor or something more? *Heart Fail. Rev.* 23, 631–639. doi: 10.1007/s10741-018-9698-8

Tan, Y., Zuo, W., Huang, L., Zhou, B., Liang, H., Zheng, S., et al. (2020). Nervilifordin F alleviates intestinal ischemia/reperfusion-induced acute lung injury via inhibiting inflammasome and mTOR pathway. *Int. Immunopharmacol.* 89:107014. doi: 10.1016/j.intimp.2020.107014

Tanimura, M., Dohi, K., Matsuda, M., Sato, Y., Sugiura, E., Kumagai, N., et al. (2015). Renal resistive index as an indicator of the presence and severity of anemia and its future development in patients with hypertension. *BMC Nephrol.* 16:45. doi: 10.1186/s12882-015-0040-6

van der Pol, A., Gil, A., Tromp, J., Silljé, H. H. W., van Veldhuisen, D. J., Voors, A. A., et al. (2018). OPLAH ablation leads to accumulation of 5-oxoproline, oxidative stress, fibrosis, and elevated fillings pressures: a murine model for heart failure with a preserved ejection fraction. *Cardiovasc. Res.* 114, 1871–1882. doi: 10.1093/cvr/cvy187

Wang, Q., Wang, C., Wang, B., Shen, Q., Qiu, L., Zou, S., et al. (2019). Identification of RyR2-PBmice and the effects of transposon insertional mutagenesis of the RyR2 gene on cardiac function in mice. *PeerJ.* 7:e6942. doi: 10.7717/peerj.6942

Winter, J. M., Joureau, B., Lee, E. J., Kiss, B., Yuen, M., Gupta, V. A., et al. (2016). Mutation-specific effects on thin filament length in thin filament myopathy. *Ann. Neurol.* 79, 959–969. doi: 10.1002/ana.24654

Wong, G. K., Heather, J. M., Barmettler, S., and Cobbold, M. (2017). Immune dysregulation in immunodeficiency disorders: the role of T-cell receptor sequencing. *J. Autoimmun.* 80, 1–9. doi: 10.1016/j.jaut.2017.04.002

Wu, H., Chen, L., Xie, J., Li, R., Li, G. N., Chen, Q. H., et al. (2016). Periostin expression induced by oxidative stress contributes to myocardial fibrosis in a rat model of high salt-induced hypertension. *Mol. Med. Rep.* 14, 776–782. doi: 10.3892/mmr.2016.5308

Wu, Q., Xu, Z., Song, S., Zhang, H., Zhang, W., Liu, L., et al. (2020). Gut microbiota modulates stress-induced hypertension through the HPA axis. *Brain Res. Bull.* 162, 49–58. doi: 10.1016/j.brainresbull.2020.05.014

Yan, F., Gao, M., Gong, Y., Zhang, L., Ai, N., Zhang, J., et al. (2020). Proteomic analysis of underlying apoptosis mechanisms of human retinal pigment epithelial ARPE-19 cells in response to mechanical stretch. *J. Cell. Physiol.* 235, 7604–7619. doi: 10.1002/jcp.29670

Yoshida, S., Kurajoh, M., Fukumoto, S., Murase, T., Nakamura, T., Yoshida, H., et al. (2020). Association of plasma xanthine oxidoreductase activity with blood pressure affected by oxidative stress level: medCity21 health examination registry. *Sci. Rep.* 10:4437. doi: 10.1038/s41598-020-61463-8

Yuan, Y., Zhang, Y., Zhang, X., Yu, Y., Li, B., Wang, P., et al. (2016). Deciphering the genetic and modular connections between coronary heart disease, idiopathic pulmonary arterial hypertension and pulmonary heart disease. *Mol. Med. Rep.* 14, 661–670. doi: 10.3892/mmr.2016.5298

Zeng, H., and Chen, J. X. (2019). Sirtuin 3, Endothelial Metabolic Reprogramming, and Heart Failure With Preserved Ejection Fraction. *J. Cardiovasc. Pharmacol.* 74, 315–323. doi: 10.1097/fjc.0000000000000719

Zhang, Y., Kong, P., Chen, Y., Yu, Y., Liu, J., Yang, L., et al. (2014). Significant overlapping modules and biological processes between stroke and coronary heart disease. *CNS Neurol. Disord. Drug Targets* 13, 652–660. doi: 10.2174/1871527312666131223115112

Zhao, T. V., Li, Y., Liu, X., Xia, S., Shi, P., Li, L., et al. (2019). ATP release drives heightened immune responses associated with hypertension. *Sci. Immunol.* 4:eaau6426. doi: 10.1126/sciimmunol.aau6426

**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: 31 January 2022
doi: 10.3389/fgene.2021.635904

# Hypoxia Induced Sex-Difference in Zebrafish Brain Proteome Profile Reveals the Crucial Role of H3K9me3 in Recovery From Acute Hypoxia

*Tapatee Das*[1,2], *Avijeet Kamle*[3], *Arvind Kumar*[2,3] *and Sumana Chakravarty*[1,2]*

[1]Applied Biology, CSIR-Indian Institute of Chemical Technology (IICT), Hyderabad, India, [2]Academy of Scientific and Innovative Research (AcSIR), Ghaziabad, India, [3]CSIR-Centre for Cellular and Molecular Biology (CCMB), Hyderabad, India

Understanding the molecular basis of sex differences in neural response to acute hypoxic insult has profound implications for the effective prevention and treatment of ischemic stroke. Global hypoxic-ischemic induced neural damage has been studied recently under well-controlled, non-invasive, reproducible conditions using a zebrafish model. Our earlier report on sex difference in global acute hypoxia-induced neural damage and recovery in zebrafish prompted us to conduct a comprehensive study on the mechanisms underlying the recovery. An omics approach for studying quantitative changes in brain proteome upon hypoxia insult following recovery was undertaken using iTRAQ-based LC-MS/MS approach. The results shed light on the altered expression of many regulatory proteins in the zebrafish brain upon acute hypoxia following recovery. The sex difference in differentially expressed proteins along with the proteins expressed in a uniform direction in both the sexes was studied. Core expression analysis by Ingenuity Pathway Analysis (IPA) showed a distinct sex difference in the disease function heatmap. Most of the upstream regulators obtained through IPA were validated at the transcriptional level. Translational upregulation of H3K9me3 in males led us to elucidate the mechanism of recovery by confirming transcriptional targets through ChIP-qPCR. The upregulation of H3K9me3 level in males at 4 h post-hypoxia appears to affect the early neurogenic markers nestin, klf4, and sox2, which might explain the late recovery in males, compared to females. Acute hypoxia-induced sex-specific comparison of brain proteome led us to reveal many differentially expressed proteins, which can be further studied for the development of novel targets for better therapeutic strategy.

**Keywords: sex difference, IPA, pathway analysis, iTRAQ, hypoxia-ischemia recovery**

## HIGHLIGHTS

- Sex disparity was observed in differentially regulated proteins; mostly downregulated in males.
- Five common transcription regulators [Myc, Mknk1, Nfe2l2 (Nrf2), Thrb, and Otx 2] have differential activation states.
- Upon CoIP, H3K9me3 targets of hypoxia were found to be totally different from normoxia.
- H3K9me3 seems to be a key player in early neurogenesis.
- Novel finding: H3K9me3 appeals to play an important role in the delayed recovery of males from acute hypoxia.

OPEN ACCESS

***Edited by:***
*Sanjay Kumar Banerjee, National Institute of Pharmaceutical Education and Research (Guwahati), India*

***Reviewed by:***
*Renu Goel, Translational Health Science and Technology Institute (THSTI), India*
*Parmeshwar Katare, University of Oslo, Norway*

***Correspondence:**
*Sumana Chakravarty*
*sumanachak@iict.res.in*

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

***Received:*** *30 November 2020*
***Accepted:*** *29 September 2021*
***Published:*** *31 January 2022*

***Citation:***
*Das T, Kamle A, Kumar A and Chakravarty S (2022) Hypoxia Induced Sex-Difference in Zebrafish Brain Proteome Profile Reveals the Crucial Role of H3K9me3 in Recovery From Acute Hypoxia. Front. Genet. 12:635904. doi: 10.3389/fgene.2021.635904*

## INTRODUCTION

Oxygenation in vertebrates is always a life-or-death necessity for any of the metabolic needs of cells and tissues (Sun, 1999). Over the last decade, we have acquired adequate information on cellular and molecular mechanisms in hypoxic-ischemic injury, survival, and death (Muller and Marks, 2014; Sekhon et al., 2017). Hypoxic-ischemic neural injury continues to be the leading cause of death and disability worldwide (Catherine and Collaborators, 2019). The degree of disability does not simply reflect the severity or distribution of the impaired blood supply (Dugan et al., 1999). The most common condition of hypoxia-ischemia leads to cerebral stroke due to the focal disruption of blood supply to a part of the brain. Other conditions include transient impairment of blood flow to the entire brain, termed global ischemia, which occurs following cardiac arrest.

A low level of oxygen and the brain's susceptibility to acute hypoxia characterizes the key factor determining critical dependency. Cerebral oxygenation is reduced in hypoxia and neuronal damage can occur during a prolonged mismatch between oxygen supply and demand (Goodall et al., 2014). All the neurons in the brain can sense and, crucially, modify, their activity in response to hypoxia. Most neurons respond to hypoxia by decreasing metabolic demand and thus the need for aerobic energy (Michiels, 2004). Deciphering cellular response to energetic challenges that occur on the onset of acute hypoxia may give insight into the ischemic condition in various diseases (Bickler and Buck, 2007). Broad high throughput approaches in global changes in protein expression allow uncovering the critical signals underlying mechanisms in the disease condition. Acute Hypoxia causes a significant perturbation in cellular energy homeostasis before a hypoxia sensing and signal transduction cascade needing energy demand initiates (Hochachka et al., 1996). An early component of the responses to acute hypoxia i.e., neural damage and recovery may have both post-transcriptional and translational mechanisms. The rapid response to acute hypoxia may preclude many pathways that require many new gene expressions suggesting the mechanism underlying recovery from acute hypoxia is mediated at least in part by the activities of the existing pool of mRNA and protein. An approach such as high throughput proteomic analysis is one of the ideally suited approaches to understand the neural changes induced by acute hypoxia with recovery (Li et al., 2019).

Previous proteomic studies have shown hypoxia-induced changes in the zebrafish (*Danio rerio*) skeletal muscle proteome (Chen et al., 2013) and have implicated a broad range of cellular functions in response to hypoxia. Another proteomics study on zebrafish brain upon chronic unpredictable stress (Chakravarty et al., 2013) has recently laid the groundwork for the analysis of neural proteome response to stressors.

A recent review article on Proteomics-Based Approaches for the Study of Ischemic Stroke (Li et al., 2019) discussed the proteomics study of ischemic stroke using *in vivo* and *in vitro* models, with and without interventions and taking tissue, cerebrospinal fluid, or plasma. Although proteomic studies have contributed with a long list of potential biomarkers for diagnosis, prognosis, and monitoring of ischemic stroke, most of these have not been implemented in clinical application successfully. The shortcomings from the existing proteomics data are small sample size, cell types, te age of experimental animals, and using single-sex experimental animals all seem to be responsible for blocking these results from achieving clinical implementation.

Like many neurological disorders, cerebral stroke is reported to have sex-specific differences in occurrence and mechanisms. However, the molecular details underlying these sex-specific differences have not yet been explored using a relevant animal model. In fact, many factors including genetics, hormones (estrogen and androgen), epigenetic regulation, and environment contribute to sex-specific differences. Since ischemic sensitivity varies over the lifespan, and the "ischemia resistant" female phenotype diminishes after menopause, hence the role of sex hormones cannot be ruled out. To understand the role of hormonal status on the cerebral vasculature in pinpointing sex-specific differences in stroke pathophysiology, a suitable, simple animal model that can help to address these complicated sex-specific differences is warranted.

Sex-specific differences in the hypoxic-ischemic brain have profound implications for effective prevention and treatment. Global hypoxic-ischemic damages and recovery are well studied under the well-controlled, non-invasive, reproducible conditions in zebrafish (Yu and Li, 2013; Braga et al., 2016; Silva et al., 2016; Das et al., 2019). In our previous study, we have reported the sex-specific difference in hypoxia-induced neural damage and recovery, where we have concluded that as compared to males, females showed a higher level of neural damage and an ability to recover faster. This interesting finding led us to explore the global proteome changes induced in recovery after the hypoxic stress, so in the present study, we performed a high throughput proteomic analysis on zebrafish brain by iTRAQ method. The iTRAQ labeling method also allows the identification of different post-translational modifications which are key to understand the aetiology and develop better treatment.

## EXPERIMENTAL PROCEDURE

### Animal Procurement and Acute-Hypoxia Treatment

Wild type strain of zebrafish was bred and raised at CSIR-IICT zebrafish facility in accordance with protocol no IICT/CB/SC/281114/30 under registration no# 97/1999/CPCSEA. All the experimental animals were maintained in a controlled environment with a 14 h light/10 h dark cycle at 28°C with three feedings and constant aeration. Zebrafish aged 5–6 month were segregated on the basis of sex and used for all the experiments. For an acute hypoxia treatment animals were placed in an air-tight glass hypoxia chamber for a period of 5 min with 0.6 mg/ltr dissolved oxygen following reoxygenation at 7 mg/ltr dissolved oxygen in a recovery tank, which is exactly described in (Das et al., 2019). After 4 h post-hypoxia, all the animals were sacrificed for brain tissue collection.

