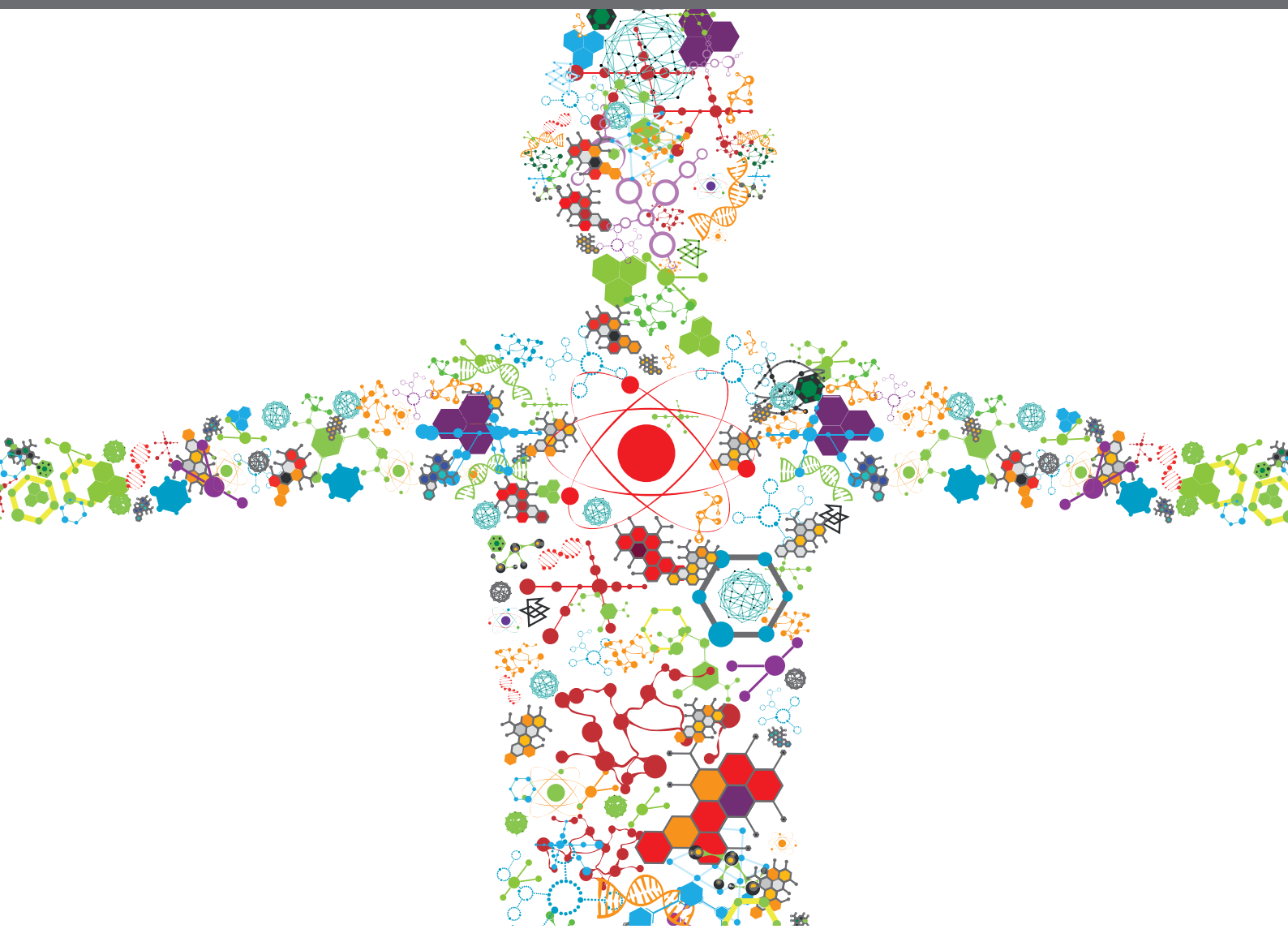
MULTI-OMICS APPROACHES TO STUDY SIGNALING PATHWAYS

EDITED BY: Jyoti Sharma, Lavanya Balakrishnan, Sandeep Kaushik and Manoj Kumar Kashyap

PUBLISHED IN: Frontiers in Bioengineering and Biotechnology and Frontiers in Genetics



frontiers Research Topics

frontiers

ISSN 1664-8714
ISBN 978-2-88966-125-1
DOI 10.3389/978-2-88966-125-1

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: researchtopics@frontiersin.org

MULTI-OMICS APPROACHES TO STUDY SIGNALING PATHWAYS

Topic Editors:
Jyoti Sharma, Institute of Bioinformatics (IOB), India
Lavanya Balakrishnan, Mazumdar Shaw Medical Centre, India
Sandeep Kaushik, University of Minho, Portugal
Manoj Kumar Kashyap, Amity University Gurgaon, India

Citation: Sharma, J., Balakrishnan, L., Kaushik, S., Kashyap, M. K., eds. (2020). Multi-Omics Approaches to Study Signaling Pathways. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-88966-125-1

Table of Contents

frontiers in Bioengineering and Biotechnology

EDITORIAL
published: 04 September 2020
doi: 10.3389/fbioe.2020.00829

Editorial: Multi-Omics Approaches to Study Signaling Pathways

Jyoti Sharma[1,2]*, *Lavanya Balakrishnan*[3], *Sandeep Kaushik*[4] and *Manoj Kumar Kashyap*[5]*

[1] Institute of Bioinformatics, International Technology Park, Bengaluru, India, [2] Manipal Academy of Higher Education, Manipal, India, [3] Mazumdar Shaw Center for Translational