## Protein Extraction for iTRAQ

The animals were euthanatized and decapitated to remove the brain. The whole brain from each animal was homogenized in a lysis buffer [50 mM ammonium bicarbonate pH 8.0, 0.1% SDS with protease inhibitor cocktail (Sigma)] and for further efficient disruption and homogenization of tissue, a mild sonication was done using Bioruptor®. The obtained lysates were cold-centrifuged at 14,000 rpm for 15 min and the supernatant was quantified using Bradford assay with BSA as standard. Further protein samples were cleaned up by acetone precipitation. For each group, 80 μg of protein was taken and six volumes of chilled acetone were added for precipitation. After decantation of acetone the samples were resuspended in dissolution buffer (Buffer pH is 8.5. Contains 0.5 M triethyammonium bicarbonate) provided with the iTRAQ® Reagents-4plex Applications kit-Protein (AB Sciex). Before trypsin digestion, all the protein samples were reduced and cysteine blocked using the reagents provided in the iTRAQ® Reagents-4plex Applications kit-Protein (AB Sciex). Digestion and labeling of proteins were done according to the manufacturer's protocol. The samples from normoxia male and female were labeled with reagents 114 and 116 and the samples from hypoxia male and female were labeled with reagents 115 and 117, respectively. Subsequently, all the labeled samples were pooled and vacuum dried, and further cleaned up using the C18 desalting column (Thermo Fisher Scientific). The final fraction was concentrated using a vacuum concentrator and reconstituted in 10 μl of 0.1% formic acid for LC-MS/MS analysis.

## LC-MS/MS Analysis

LC-MS/MS analysis of the trypsin digested iTRAQ labeled and purified fractions were performed in LTQ - Orbitrap Velos (Thermo Scientific, Germany). The fragmentation was carried out using higher-energy collision dissociation (HCD) with 50% normalized collision energy. The MS data were analyzed using Proteome Discoverer (Thermo Fisher Scientific, Version 1.4). MS/MS search was carried out using the SEQUEST search engine against the NCBI zebrafish protein database. Search parameters included trypsin as an enzyme with a maximum of two missed cleavage allowed; precursor and fragment mass tolerance were set to 10 ppm and 0.2 Da respectively; Methionine oxidation was set as a dynamic modification while methylthio modification at cysteine and iTRAQ modification at N-terminus of the peptide were set as static modifications. The FDR was calculated by enabling the peptide sequence analysis using a decoy database. High confidence peptide identifications were obtained by setting a target FDR threshold of 1% at the peptide level. Relative quantitation of proteins was determined based on the ratios of relative intensities of the reporter ions from hypoxia treated and untreated samples released during MS/MS fragmentation of each peptide. Appropriate quality control filters at the level of peptides/peptide spectral matches (PSMs) and then at the protein level were applied to the iTRAQ data. Proteins identified from the triplicate runs as having more than 1.5-log-fold changes in the hypoxia samples against the normoxia samples were selected for upregulation and having less than 0.5-log fold change considered to be downregulated for its differential expression. Proteins based on their regulation were analyzed for putative associations in different network pathways.

The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE partner repository with the dataset identifier PXD027528".

## Protein Enrichment Analysis

To perform the functional enrichment tests of the candidate proteins, we used Ingenuity Pathway Analysis (IPA) software for both canonical pathways and molecular networks altered. The IPA system provides a more comprehensive pathway resource based on manual collection and curation. The rich information returned by IPA is also suitable for pathway crosstalk analysis as it has more molecules and their connections included. For analysis, we have provided the identified peptides with relative and absolute expression fold change values and performed core IPA analysis, biomarkers, and molecular and functional comparison analysis.

## Co-Immunoprecipitation

Zebrafish brain tissue was homogenized in nuclear extraction buffer [50 mM HEPES (pH 7.8), 50 mM KCl, 300 mM NaCl, 0.1 M EDTA, 1 mM DTT, 10% (v/v) Glycerol and 1X protease inhibitor] and further washed with PBS and incubated in RIPA buffer [20 mM Tris (pH 7.5), 150 mM NaCl, 1% NP-40, 5 mM EDTA, protease and phosphatase inhibitors] for 15 min on ice. After centrifugation, the supernatant was collected and pre-cleared with protein A agarose beads (Santa Cruz) at 4°C for 30 min. The pre-cleared lysate was then incubated with Anti-Histone H3 (tri methyl K9) antibody (H3K9me3) (AB8898 1:250) complexed to protein A beads at 4°C for 5–6 h, followed by washes with a buffer containing 10 mM Tris (pH 7.5), 150 mM NaCl, and 1 mM EDTA. The beads complexed with the immunoprecipitated proteins were then boiled at 100°C in 3X Laemmli buffer for 5 min. 2.5% of whole tissue lysate was taken as input for each immunoprecipitation. Western blotting was carried out by loading equal amounts of the immunoprecipitated proteins.

## Immunoblotting Analysis

For immunoblotting experiments, cells were lysed in 3X Laemmli buffer [180 mM Tris (pH 6.8), 6% SDS, 15% glycerol, 7.5% β-mercaptoethanol, and 0.01% bromophenol blue]. Images were captured using Chemicapt (Vilber-Lourmat, Germany). Densitometry analysis for blots was performed using ImageJ software (NIH) and images were processed in Adobe Photoshop CS3. The intensity values plotted or mentioned are average values from the number of biological replicates indicated in the legend.

## Chromatin Immunoprecipitation Assay

ChIP was performed as described in (Weidemann et al., 2013) with required minor modifications. Briefly, for each ChIP, cross-linked samples from three animals were pooled together both in the normoxia and hypoxia groups. The 30 μg of chromatin from each sample was pre-cleared with Dynabeads (Invitrogen) before incubation with an anti-rabbit

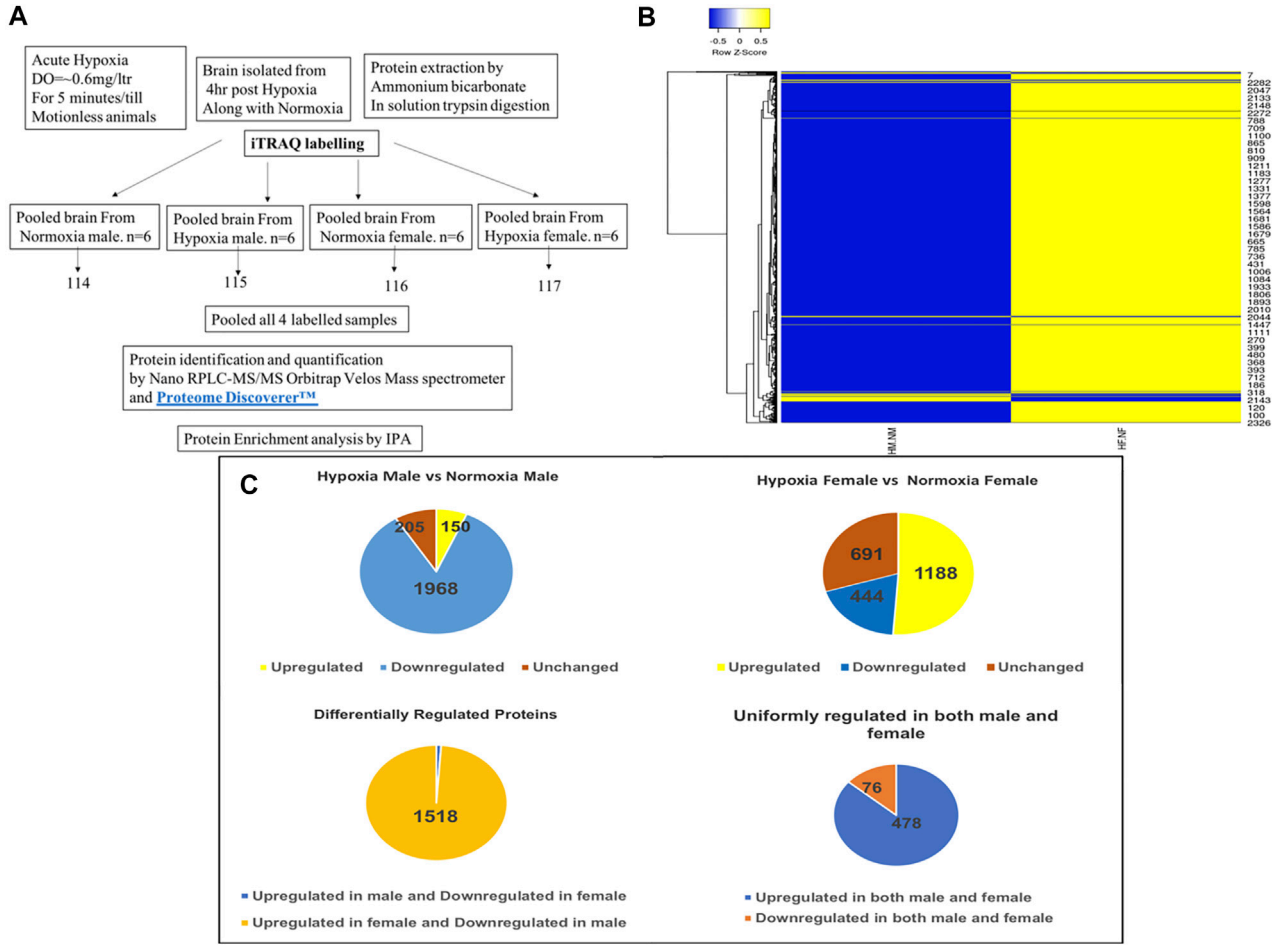



**FIGURE 1 |** Analysis of zebrafish brain proteome by iTRAQ. Schematic representation of brain proteome analysis by iTRAQ labeling **(A)**, Cluster heatmap of proteins obtained from iTRAQ analysis for hypoxia male [HM] vs. normoxia male [NM] and hypoxia female [HF] vs. normoxia female [NF] **(B)**, Pie chart showing the analysis of expression of proteins resulted from iTRAQ in hypoxia male and female brain compared with normoxia male and female brain **(C)**.

H3K9me3 antibody (EPR16601) keeping non-immune rabbit IgG antibody as a negative control. After reverse cross-linking and sequential washes with different concentration salt buffers, DNA was purified using phenol-chloroform-isoamyl alcohol (25:24:1 ratio, SIGMA). Specific primers for the gene-specific 5′ upstream region of the transcription start site were used for quantifying the enrichment of the histone mark H3K9me3, for 10% input, in SYBR Green-based qPCR assays.

## qPCR

The total RNA was isolated using TRIzol Reagent as per the manufacturer's instruction. The cDNA was synthesized employing RevertAid H Minus First Strand cDNA Synthesis Kit according to the manufacturer's protocol. The primer sequences are available on a request basis. Real-time PCR was performed in triplicate using SYBR Green PCR Master Mix Detection System (Applied Biosystems). Normalization of mRNA expression levels was carried out using β-actin as the housekeeping gene. Gene expression was normalized against the ubiquitously expressed beta actin gene. Data were analyzed using the Δ(ΔCT) method.

## Statistical Analysis

Statistical analysis was performed using Microsoft Excel. Mean differences between the normoxia and hypoxia groups were determined using a two-tailed unpaired Student's *t*-test with confidence intervals of 95% since only two groups were used to compare a single variable i.e., normoxia/hypoxia. A $p$-value of $\leq 0.05$ was considered significant.

# RESULTS AND DISCUSSION

## Analysis of Zebrafish Brain Proteome Induced by Acute Hypoxia and During Recovery Using iTRAQ Based LC-MS/MS

As described previously by us (Das et al., 2019), 4 h post-acute hypoxia treatment (for 5 min at DO = ±0.6 mg/litre) the brain tissues from both hypoxia-treated and untreated male and female zebrafish were isolated ($n$ = 6 per group) and subjected to comprehensive proteomic profiling. For this, male and female

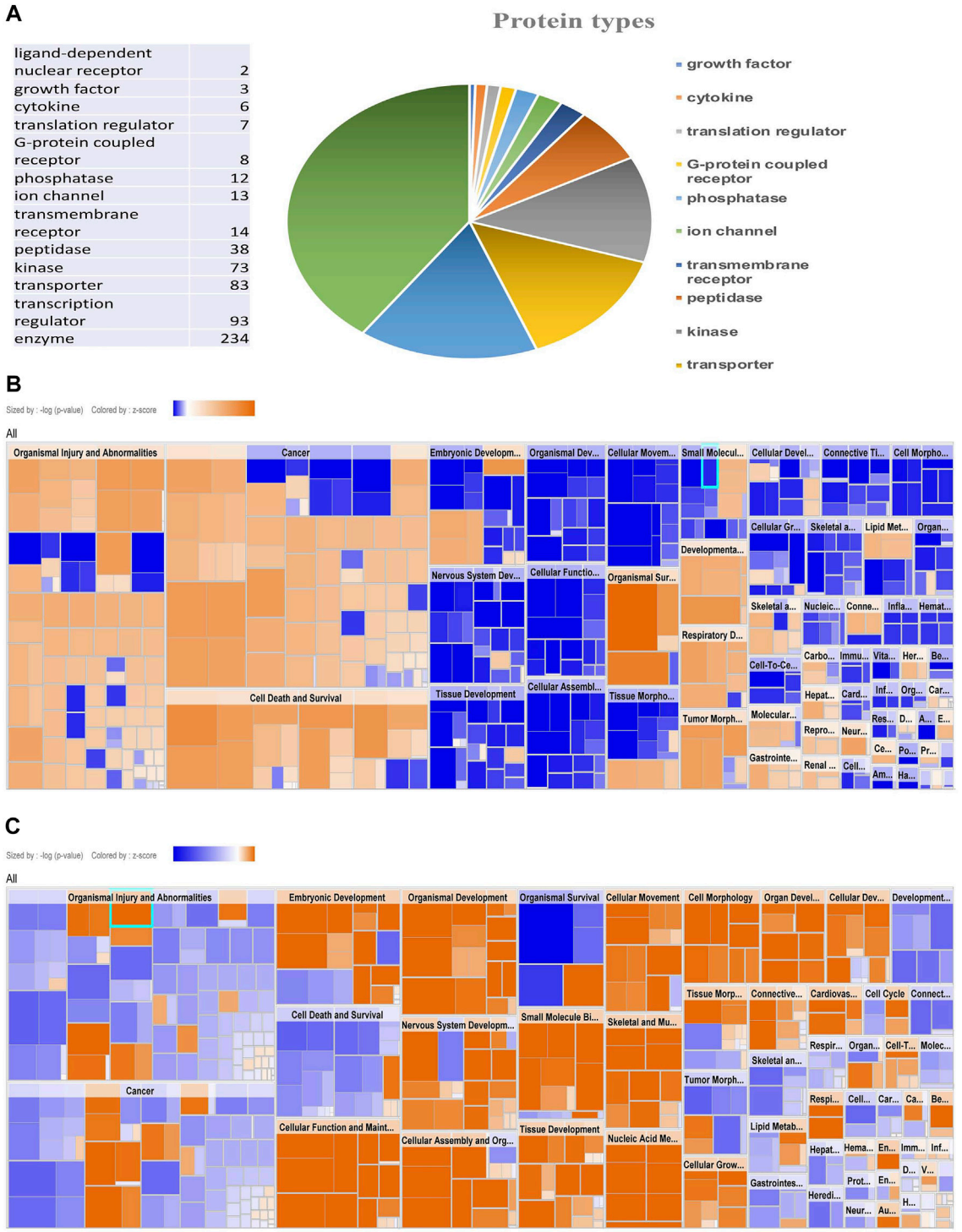



**FIGURE 2 |** Protein enrichment analysis by IPA showing types of protein and Comparative heat map. Pie chart showing types of protein mapped by IPA **(A)**, Disease function Heat map of male **(B)** and female **(C)** zebrafish brain proteome upon hypoxia treatment.

zebrafish brains of nine individuals from each sex were pooled together for iTRAQ based LC-MS/MS method and three technical replicates were run in LC-MS/MS, as depicted in **Figure 1A**. Differentially expressed proteins were identified using iTRAQ and LC-MS/MS analysis on an LTQ Orbitrap Velos mass spectrometer, by comparing hypoxia male (HM) vs. normoxia male (NM) and hypoxia female (HF) vs. normoxia female (NF) (**Figure 1A**). A total of 2,323 proteins

**TABLE 1 |** IPA generated disease function analysis for zebrafish male and female brain proteome induced by hypoxia following recovery.

**Male ZF brain proteome analysis (HM/NM)**

| Categories | Diseases or functions annotation | *p*-value | Predicted activation state | Activation z-score | # Molecules |
|---|---|---|---|---|---|
| Organismal survival | Organismal death | 2.31E-11 | Increased | 12.434 | 251 |
| Cell death and survival | Cell death | 2.51E-11 | Increased | 4.843 | 218 |
| Cell death and survival | Necrosis | 3.27E-09 | Increased | 4.131 | 154 |
| Cancer, organismal injury and abnormalities, respiratory disease | Development of lung tumor | 9.13E-08 | Increased | 3.302 | 27 |
| Cancer, organismal injury and abnormalities | Incidence of tumor | 0.000000893 | Increased | 2.199 | 87 |
| Cell death and survival | Apoptosis | 0.000000991 | Increased | 4.38 | 153 |
| Cancer, organismal injury and abnormalities | Malignant genitourinary solid tumor | 0.00000154 | Increased | 2.272 | 33 |
| Cancer, organismal injury and abnormalities | Frequency of tumor | 0.00000449 | Increased | 2.133 | 77 |
| Cancer, organismal injury and abnormalities, respiratory disease | Lung carcinoma | 0.0000213 | Increased | 2.584 | 23 |
| Cancer, organismal injury and abnormalities | Tumorigenesis of epithelial neoplasm | 0.0000284 | Increased | 2.18 | 52 |
| Cancer, organismal injury and abnormalities | Development of adenocarcinoma | 0.0000307 | Increased | 2.525 | 28 |
| Cell death and survival | Apoptosis of neurons | 0.0000414 | Increased | 3.077 | 39 |
| Gastrointestinal disease, hepatic system disease, organismal injury and abnormalities | Liver lesion | 0.000048 | Increased | 2.594 | 53 |
| Developmental disorder, embryonic development, organismal survival | Death of embryo | 0.0000693 | Increased | 4.258 | 22 |
| Cancer, organismal injury and abnormalities | Epithelial neoplasm | 0.0000849 | Increased | 2.115 | 65 |
| Cancer, organismal injury and abnormalities | Development of carcinoma | 0.0000998 | Increased | 2.229 | 39 |
| Cancer, organismal injury and abnormalities, respiratory disease | Development of lung carcinoma | 0.000105 | Increased | 2.559 | 17 |
| Cancer, organismal injury and abnormalities | Adenocarcinoma | 0.000142 | Increased | 2.587 | 30 |
| Cancer, cell death and survival, organismal injury and abnormalities | Cell death of tumor | 0.000151 | Increased | 3.66 | 30 |
| Cancer, cell death and survival, organismal injury and abnormalities, tumor morphology | Necrosis of tumor | 0.000254 | Increased | 3.66 | 29 |
| Cancer, organismal injury and abnormalities, respiratory disease | Lung adenocarcinoma | 0.000393 | Increased | 2.375 | 15 |
| Cancer, organismal injury and abnormalities, respiratory disease | Non-small cell lung carcinoma | 0.000584 | Increased | 2.559 | 17 |
| Cancer, organismal injury and abnormalities | Adenoma | 0.00077 | Increased | 2.042 | 25 |
| Developmental disorder, embryonic development | Degeneration of embryo | 0.000814 | Increased | 2.804 | 8 |
| Cancer, organismal injury and abnormalities | Carcinoma | 0.00101 | Increased | 2.006 | 48 |
| Developmental disorder, embryonic development, tissue morphology | Degeneration of embryoblast | 0.00135 | Increased | 2.433 | 6 |
| Cancer, cell death and survival, organismal injury and abnormalities, tumor morphology | Cell death of tumor cells | 0.00151 | Increased | 3.536 | 26 |
| Connective tissue disorders, developmental disorder, organismal injury and abnormalities, skeletal and muscular disorders | Dysplasia of skeleton | 0.00204 | Increased | 2.2 | 7 |
| Organismal survival | Perinatal death | 0.00223 | Increased | 6.322 | 60 |
| Developmental disorder, embryonic development, tissue morphology | Degeneration of embryonic tissue | 0.00329 | Increased | 2.63 | 7 |
| Neurological disease, organismal injury and abnormalities | Hydrocephalus | 0.00329 | Increased | 3.138 | 11 |
| Carbohydrate metabolism | Glycolysis of cells | 0.00522 | Increased | 2 | 8 |
| Lipid metabolism, molecular transport, small molecule biochemistry | Concentration of acylglycerol | 0.00562 | Increased | 2.147 | 33 |
| Cellular compromise | Dysfunction of mitochondria | 0.00654 | Increased | 2.213 | 5 |
| Cancer, cell death and survival, organismal injury and abnormalities, tumor morphology | Cell death of cancer cells | 0.00661 | Increased | 3.252 | 20 |
| Cancer, cell death and survival, organismal injury and abnormalities, tumor morphology | Cell death of osteosarcoma cells | 0.00721 | Increased | 3.742 | 14 |
| Cancer, organismal injury and abnormalities | Development of head and neck tumor | 0.00724 | Increased | 2.189 | 11 |
| Cellular assembly and organization, cellular function and maintenance | Organization of cytoskeleton | 6.48E-08 | Decreased | –3.001 | 101 |
| Nervous system development and function, tissue morphology | Quantity of neurons | 0.000000822 | Decreased | –2.239 | 53 |
| | Organization of cytoplasm | 0.000000842 | Decreased | –3.001 | 103 |

(Continued on following page)

**TABLE 1 |** (*Continued*) IPA generated disease function analysis for zebrafish male and female brain proteome induced by hypoxia following recovery.

**Male ZF brain proteome analysis (HM/NM)**

| Categories | Diseases or functions annotation | *p*-value | Predicted activation state | Activation z-score | # Molecules |
|---|---|---|---|---|---|
| Cellular assembly and organization, cellular function and maintenance | | | | | |
| Cellular assembly and organization, cellular function and maintenance | Microtubule dynamics | 0.00000115 | Decreased | −3.268 | 86 |
| Cell morphology, cellular assembly and organization, cellular function and maintenance | Formation of cellular protrusions | 0.00000314 | Decreased | −2.624 | 70 |
| Cancer, organismal injury and abnormalities | Growth of tumor | 0.00000351 | Decreased | −3.22 | 75 |
| Cellular movement | Cell movement | 0.00000707 | Decreased | −4.559 | 132 |
| Tissue morphology | Quantity of cells | 0.00001 | Decreased | −3.306 | 171 |
| Cellular growth and proliferation, connective tissue development and function, tissue development | Proliferation of connective tissue cells | 0.0000108 | Decreased | −2.856 | 49 |
| Cellular movement | Migration of cells | 0.0000139 | Decreased | −4.26 | 117 |
| Cellular movement, nervous system development and function | Migration of neurons | 0.0000177 | Decreased | −2.055 | 30 |
| Connective tissue development and function, tissue development | Growth of connective tissue | 0.0000256 | Decreased | −2.686 | 50 |
| Cellular assembly and organization | Quantity of intermediate filaments | 0.000211 | Decreased | −2 | 4 |
| Nervous system development and function | Sensation | 0.00041 | Decreased | −2.482 | 30 |
| Cancer, organismal injury and abnormalities | Neoplasia of tumor cell lines | 0.000582 | Decreased | −2.44 | 15 |
| Cellular development, cellular growth and proliferation, nervous system development and function, tissue development | Development of neurons | 0.000782 | Decreased | −2.282 | 63 |
| Cellular function and maintenance | Cellular homeostasis | 0.000792 | Decreased | −2.034 | 103 |
| Organismal development | Size of animal | 0.00102 | Decreased | −2.322 | 22 |
| Cell-to-cell signaling and interaction, nervous system development and function | Auditory evoked potential | 0.00126 | Decreased | −2.725 | 12 |
| Behavior | Learning | 0.00151 | Decreased | −2.11 | 41 |
| Lipid metabolism, small molecule biochemistry, vitamin and mineral metabolism | Synthesis of steroid hormone | 0.00168 | Decreased | −2.219 | 6 |
| Tissue development | Formation of gland | 0.00313 | Decreased | −2.088 | 25 |
| Embryonic development, organismal development | Development of body trunk | 0.00377 | Decreased | −3.116 | 95 |
| Embryonic development, organ development, organismal development, skeletal and muscular system development and function, tissue development | Formation of muscle | 0.00389 | Decreased | −2.398 | 32 |
| Cancer, organismal injury and abnormalities | Metastasis of tumor cell lines | 0.00431 | Decreased | −2.556 | 10 |
| Organismal development | Development of genitourinary system | 0.0049 | Decreased | −3.43 | 86 |
| Auditory and vestibular system development and function, nervous system development and function | Hearing | 0.00506 | Decreased | −2.157 | 15 |
| Amino acid metabolism,post-translational modification, small molecule biochemistry | Phosphorylation of L-amino acid | 0.00686 | Decreased | −2 | 13 |

**Female ZF brain proteome analysis (HF/NF)**

| Categories | Diseases or functions annotation | *p*-value | Predicted activation state | Activation z-score | # Molecules |
|---|---|---|---|---|---|
| Cellular assembly and organization, cellular function and maintenance | Organization of cytoskeleton | 6.83E-08 | Increased | 2.675 | 101 |
| Cellular assembly and organization, cellular function and maintenance | Organization of cytoplasm | 0.000000884 | Increased | 2.675 | 103 |
| Cellular assembly and organization, cellular function and maintenance | Microtubule dynamics | 0.0000012 | Increased | 2.941 | 86 |
| Cell morphology, cellular assembly and organization, cellular function and maintenance | Formation of cellular protrusions | 0.00000326 | Increased | 2.483 | 70 |
| Cancer, organismal injury and abnormalities | Growth of tumor | 0.00000365 | Increased | 2.325 | 75 |
| Cellular movement | Cell movement | 0.00000746 | Increased | 2.14 | 132 |
| Nucleic acid metabolism | Metabolism of nucleic acid component or derivative | 0.000087 | Increased | 2.209 | 27 |
| Cell morphology, cellular function and maintenance | Autophagy | 0.000186 | Increased | 2.393 | 24 |
| Nucleic acid metabolism, small molecule biochemistry | Metabolism of nucleotide | 0.000212 | Increased | 2.209 | 23 |
| Organismal survival | Viability | 0.000296 | Increased | 3.302 | 14 |
| Nervous system development and function | Sensation | 0.000417 | Increased | 2.058 | 30 |
| Cellular function and maintenance | Cellular homeostasis | 0.000821 | Increased | 2.663 | 103 |

(Continued on following page)

**TABLE 1 |** (*Continued*) IPA generated disease function analysis for zebrafish male and female brain proteome induced by hypoxia following recovery.

| Male ZF brain proteome analysis (HM/NM) | | | | | |
|---|---|---|---|---|---|
| **Categories** | **Diseases or functions annotation** | ***p*-value** | **Predicted activation state** | **Activation z-score** | **# Molecules** |
| Organismal development | Size of animal | 0.00104 | Increased | 2.322 | 22 |
| Embryonic development, organismal development | Growth of embryo | 0.00133 | Increased | 2.454 | 46 |
| Cellular movement, embryonic development | Cell movement of embryonic cells | 0.00172 | Increased | 2.2 | 12 |
| Embryonic development, organismal development | Development of body trunk | 0.00389 | Increased | 2.736 | 95 |
| Cancer, organismal injury and abnormalities | Metastasis of tumor cell lines | 0.00434 | Increased | 2.008 | 10 |
| Nervous system development and function, tissue morphology | Quantity of neuroglia | 0.00449 | Increased | 2.402 | 14 |
| Nucleic acid metabolism, small molecule biochemistry | Synthesis of nucleotide | 0.00583 | Increased | 2.019 | 15 |
| Respiratory system development and function | Respiration of mice | 0.00693 | Increased | 2 | 9 |
| Embryonic development, organ development, organismal development, skeletal and muscular system development and function, tissue development | Development of striated muscle | 0.00767 | Increased | 2.236 | 17 |
| Organismal survival | Organismal death | 2.59E-11 | Decreased | −7.796 | 251 |
| Developmental disorder, embryonic development, organismal survival | Death of embryo | 0.0000704 | Decreased | −2.198 | 22 |
| Cancer, cell death and survival, organismal injury and abnormalities | Cell death of tumor | 0.000153 | Decreased | −2.141 | 30 |
| Cancer, cell death and survival, organismal injury and abnormalities, tumor morphology | Necrosis of tumor | 0.000258 | Decreased | −2.141 | 29 |
| Cancer, cell death and survival, organismal injury and abnormalities, tumor morphology | Cell death of tumor cells | 0.00154 | Decreased | −2.363 | 26 |
| Connective tissue disorders, developmental disorder, organismal injury and abnormalities, skeletal and muscular disorders | Dysplasia of skeleton | 0.00205 | Decreased | −2.2 | 7 |
| Organismal survival | Perinatal death | 0.00228 | Decreased | −3.588 | 60 |
| Organismal survival | Death of perinatal stage organism | 0.00619 | Decreased | −2.137 | 11 |
| Cellular compromise | Dysfunction of mitochondria | 0.00657 | Decreased | −2.213 | 5 |

were identified to be regulated differentially after the data analysis from experimental runs in triplicate. The resuling proteins (2,323) were used to generate a clustered heatmap for showing the sex difference in expression patterns (**Figure 1B**). A majority (i.e., 84%) of differentially regulated proteins, 1,968 in total, were found downregulated in HM versus NM. In contrast, more than half (51%) of differentially regulated proteins, 1,188 in total, were found upregulated in HF versus NF (**Figure 1C**). 1,535 proteins were differentially regulated, among those 1,518 (98%) proteins were upregulated in females and downregulated in males whereas only 17 (2%) proteins were upregulated in males and downregulated in females, showing a clear differential regulation in a sex-specific manner. Another 554 proteins showed a similar expression pattern in both sexes, where 478 proteins were upregulated and 76 proteins downregulated. A list of differentially regulated proteins is provided in **Table 1**.

## Protein Enrichment Analysis for Zebrafish Brain Proteome Induced by Acute Hypoxia and During Recovery

Based on the zebrafish annotated database, IPA mapped 994 proteins out of 2,323 proteins identified in the iTRAQ analysis. These 994 proteins included different types of proteins i.e., transporters, transmembrane receptors, translation and transcription regulators, phosphatases, peptidase, kinases, enzymes, G-protein coupled receptors, ligand-binding receptors, and cytokines (**Figure 2A**). Out of all the groups, the majority of proteins belonged to the group "transcription regulators."

The protein enrichment analysis on the altered proteome was performed using the Ingenuity Pathway Analysis (IPA) software. The disease function pathway-based heat map generated by the IPA (**Figures 2B,C**) clearly showed a sex-specific difference in the altered expression of proteins in different disease pathway conditions.

The IPA analysis for disease function annotation showed a predicted activation state with activation z-score and *p*-values and molecules involved in each category of disease function. Among all the 502 proteins mapped in IPA for the disease and function analysis, in males 65 categories of disease function showed the predicted activation state: 37 categories showed increased activation states while 28 categories exhibited decreased activation states. In females only 30 categories of disease function showed the predicted activation state and among those, 21 categories of disease function showed increased activation states and only 9 categories showed decreased activation states.

The most striking feature was the contrasting regulation between male and female in one of the disease and function categories named "organismal survival" and "leads to organismal death" with a very significant ($p$ = 2.31E-11) activation z-score (12.434) identifying 251 molecules with an increased activation state, but the same 251 molecules showed a significantly (2.59E-11) decreased activation state with a z-score (−7.796) in females (**Table 1**). In males most of the increased activation state was

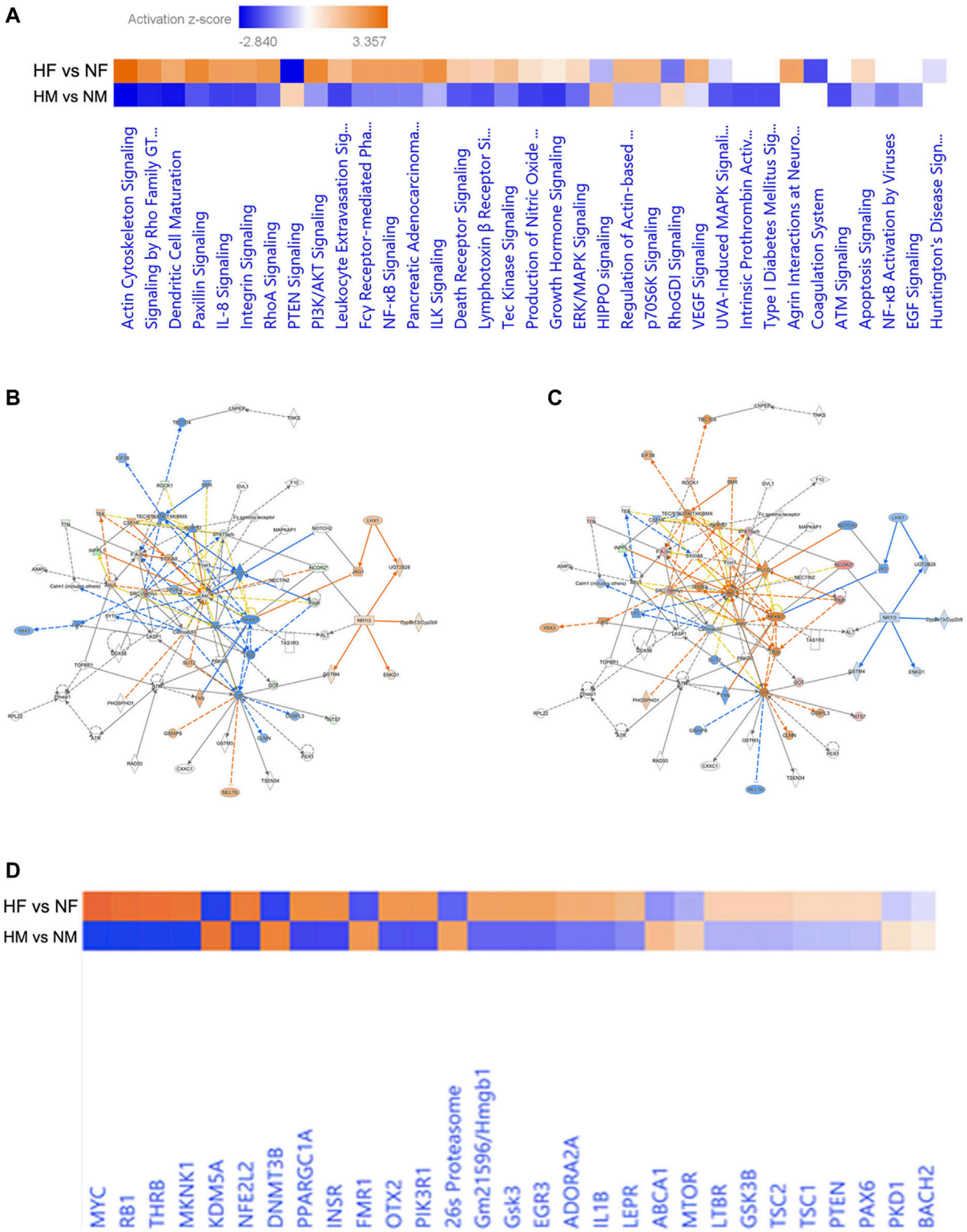



**FIGURE 3 |** Pathway based on the top networks. Comparative heatmap for canonical pathways of male and female brain proteome induced by hypoxia **(A)**, Predictive pathway for zebrafish male **(B)** and female **(C)** brain proteome induced by hypoxia, comparative heatmap for upstream regulators **(D)**.

**TABLE 2 |** IPA generated upstream regulators analysis for male and female zebrafish brain proteome induced by hypoxia following recovery.

**Upstream regulators in male zebrafish brain with predicted activation state**

| S. No. | Upstream regulator | Molecule type | Predicted activation state | Activation z-score | *p*-value of overlap | Target molecules in dataset |
|---|---|---|---|---|---|---|
| 1 | MYC | Transcription regulator | Inhibited | −2.24 | 4.52E-05 | BUB1, CCNA2, ENO1, EZH1, GLS, GOT1, GOT2, GPI, HIF1A, LDHB, MCM6, MCM7, PA2G4, PDHA1, PDK1, PGAM1, PGK1, PKM, TCF3 |
| 2 | RB1 | Transcription regulator | Inhibited | −3.183 | 0.000472 | ACO2, Actn3, APAF1, ATP1A1, ATP5F1A, BCKDHA, CASP9, CCNA2, CKM, DLST, FABP2, HSBP1, KRT18, MCM7, MFN2, MLYCD, MYH4, MYH6, MYH7, NDUFB10, RBL1, TIMM22 |
| 3 | TSC2 | Other | Inhibited | −2.433 | 0.000695 | IRS1, IRS2, MCL1, PDGFRB, PRKCA, PSMC3 |
| 4 | CSF2 | Cytokine | Inhibited | −3.141 | 0.00347 | BUB1, C3, CCNA2, CHTF18, E2F8, EXO1, FBXO5, FIGNL1, IL12B, KNTC1, MCM6, MNS1, NOS2, POLD1, POLE, SMC2, TLR4 |
| 5 | MKNK1 | Kinase | Inhibited | −3.317 | 0.00792 | APC, ATP1A3, CRMP1, DPYSL3, HADHA, KIF5A, MYO6, PRKAR1B, SNAP25, STXBP1, THRA |
| 6 | Gsk3 | Group | Inhibited | −2.219 | 0.00826 | COL2A1, KDR, MYH6, NOS2, STAT1 |
| 7 | NFE2L2 | Transcription regulator | Inhibited | −2.066 | 0.00931 | ACTG1, ALAS2, APOA4, ATP1A1, CDKN2C, ESD, FKBP5, G6PD, GFAP, GSTP1, HMOX1, L1CAM, MCFD2, NCKAP1, NOS2, NQO1, PDIA3, PFN2, PSMC1, PSMC3, PSMD11, RAN, SCG2, SREBF1, SYP, TTR, VCP |
| 8 | THRB | Ligand-dependent nuclear receptor | Inhibited | −3.054 | 0.0121 | ABCD3, CSHL1, DDC, DIO1, FGFR3, IGFBP2, MAPK8, MYH6, MYH7, NCOR2, STAT5B, WNT4, YWHAE |
| 9 | LEPR | Transmembrane receptor | Inhibited | −2.791 | 0.0139 | APOA1, APOA4, CREB3L2, CSHL1, EXOC4, GFAP, HIF1A, INPPL1, IRS2, MMP14, NBN, PLCB3, SNAP25, SREBF1 |
| 10 | IL1B | Cytokine | Inhibited | −2.59 | 0.0182 | A2M, ATP1A1, C3, COL2A1, FKBP5, FOXO1, HAS2, HIF1A, KIF15, MMP9, NOS2, STAT1 |
| 11 | EGFR | Kinase | Inhibited | −2.314 | 0.0258 | ACY1, ATAD3A, CCNA2, CCT5, GFAP, HAS2, MMP14, MMP9, PA2G4, TUBA4A, UBA1 |
| 12 | OTX2 | Transcription regulator | Inhibited | −2.219 | 0.0435 | A2M, EN1, PRDM1, SIX3, TF, TTR |
| 13 | UCHL1 | Peptidase | Inhibited | −2 | 0.366 | ANXA6, LDHB, MAPK6, SCP2 |
| 14 | STAT6 | Transcription regulator | Inhibited | −2.433 | 1 | BCL6, Cmah, IL12B, IRS2, MMP14, MMP9, MYO6, NCOA3, SERPINA1 |
| 15 | KDM5A | Transcription regulator | Activated | 2.688 | 0.00249 | ACO2, Actn3, ATP1A1, ATP5F1A, BCKDHA, DLST, HSBP1, MFN2, MLYCD, MYH4, MYH6, MYH7, NDUFB10, TIMM22 |
| 16 | 26s Proteasome | Complex | Activated | 2.236 | 0.00826 | APAF1, BHLHE22, FOXO1, NOTCH1, PRKCA |
| 17 | HAND1 | Transcription regulator | Activated | 2 | 0.0179 | KDR, MLYCD, NOTCH1, NRP1 |
| 18 | SATB1 | Transcription regulator | Activated | 2 | 0.472 | APC, ETS1, NCOR1, NR2C2 |

**Upstream regulators in female zebrafish brain with predicted activation state**

| S. No. | Upstream regulator | Molecule type | Predicted activation state | Activation z-score | *p*-value of overlap | Target molecules in dataset |
|---|---|---|---|---|---|---|
| 1 | MYC | Transcription regulator | Activated | 2.801 | 4.58E-05 | BUB1, CCNA2, ENO1, EZH1, GLS, GOT1, GOT2, GPI, HIF1A, LDHB, MCM6, MCM7, PA2G4, PDHA1, PDK1, PGAM1, PGK1, PKM, TCF3 |
| 2 | INSR | Kinase | Activated | 2.124 | 0.000466 | ACO2, ACTA1, ACTN4, ALDH6A1, ATP5F1A, ATP5F1B, CS, DCTN4, FLNC, GOT2, HADHA, HSPD1, IDH3A, IGF2R, INSR, MDH2, MMP9, MPEG1, MYH7, NAMPT, OGDH, PDHA1, PDHB, PKLR, SCP2, SREBF1 |
| 3 | Gm21596/Hmgb1 | Transcription regulator | Activated | 2.219 | 0.00101 | HIF1A, NOS2, PKM, SIGIRR, TLR4 |
| 4 | PIK3R1 | Kinase | Activated | 2.621 | 0.00711 | FOXO1, HIF1A, HMOX1, IL12B, NOS2, PDHA1, PDK1, PKM |
| 5 | MKNK1 | Kinase | Activated | 2.111 | 0.00798 | APC, ATP1A3, CRMP1, DPYSL3, HADHA, KIF5A, MYO6, PRKAR1B, SNAP25, STXBP1, THRA |
| 6 | NFE2L2 | Transcription regulator | Activated | 2.705 | 0.00943 | ACTG1, ALAS2, APOA4, ATP1A1, CDKN2C, ESD, FKBP5, G6PD, GFAP, GSTP1, HMOX1, L1CAM, MCFD2, NCKAP1, NOS2, NQO1, PDIA3, PFN2, PSMC1, PSMC3, PSMD11, RAN, SCG2, SREBF1, SYP, TTR, VCP |
| 7 | EGR3 | Transcription regulator | Activated | 2.219 | 0.0116 | BCL6, ESD, LMO7, NOTCH1, PABPC1L |
| 8 | THRB | Ligand-dependent nuclear receptor | Activated | 3.054 | 0.0122 | ABCD3, CSHL1, DDC, DIO1, FGFR3, IGFBP2, MAPK8, MYH6, MYH7, NCOR2, STAT5B, WNT4, YWHAE |
| 9 | MYB | Transcription regulator | Activated | 2 | 0.019 | CLTA, HSPA8, MAD1L1, NOTCH1, RGS8, SLC27A2, TULP4 |
| 10 | OTX2 | Transcription regulator | Activated | 2.219 | 0.0437 | A2M, EN1, PRDM1, SIX3, TF, TTR |

(Continued on following page)

**TABLE 2 |** (*Continued*) IPA generated upstream regulators analysis for male and female zebrafish brain proteome induced by hypoxia following recovery.

| Upstream regulators in male zebrafish brain with predicted activation state | | | | | | |
|---|---|---|---|---|---|---|
| **S. No.** | **Upstream regulator** | **Molecule type** | **Predicted activation state** | **Activation z-score** | ***p*-value of overlap** | **Target molecules in dataset** |
| 11 | SRF | Transcription regulator | Activated | 2.957 | 0.0551 | ACTA1, ACTG1, CKM, ETS1, ITGA2B, ITGB1, KDR, MYH4, MYH6, MYH7, PRSS57, TTN, VCL |
| 12 | HIF1A | Transcription regulator | Activated | 2.008 | 0.165 | ENO1, IL12B, MIF, NOS2, NOTCH1, PDHA1, PDK1, PKM, TTN |
| 13 | HOXD10 | Transcription regulator | Activated | 2 | 0.193 | DAPP1, RSAD2, TFR2, WDR5 |
| 14 | Creb | Group | Activated | 2 | 0.296 | ARHGEF9, CBWD1, GABBR1, GRK3, INTS7, MCL1, PGK1, POLE, PRKCA |
| 15 | HNF4A | Transcription regulator | Activated | 2.382 | 0.329 | CCNA2, ELMO1, FABP2, HIF1A, HSPA8, KRT8, NBEA, PFN2, PKM, RSPH4A, SCP2, SERPINA1, TF, TFR2, WNT4 |
| 16 | mir-223 | MicroRNA | Activated | 2 | 1 | ALCAM, CRHBP, NQO1, TLR4 |
| 17 | DNMT3B | Enzyme | Inhibited | −2.53 | 0.00754 | ACTA1, CKM, GRK3, KDR, MYH6, MYH7, PIK3C2B, PRKAR1B, SLC8A2, STAT1 |
| 18 | MAT1A | Enzyme | Inhibited | −2 | 0.0149 | APOA1, KRT18, MAT1A, PRDX6 |
| 19 | CHADL | Other | Inhibited | −2 | 0.061 | COL2A1, CSPG4, MN1, NTRK3 |
| 20 | PTPN1 | Phosphatase | Inhibited | −2.219 | 0.0762 | HHEX, IRS1, IRS2, NOS2, TMEM26 |
| 21 | DNMT3A | Enzyme | Inhibited | −2.121 | 0.142 | ACTA1, CKM, GRK3, IRS1, MYH6, MYH7, PIK3C2B, SLC8A2 |
| 22 | ZNF106 | Other | Inhibited | −2 | 0.204 | ALAS2, C3, NDUFB10, PRDX2 |
| 23 | BTNL2 | Transmembrane receptor | Inhibited | −2 | 0.55 | CDKN2C, DAPP1, NTRK3, S100A4 |
| 24 | DICER1 | Enzyme | Inhibited | −2.138 | 1 | CCNG1, CRH, ERBB2, HNRNPH1, IGF2R, ITGB1, MMP9, NOTCH1, PRKCA |

observed in organismal injury, abnormalities, cell death, connective tissue disorders, developmental disorder, skeletal and muscular disorders, neurological, respiratory diseases, and cancer; whereas in females cellular assembly and organization, cellular function and maintenance, cell morphology, nucleic acid metabolism, and cellular movement showed an increased activation state. The disease function analysis clearly showed a sex-specific difference in hypoxia-induced neural damage and recovery as seen in our earlier studies (Das et al., 2019).

The IPA generated core expression analysis shed light on sex differences in canonical pathways with their predictive upregulation and downregulation in expression (**Figure 3A**). Among all, the top five canonical pathways were Actin Cytoskeleton Signaling, TCA Cycle II (Eukaryotic) 14-3-3-mediated Signaling, Remodeling of Epithelial Adherens Junctions, and Huntington's Disease Signaling showed negative score in males while in females, a positive score was observed.

Hypoxia can induce cytoskeletal injury and remodeling through the activation of hypoxia-inducible factor-1α (HIF-1α) and HIF-1α activation results in actin cytoskeleton signaling (Weidemann et al., 2013; Huang et al., 2019). F-actin in non-muscle cells is to organize the actin cytoskeleton, which is utilized for cell locomotion, adhesion, and cell proliferation and we have observed activation of Factin in females (**Figures 3B,C**) indicating early proliferation in response to neural damage induced by hypoxia.

The core expression analysis of IPA led us to decipher many regulatory networks, disease function pathways, top upstream regulators and their predicted activation state, and also some biomarkers. Upon reviewing all the pathways involved, two individual pathways were found very interesting in males and females (**Figures 3A,B**), which clearly showed a gender-specific difference in the expression of a number of proteins in the pathway such as Rock1, Inppl1, Factin, Stat5ab, Ncor2, SRC-family, Got, Ints7 and Pdgfr, which were downregulated in the male brain but upregulated in the female brain. Though the pathways involved many molecules and networks, they were still centered around the AKT signaling pathway, which regulates a wide range of cellular functions and is involved in the resistance response to hypoxia-ischemia through the activation of proteins associated with cell survival, proliferation, and regulation of HIF-1α (Zhang et al., 2018). The growth factors and inflammation markers noticed in the pathway were studied in our previous study reported in (Das et al., 2019; Das et al., 2020).

The upstream regulators analyzed in IPA were 155 in male and 165 in female; among these 18 upstream regulators in male and 24 in female showed the predicted activation state (**Table 2**). Among the 18 upstream regulators in males, 14 were inhibited and only 4 were activated, and most of these were transcription regulators. Among the 24 upstream regulators in females, 16 were activated, with a majority of transcription regulators, and only with 8 were inhibited, which were not found in the males. Five upstream regulators [Myc, Mknk1, Nfe2l2 (Nrf2), Thrb, and otx2] were found to be common in both males and females with differential activation states, and interestingly these were in opposite directions, i.e., inhibited in males while activated in females.

**TABLE 3 |** IPA generated regulatory effect analysis for male and female zebrafish brain proteome induced by hypoxia following recovery.

**Male hypoxia regulator effects**

| ID | Consistency score | Regulators | Target total | Target molecules in dataset | Diseases & functions | Known regulator-disease/function relationship |
|---|---|---|---|---|---|---|
| 1 | 2.309 | HAND1,THRB | 12 | CSHL1,FGFR3,IGFBP2,KDR,MAPK8,MLYCD,MYH6,NCOR2,NOTCH1,NRP1,STAT5B,WNT4 | Cellular homeostasis, development of body trunk, development of genitourinary system | 67% (4/6) |
| 2 | 2.111 | MYC | 11 | BUB1,CCNA2,ENO1,GLS,GPI,HIF1A,PA2G4,PDHA1,PDK1,PKM,TCF3 | Carcinoma, frequency of tumor, growth of tumor, incidence of tumor | 100% (4/4) |
| 3 | 2 | NFE2L2 | 4 | PSMC1,PSMD11,RAN,VCP | Cell death of tumor cells | 0% (0/1) |
| 4 | 1.789 | OTX2 | 5 | A2M,EN1,PRDM1,SIX3,TF | Quantity of cells | 100% (1/1) |
| 5 | −5.715 | MYC | 6 | ENO1,GPI,HIF1A,PDK1,PGK1,PKM | Glycolysis of cells | 100% (1/1) |
| 6 | −7.506 | HAND1 | 3 | KDR,NOTCH1,NRP1 | Migration of cells | 0% (0/1) |
| 7 | −16.743 | HAND1 | 3 | KDR,NOTCH1,NRP1 | Organization of cytoplasm | 0% (0/1) |
| 8 | −19.23 | MYC | 5 | BUB1,CCNA2,GPI,HIF1A,PDK1 | Growth of connective tissue | 100% (1/1) |

**Female hypoxia regulator effects**

| ID | Consistency score | Regulators | Target total | Target molecules in dataset | Diseases & functions | Known regulator-disease/function relationship |
|---|---|---|---|---|---|---|
| 1 | 3.051 | Gm21596/ Hmgb1,PIK3R1,THRB | 13 | CSHL1,FGFR3,FOXO1,HIF1A,HMOX1,IL12B,MAPK8,MYH6,NCOR2,NOS2,STAT5B,TLR4,WNT4 | Autophagy, development of body trunk | 50% (3/6) |
| 2 | −4.082 | MKNK1 | 6 | ATP1A3,HADHA,KIF5A,SNAP25,STXBP1,THRA | Perinatal death | 0% (0/1) |
| 3 | −4.491 | PIK3R1 | 6 | FOXO1,HIF1A,HMOX1,IL12B,NOS2,PDK1 | Cell movement | 100% (1/1) |
| 4 | −5.367 | MKNK1 | 5 | APC,CRMP1,DPYSL3,KIF5A,MYO6 | Microtubule dynamics | 0% (0/1) |
| 5 | −6.5 | Gm21596/Hmgb1 | 4 | HIF1A,NOS2,PKM,TLR4 | Growth of tumor | 100% (1/1) |
| 6 | −7.5 | NFE2L2 | 4 | G6PD,NOS2,NQO1,VCP | Metabolism of nucleotide | 0% (0/1) |

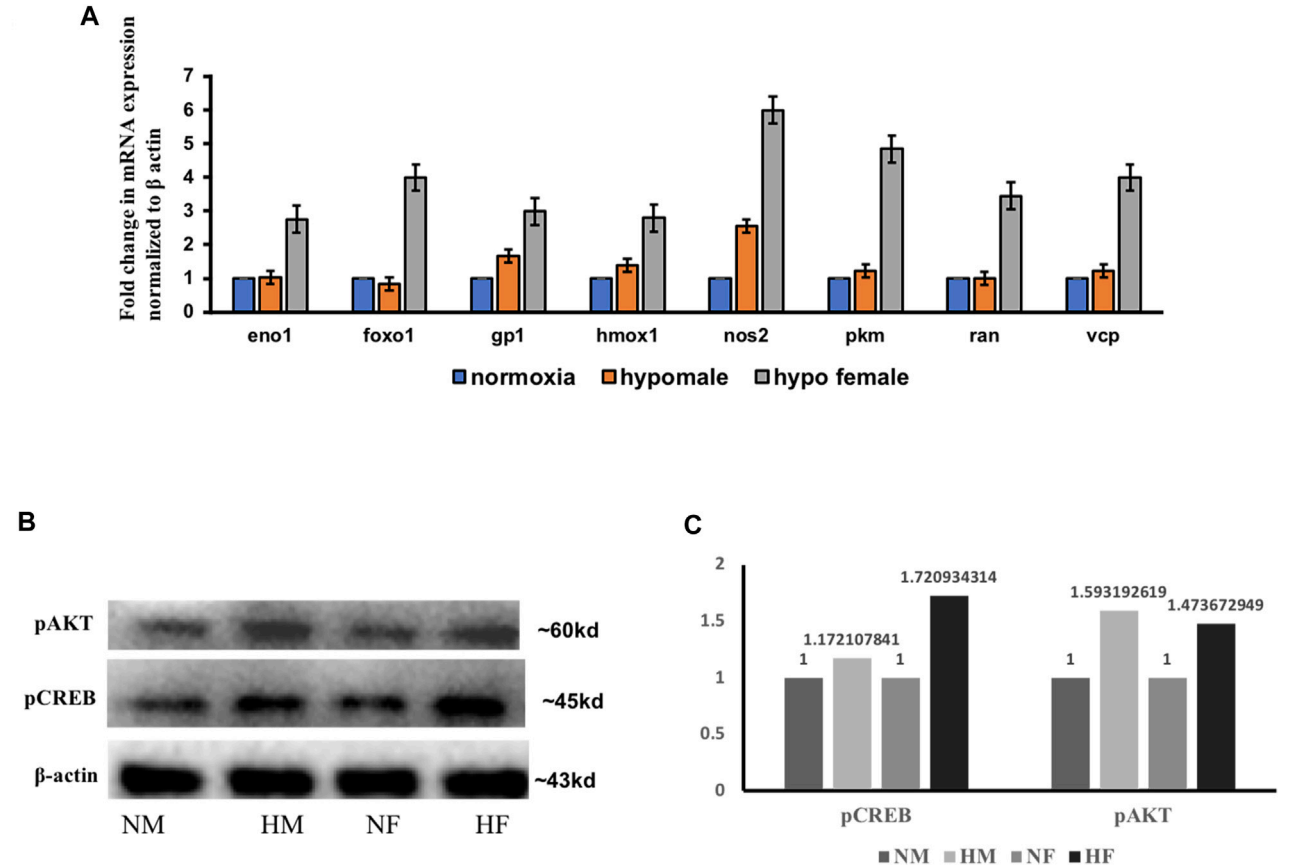



**FIGURE 4 |** Validation of few regulatory target molecules. Graph showing mRNA expression of eno1, foxo1, gp1, hmox1, nos2, pkm, ran, and vcp **(A)** The data are expressed as the mean ± SEM, ($n$ = 6 pooled brains). Immunoblot showing expression of pCREB and pAKT in male and female brain **(B)**, Densitometry for comparison of protein expression in HM vs. NM and HF vs. NF **(C)**.

**TABLE 4 |** Top 5 upregulated proteins retrieved from uniformly regulated (upregulated) in both male and female zebrafish brain induced by acute hypoxia.

| S. No. | Accession No. | Description | HM/NM | HF/NF |
|---|---|---|---|---|
| 1 | 56693350 | Very long-chain acyl-CoA synthetase | 3.234 | 1.690 |
| **2** | **71834420** | **Histone-lysine N-methyltransferase, H3 lysine-9 specific 5** | **3.188** | **1.603** |
| 3 | 189531944 | Predicted: hypothetical protein LOC100148665 | 3.057 | 1.966 |
| 4 | 326666355 | Predicted: zinc finger protein 208-like | 2.984 | 2.606 |
| 5 | 326664965 | Predicted: protein FAM5C | 2.742 | 3.703 |

*H3K9 which was the focus of manuscript and taken for further mechanistic analysis so made it bold.*

In **Table 3** the regulator effects of males and females are shown, where only one of the regulators, Nfe2l2, is common in both, but with only one common target molecule VCP (valosin containing protein), and all different target molecules in both the dataset VCP was earlier reported to be an AKT binding protein, and its expression was found enhanced in hypoxia (Klein et al., 2005).

## Validation of Few Selected Regulator Effects Molecules From IPA Analysis

A few of the target molecules (Eno1, Foxo1, Gp1, Hmox1, Nos2, Pkm, Ran, and Vcp) from both the data sets were considered for validation by quantitative Realtime PCR (**Figure 4**). eno1 (enolase 1) is one of the HIF target genes (Benita et al., 2009). The qPCR analysis revealed more than ~2-fold increase in eno1 and ran (ras-related nuclear protein) in females but it remained unchanged in males. The foxo transcriptional factors are important regulators of cell survival in response to various stresses including oxidative stress (Bakker et al., 2007). foxo1 was upregulated ~4-fold in females but unaffected in males thus indicating better survival response after hypoxia in females. The expression of gp1 (Glycoprotein 1) and hmox1 (Heme Oxygenase 1) showed a similar kind of expression pattern in both sexes, a mild upregulation in males, and ~3-fold upregulation in females. (gp1) acts as a glycolytic

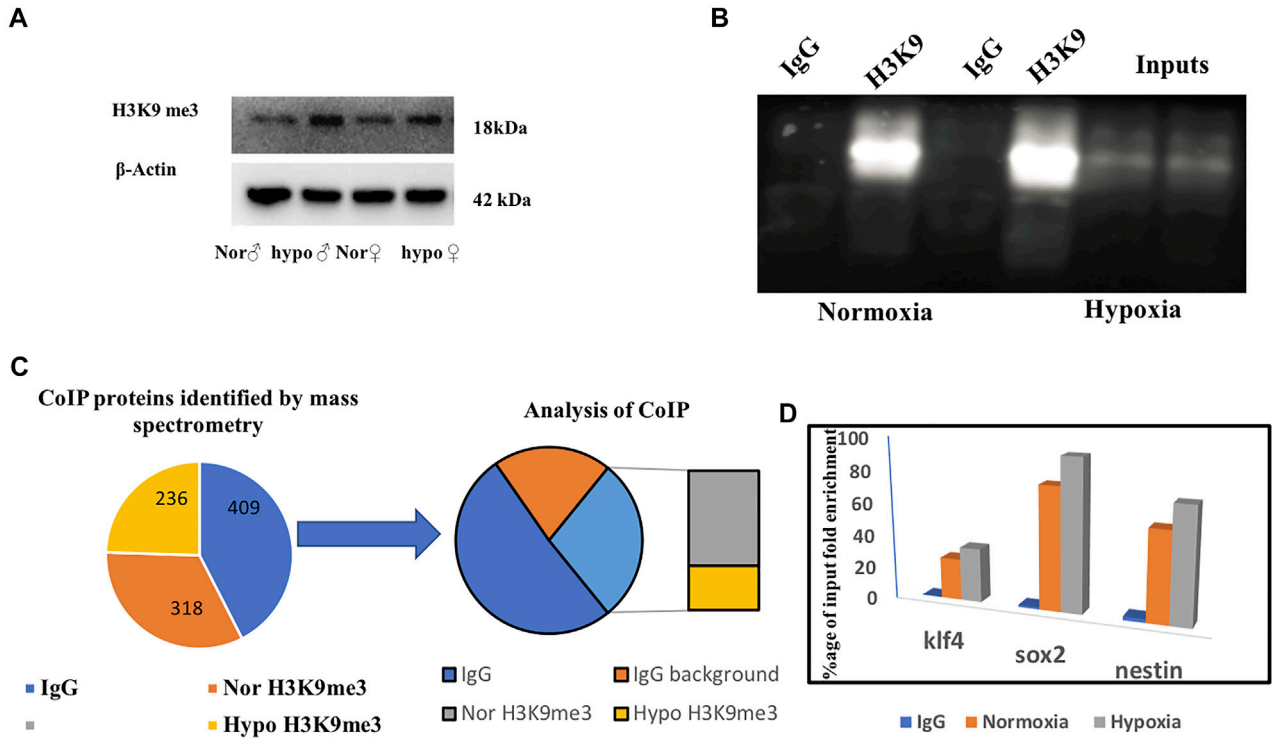



**FIGURE 5 |** Deciphering the role of H3K9me3 by co-immunoprecipitation. Immunoblot of H3K9me3 showing upregulation in hypoxia **(A)**, CoIP of H3K9me3 in zebrafish male brain nuclear protein **(B)**, Schematic representation of CoIP analysis **(C)**. ChIP qPCR showing H3K9 occupancy on target gene promoter. ChIP qPCR showing the increase in H3K9me3 enrichment on the promoter region of klf4, sox2, and nestin in hypoxic male brain **(D)**.

enzyme, as well as functioning as a tumor-secreted cytokine and an angiogenic factor (AMF) that stimulates endothelial cell motility, GPI is also a neurotrophic factor (Neuroleukin) for spinal and sensory neurons. The role of neurotrophic factors in repair mechanisms are well evident, therefore in our study 4 h post-hypoxia females are better in recovery speed as compared to males. Hmox1 has been shown to be induced by various stresses including hypoxia (Panchenko et al., 2000); our study also revealed an increase in its expression. The expression of nos2 (nitric oxide synthase) and pkm (pyruvate kinase M) is inducible with hypoxia and hif1 targets showed a higher fold upregulation in females as compared to males.

Various mediators like growth factors, glucose transporters, solute carriers, neurotransmitters, inflammatory molecules, and stress signals as well as factors known to modulate the intracellular cAMP or $Ca^{2+}$ levels can activate cAMP-responsive element-binding protein (CREB) through phosphorylation of serine 133 ($Ser^{133}$) by protein kinase A (PKA) and protein kinase B (PKB/AKT) (Steven et al., 2016). The pathway generated by IPA in **Figures 3A,B** was also centered on the AKT pathway with many downstream interacting molecules involved in cell death and repair mechanisms. Among all the molecules, we evaluated the expression of two of the major molecules in acute hypoxia recovery with respect to sex difference and found a sex difference in the expression of PCREB and PAKT (**Figures 4B,C**) where in females activation of CREB and AKT leads to early cell death survival and repair.

## Analysis of Proteins Showing Uniform (Either Upregulated or Downregulated) Expression in Both the Sexes

Throughout the protein enrichment analysis by IPA sex-specific, global proteome changes in the acute hypoxia zebrafish model were observed, which is in concurrence with our previous study (Das et al., 2019). But the question which remains unsolved is why the recovery in females is quicker than in males. At 4 h post-hypoxia when both the sexes survived coping up with the neural damage then there must be some common mechanism involved for recovery. So rather than looking further into the differentially expressed markers, we looked into the shared regulation of proteins. In **Figure 1B** we have shown the analysis of proteins resulted from iTRAQ where 554 proteins have a common expression pattern in both the sexes and among them 478 proteins were found regulated in one direction i.e., upregulated in both male and female brain in response to acute hypoxia. We hypothesized that as animals from both sexes are in the recovery process, therefore, a common mechanism of regulation may help to elucidate the mechanism behind the later recovery of the males from neural damage induced by hypoxia-ischemia. While looking into the 478 upregulated proteins, among the top five upregulated proteins in males, we identified histone-lysine N-methyltransferase H3 lysine-9 specific 5 protein, an epigenetic regulator displaying ~3-fold upregulation in the male brains and ~1.6-fold upregulation in the female brains (**Table 4**). Post-translational modifications of histones are widely recognized as an important epigenetic mechanism in the

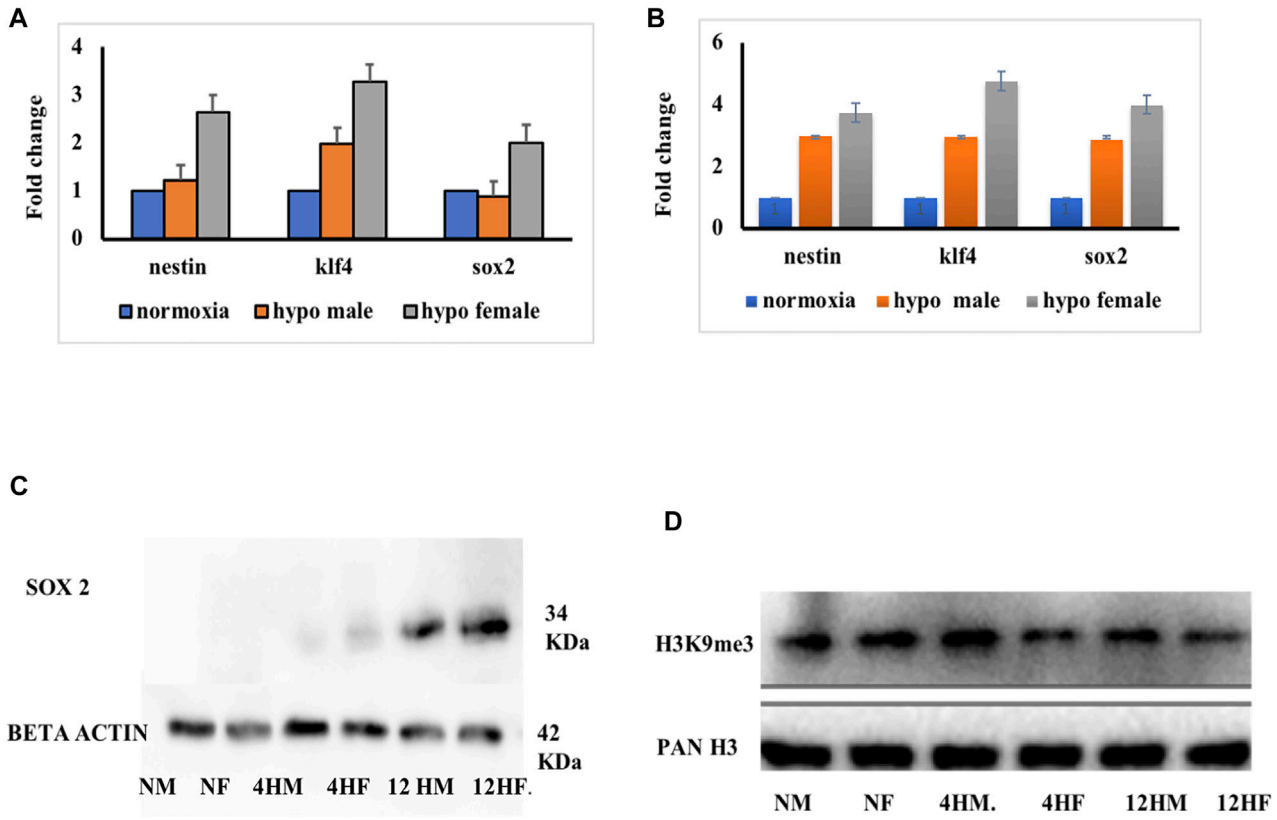



**FIGURE 6 |** Validation ChIP analysis. mRNA expression at 4 h **(A)** and 10 h **(B)** post hypoxia in both male and female (*n* = pooled 6 brain). Expression of sox2 protein at different times of recovery from hypoxia **(C)**. Expression of H3K9me3 at different times of recovery from hypoxia **(D)**.

organization of chromosomal domains and gene regulation. Methylation of lysine 4 and acetylation of lysine 9 of histone H3 has been associated with regions of active transcription, whereas methylation of H3K9 and H3K27 are generally associated with gene repression (Litt et al., 2001; Nakayama et al., 2001; Maison et al., 2002; Peters et al., 2002; Vakoc et al., 2006). Recently, hypoxia-induced histone modifications in neural gene regulation have been reported, and these were found on both hypoxia-activated and hypoxia-repressed genes (Johnson et al., 2008). H3K9 methylation is a critical epigenetic mark for gene repression and silencing. Hypoxia induces H3K9 methylation at different gene promoters, which is correlated with the repression and silencing of those genes following hypoxia (Lu et al., 2011).

## Deciphering the Role of H3K9me3 by Co-Immunoprecipitation and ChIP qPCR

Based on the previous literature (Lindeman et al., 2010), we hypothesized that H3K9 can be our prime target for deciphering late recovery in males as it was significantly upregulated in the male brain following hypoxia and being a repressive epigenetic mark in nature its high level can repress and/or silence a number of critical neural genes. Considering the role of H3K9me3 in hypoxia (Chakravarty et al., 2016) we immunoblotted for H3K9me3 using a specific antibody and performed a co-immunoprecipitation (CoIP) to identify the interacting proteins of H3K9me3 in hypoxic condition (**Figures 5A–C**). We could validate the expression of H3K9me3 through immunoblotting with an upregulation of H3K9me3 in hypoxia males when compared to normoxia males (**Figure 5A**). For CoIP experiment, nuclear extract was isolated from male zebrafish brains. The eluted proteins were then detected for immunoprecipitated and co-immunoprecipitated proteins by SDS-PAGE followed by western blotting. Then, 5% of the initial lysates were used as the input (**Figure 5B**). A mass spectrometric approach was used to identify the co-immunoprecipitated proteins obtained by the pull-down of the target antibody. The resultant peptides from MS/MS for four groups (normoxia IgG, normoxia and hypoxia H3k9me3 pull-down) were analyzed and after removing the background of IgG pooled proteins we could obtain 153 proteins identified in male normoxia H3K9me3 pull-down and 72 proteins identified in male hypoxia H3K9me3 pull down. Surprisingly, there were no common proteins in the normoxia and hypoxia H3K9me3 pull-downs, showing hypoxia stress may lead to alteration in interacting proteins. Further, we went through our iTRAQ data and tried to see whether these co-immunoprecipitated proteins were also found altered post-hypoxia in our high throughput proteomics data where almost all the proteins were found to overlap (**Figure 5C**).

The co-interacted proteins of H3K9me3 could not answer the unresolved question of why the males are recovering later. Therefore, we thought of evaluating the transcriptional targets of H3K9me3 to get an answer to our question, as H3K9me3 is a repressive marker so its upregulation in males may repress any neurogenic marker needed for recovery from hypoxia-induced neural damage. An earlier report on chromatin state of the developmentally regulated genes (Lindeman et al., 2010) led us to explore the striking upregulation of transcriptionally repressive epigenetic marker H3K9me3 at 4 h post hypoxia in zebrafish male brain.

The ChIP-qPCR data showed the repression of early neurogenesis markers nestin, klf4, and sox2 in the zebrafish male brain 4 h post-hypoxia (**Figure 5D**). It is pertinent to mention here that the ChIP assay was not performed on female zebrafish brains as the male brain proteome showed a higher fold upregulation in H3K9me3. For further validating the data the mRNA expression levels for nestin, klf4, and sox2 at two-time points of recovery i.e., at 4 h (**Figure 6A**) and 12 h (**Figure 6B**) post hypoxia, was assessed. The qPCR analysis showed at 4 h post-hypoxia the expression of early neurogenic markers showed mild activation in males and later at 12 h post-hypoxia the expression was much higher. The protein level expression of Sox2 was evaluated at low concentration (25 μg) of protein which showed in both males and females at 4 h post-hypoxia but the expression was quite low in both the sexes; in the male it was almost negligible, however in female a mild expression was observed, which at the later time point i.e., 12 h post-hypoxia showed noticeable upregulation in both male and female brain (**Figure 6C**).

To further validate if the sox2 expression is dependent on H3K9me3 level, the expression level of H3K9me3 was assessed and predictably it was found upregulated in the male brain as shown in the previous experiment, compared to the female brain at 4 h post-hypoxia. Later at 12 h post-hypoxia, the level of H3K9me3 was much less in males than what it was at 4 h post-hypoxia (**Figure 6D**). This result suggested that with the activation of H3K9me3 the expression levels of early neurogenic markers are getting repressed. This could be the possible reason for late recovery in males as early neurogenic markers are not fully activated in response to hypoxia insult, in contrast to the female brain.

Among Cerebral strokes, ischemic stroke is the most common type of stroke and a major cause of death and/or disability worldwide, though there are continuous efforts to establish a proper diagnosis and efficient therapy. The proteomics study complements both genomics and transcriptomics and simultaneously provides information about the proteins that can be implemented for main functional mediators of cells such as their post-translational modification and their interactions with biological molecules. However, post-stroke is mostly related to protein function which can be a direct target for therapeutic intervention. Therefore in the present study, we performed a quantitative proteomics approach for hypoxia-induced brain to identify favorable biomarkers involved in neuronal injury and recovery (Li et al., 2019). In our previous study, clear sex-specific differences were observed in acute hypoxia-induced neural damage and recovery but to explore more about the mechanism of recovery in the present study we have focused on a 4 h post hypoxia timepoint, predicting this could possibly be a viable therapeutic window. To date, many high throughput studies on hypoxia (Durukan and Tatlisumak, 2007; Cuadrado et al., 2010; Goldenberg et al., 2014; Chen et al., 2015; Durukan and Tatlisumak, 2007; Cuadrado et al., 2010; Goldenberg et al., 2014; Chen et al., 2015; Ton et al., 2003; van der Meer et al., 20052005; Shah et al., 2019) gave sufficient information about the genes and proteins involved in hypoxia and related diseases but the roles of these hallmarked hypoxia markers are not well studied in a sex-specific context. Therefore, we have attempted to emphasize more on the sex-specific neural regulation post-hypoxia, which will provide a better insight into designing efficient therapeutics for patients who suffered acute hypoxic insults. The prevalence of hypoxic brain damage is increasing and prognostic factors for either poor or good outcome are lacking (Heinz and Rollnik, 2015).

The advantage over traditional proteomics and iTRAQ based proteomics is that in iTRAQ all four groups can be simultaneously processed to reduce the error rate and post-translational modifications can also be quantified. The present study on whole zebrafish brain proteome upon global acute hypoxia sheds light on many differential roles of protein markers which can be further validated. Solute carrier (SLC) transporters are well-known therapeutic targets (Lin et al., 2015) and in our study too we have observed a very high activation in recovery. We tried to elucidate the role of one of the histone-based epigenetic regulatory mechanism (H3K9me3) that controls adult neurogenesis during the recovery phase post-hypoxia-ischemia. There is hardly any study on the epigenetic mechanisms in the zebrafish brain to date. Here, we identified hundreds of transcription factors involved in post-hypoxia recovery in a gender-specific manner, which can add to the development of a better therapeutic strategy.

## CONCLUSION

To conclude we have studied the sex-specific difference in global proteome changes in zebrafish brain induced by acute hypoxia and during the recovery. We elucidated the unresolved question from our previous study (Das et al., 2019) regarding the delayed recovery in males following hypoxic insult. With the striking upregulation of H3K9me3 in males at 4 h post-hypoxia, the early neurogenic markers like nestin, klf4, and sox2 expression level got affected, which might be the reason for late recovery in males, compared to females. Acute hypoxia-induced sex-specific comparison of brain proteome led us to reveal many differentially expressed proteins including the novel ones, which can be further studied for the development of novel targets and a better therapeutic strategy.

## DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: http://proteomecentral.proteomexchange.org/, PXD027528.

## ETHICS STATEMENT

The animal study was reviewed and approved by: All applicable international, national, and/or institutional guidelines for the care and use of animals were followed. The approved institutional protocol number was IICT/CB/SC/281114/30 under registration no#97/1999/CPCSEA. Further, this article does not contain any studies with human participants performed by any of the authors.

## AUTHOR CONTRIBUTIONS

TD: Conceptualization, Methodology, Validation, Formal analysis, Investigation, Writing—Original Draft. AvK: Methodology, Validation, Formal analysis. ArK: Conceptualization, Resources, Supervision, Project administration, Funding acquisition, Writing—Review & Editing. SC: Conceptualization, Resources, Supervision, Project administration, Funding acquisition, Writing—Review & Editing.

## FUNDING

This study was funded jointly by the Council of Scientific and Industrial Research (CSIR), India network project (BSC0103-UNDO to SC and AK), and Department of Biotechnology, Government of India project (BT/PR14338/MED/30/495/2010 to SC).

## ACKNOWLEDGMENTS

The authors thank the Director, CSIR-IICT for the necessary support and the Knowledge and Information Management (KIM) Department for generating the official communication number (IICT/Pubs/2019/436).

## REFERENCES

Bakker, W. J., Harris, I. S., and Mak, T. W. (2007). FOXO3a Is Activated in Response to Hypoxic Stress and Inhibits HIF1-Induced Apoptosis via Regulation of CITED2. *Mol. Cel.* 28 (6), 941–953. S1097-2765(07)00826-X [pii]. doi:10.1016/j.molcel.2007.10.035

Benita, Y., Kikuchi, H., Smith, A. D., Zhang, M. Q., Chung, D. C., and Xavier, R. J. (2009). An Integrative Genomics Approach Identifies Hypoxia Inducible Factor-1 (HIF-1)-Target Genes that Form the Core Response to Hypoxia. *Nucleic Acids Res.* 37 (14), 4587–4602. gkp425 [pii]. doi:10.1093/nar/gkp425

Bickler, P. E., and Buck, L. T. (2007). Hypoxia Tolerance in Reptiles, Amphibians, and Fishes: Life with Variable Oxygen Availability. *Annu. Rev. Physiol.* 69, 145–170. doi:10.1146/annurev.physiol.69.031905.162529

Braga, M. M., Silva, E. S., Moraes, T. B., Schirmbeck, G. H., Rico, E. P., Pinto, C. B., et al. (2016). Brain Zinc Chelation by Diethyldithiocarbamate Increased the Behavioral and Mitochondrial Damages in Zebrafish Subjected to Hypoxia. *Sci. Rep.* 6, 20279, 2016 srep20279 [pii]. doi:10.1038/srep20279

Catherine, O. J., and Collaborators, G. S. (2019). Global, Regional, and National burden of Stroke, 1990-2016: a Systematic Analysis for the Global Burden of Disease Study 2016. *Lancet Neurol.* 18 (5), 439–458. S1474-4422(19)30034-1 [pii]. doi:10.1016/S1474-4422(19)30034-1

Chakravarty, S., Jhelum, P., Bhat, U. A., Rajan, W. D., Maitra, S., Pathak, S. S., et al. (2016). Insights into the Epigenetic Mechanisms Involving Histone Lysine Methylation and Demethylation in Ischemia Induced Damage and Repair Has Therapeutic Implication. *Biochim. Biophys. Acta Mol. Basis Dis.* 1863 (1), 152–164. S0925-4439(16)30238-1 [pii]. doi:10.1016/j.bbadis.2016.09.014

Chakravarty, S., Reddy, B. R., Sudhakar, S. R., Saxena, S., Das, T., Meghah, V., et al. (2013). Chronic Unpredictable Stress (CUS)-induced Anxiety and Related Mood Disorders in a Zebrafish Model: Altered Brain Proteome Profile Implicates Mitochondrial Dysfunction. *PLoS One* 8 (5), e63302, 2013. [pii]. doi:10.1371/journal.pone.0063302.PONE-D-12-39866

Chen, J. H., Kuo, H. C., Lee, K. F., and Tsai, T. H. (2015). Global Proteomic Analysis of Brain Tissues in Transient Ischemia Brain Damage in Rats. *Int. J. Mol. Sci.* 16 (6), 11873–11891. ijms160611873 [pii]. doi:10.3390/ijms160611873

Chen, K., Cole, R. B., and Rees, B. B. (2013). Hypoxia-induced Changes in the Zebrafish (*Danio rerio*) Skeletal Muscle Proteome. *J. Proteomics* 78, 477–485. S1874-3919(12)00712-9 [pii]. doi:10.1016/j.jprot.2012.10.017

Cuadrado, E., Rosell, A., Colome, N., Hernandez-Guillamon, M., Garcia-Berrocoso, T., Ribo, M., et al. (2010). The Proteome of Human Brain after Ischemic Stroke. *J. Neuropathol. Exp. Neurol.* 69 (11), 1105–1115. doi:10.1097/NEN.0b013e3181f8c539

Das, T., Soren, K., Yerasi, M., Kamle, A., Kumar, A., and Chakravarty, S. (2020). Molecular Basis of Sex Difference in Neuroprotection Induced by Hypoxia Preconditioning in Zebrafish. *Mol. Neurobiol.* 57 (12), 5177–5192. [pii]. doi:10.1007/s12035-020-02091-1

Das, T., Soren, K., Yerasi, M., Kumar, A., and Chakravarty, S. (2019). Revealing Sex-specific Molecular Changes in Hypoxia-Ischemia Induced Neural Damage and Subsequent Recovery Using Zebrafish Model. *Neurosci. Lett.* 712, 134492, 2019 . S0304-3940(19)30595-6 [pii]. doi:10.1016/j.neulet.2019.134492

Dugan, L. L., and Choi, D. W. (1999). "Hypoxia-Ischemia and Brain Infarction," in *Basic Neurochemistry: Molecular, Cellular and Medical Aspects*. GJ AB Siegel RW Albers, et al. (Philadelphia: Lippincott-Raven).

Durukan, A., and Tatlisumak, T. (2007). Acute Ischemic Stroke: Overview of Major Experimental Rodent Models, Pathophysiology, and Therapy of Focal Cerebral Ischemia. *Pharmacol. Biochem. Behav.* 87 (1), 179–197. S0091-3057(07)00137-2 [pii]. doi:10.1016/j.pbb.2007.04.015

Goldenberg, N. A., Everett, A. D., Graham, D., Bernard, T. J., and Nowak-Gottl, U. (2014). Proteomic and Other Mass Spectrometry Based "omics" Biomarker Discovery and Validation in Pediatric Venous Thromboembolism and Arterial Ischemic Stroke: Current State, Unmet Needs, and Future Directions. *Proteomics Clin. Appl.* 8 (11-12), 828–836. doi:10.1002/prca.201400062

Goodall, S., Twomey, R., and Amann, M. (2014). Acute and Chronic Hypoxia: Implications for Cerebral Function and Exercise Tolerance. *Fatigue: Biomed. Health Behav.* 2 (2), 73–92. doi:10.1080/21641846.2014.909963

Heinz, U. E., and Rollnik, J. D. (2015). Outcome and Prognosis of Hypoxic Brain Damage Patients Undergoing Neurological Early Rehabilitation. *BMC Res. Notes* 8, 243, 2015 . [pii]. doi:10.1186/s13104-015-1175-z

Hochachka, P. W., Buck, L. T., Doll, C. J., and Land, S. C. (1996). Unifying Theory of Hypoxia Tolerance: Molecular/metabolic Defense and rescue Mechanisms for Surviving Oxygen Lack. *Proc. Natl. Acad. Sci.* 93 (18), 9493–9498. doi:10.1073/pnas.93.18.9493

Huang, D., Cao, L., Xiao, L., Song, J.-x., Zhang, Y.-j., Zheng, P., et al. (2019). Hypoxia Induces Actin Cytoskeleton Remodeling by Regulating the Binding of CAPZA1 to F-Actin via PIP2 to Drive EMT in Hepatocellular Carcinoma. *Cancer Lett.* 448, 117–127. S0304-3835(19)30072-2 [pii]. doi:10.1016/j.canlet.2019.01.042

Johnson, A. B., Denko, N., and Barton, M. C. (2008). Hypoxia Induces a Novel Signature of Chromatin Modifications and Global Repression of Transcription. *Mutat. Res.* 640 (1-2), 174–179. S0027-5107(08)00017-1 [pii]. doi:10.1016/j.mrfmmm.2008.01.001

Klein, J. B., Barati, M. T., Wu, R., Gozal, D., Sachleben, L. R., Jr., Kausar, H., et al. (2005). Akt-mediated Valosin-Containing Protein 97 Phosphorylation Regulates its Association with Ubiquitinated Proteins. *J. Biol. Chem.* 280 (36), 31870–31881. M501802200 [pii]. doi:10.1074/jbc.m501802200

Li, H., You, W., Li, X., Shen, H., and Chen, G. (2019). Proteomic-Based Approaches for the Study of Ischemic Stroke. *Transl. Stroke Res.* 10 (6), 601–606. [pii]. doi:10.1007/s12975-019-00716-9

Lin, L., Yee, S. W., Kim, R. B., and Giacomini, K. M. (2015). SLC Transporters as Therapeutic Targets: Emerging Opportunities. *Nat. Rev. Drug Discov.* 14 (8), 543–560. nrd4626 [pii]. doi:10.1038/nrd4626

Lindeman, L. C., Winata, C. L., Aanes, H., Mathavan, S., Alestrom, P., and Collas, P. (2010). Chromatin States of Developmentally-Regulated Genes Revealed by DNA and Histone Methylation Patterns in Zebrafish Embryos. *Int. J. Dev. Biol.* 54 (5), 803–813. 103081ll [pii]. doi:10.1387/ijdb.103081ll

Litt, M. D., Simpson, M., Gaszner, M., Allis, C. D., and Felsenfeld, G. (2001). Correlation between Histone Lysine Methylation and Developmental Changes at the Chicken Beta Globin Locus. *Science* 293 (5539), 2453–2455. doi:10.1126/science.1064413

Lu, Y., Chu, A., Turker, M. S., and Glazer, P. M. (2011). Hypoxia-induced Epigenetic Regulation and Silencing of the BRCA1 Promoter. *Mol. Cel. Biol.* 31 (16), 3339–3350. MCB.01121-10 [pii. doi:10.1128/MCB.01121-10

Maison, C., Bailly, D., Peters, A. H. F. M., Quivy, J.-P., Roche, D., and Taddei, A. (2002). Higher-order Structure in Pericentric Heterochromatin Involves a Distinct Pattern of Histone Modification and an RNA Component. *Nat. Genet.* 30 (3), 329–334. doi:10.1038/ng843

Michiels, C. (2004). Physiological and Pathological Responses to Hypoxia. *Am. J. Pathol.* 164 (6), 1875–1882. S0002-9440(10)63747-9 [pii]. doi:10.1016/S0002-9440(10)63747-9

Muller, A. J., and Marks, J. D. (2014). Hypoxic-Ischemic Brain Injury: Potential Therapeutic Interventions for the Future. *Neoreviews* 15 (5), e177–e186. doi:10.1542/neo.15-5-e177

Nakayama, J., Rice, J. C., Strahl, B. D., Allis, C. D., and Grewal, S. I. (2001). Role of Histone H3 Lysine 9 Methylation in Epigenetic Control of Heterochromatin Assembly. *Science* 292 (5514), 110–113. doi:10.1126/science.1060118

Panchenko, M. V., Farber, H. W., and Korn, J. H. (2000). Induction of Heme Oxygenase-1 by Hypoxia and Free Radicals in Human Dermal Fibroblasts. *Am. J. Physiology-Cell Physiol.* 278 (1), C92–C101. doi:10.1152/ajpcell.2000.278.1.C92

Peters, A. H., Mermoud, J. E., O'Carroll, D., Pagani, M., Schweizer, D., Brockdorff, N., et al. (2002). Histone H3 Lysine 9 Methylation Is an Epigenetic Imprint of Facultative Heterochromatin. *Nat. Genet.* 30 (1), 77–80. doi:10.1038/ng789

Sekhon, M. S., Ainslie, P. N., and Griesdale, D. E. (2017). Clinical Pathophysiology of Hypoxic Ischemic Brain Injury after Cardiac Arrest: a "Two-Hit" Model. *Crit. Care* 21 (1), 90, 2017 . [pii]. doi:10.1186/s13054-017-1670-9

Shah, F. A., Zeb, A., Ali, T., Muhammad, T., Faheem, M., Alam, S. I., et al. (2019). Identification of Proteins Differentially Expressed in the Striatum by Melatonin in a Middle Cerebral Artery Occlusion Rat Model-A Proteomic and In Silico Approach. *Front. Neurosci.* 12, 888. doi:10.3389/fnins.2018.00888

Silva, E. S., Rocha, J. B., Souza, D. O., and Braga, M. M. (2016). How Does Zebrafish Support New Strategies for Neuroprotection and Neuroregeneration in Hypoxia-Related Diseases? *Neural Regen. Res.* 11 (7), 1069–1070. [pii]. doi:10.4103/1673-5374.187030.NRR-11-1069

Steven, A., Leisz, S., Sychra, K., Hiebl, B., Wickenhauser, C., Mougiakakos, D., et al. (2016). Hypoxia-mediated Alterations and Their Role in the HER-2/neuregulated CREB Status and Localization. *Oncotarget* 7 (32), 52061–52084. 10474 [pii]. doi:10.18632/oncotarget.10474

Sun, M.-K. (1999). Hypoxia, Ischemic Stroke, and Memory Deficits: Prospects for Therapy. *Tbmb* 48 (4), 373–378. doi:10.1080/713803535

Ton, C., Stamatiou, D., and Liew, C. C. (2003). Gene Expression Profile of Zebrafish Exposed to Hypoxia during Development. *Physiol. Genomics* 13 (2), 97–106. 00128.2002 [pii]. doi:10.1152/physiolgenomics.00128.2002

Vakoc, C. R., Sachdeva, M. M., Wang, H., and Blobel, G. A. (2006). Profile of Histone Lysine Methylation across Transcribed Mammalian Chromatin. *Mol. Cel. Biol.* 26 (24), 9185–9195. MCB.01529-06 [pii]. doi:10.1128/MCB.01529-06

van der Meer, D. L., van den Thillart, G. E., Witte, F., de Bakker, M. A., Besser, J., Richardson, M. K., et al. (20052005). Gene Expression Profiling of the Long-Term Adaptive Response to Hypoxia in the Gills of Adult Zebrafish. *Am. J. Physiol. Regul. Integr. Comp. Physiol.* 289 (5), R1512–R1519. [pii]. doi:10.1152/ajpregu.0008910.1152/ajpregu.00089.2005

Weidemann, A., Breyer, J., Rehm, M., Eckardt, K.-U., Daniel, C., Cicha, I., et al. (2013). HIF-1α Activation Results in Actin Cytoskeleton Reorganization and Modulation of Rac-1 Signaling in Endothelial Cells. *Cell Commun. Signal* 11, 80, 2013. 1478-811X-11-80 [pii]. doi:10.1186/1478-811X-11-80

Yu, X., and Li, Y. V. (2013). Neuroprotective Effect of Zinc Chelator DEDTC in a Zebrafish (*Danio rerio*) Model of Hypoxic Brain Injury. *Zebrafish* 10 (1), 30–35. doi:10.1089/zeb.2012.0777

Zhang, Z., Yao, L., Yang, J., Wang, Z., and Du, G. (2018). PI3K/Akt and HIF-1 S-ignaling P-athway in H-ypoxia-ischemia (Review). *Mol. Med. Rep.* 18 (4), 3547–3554. doi:10.3892/mmr.2018.9375

**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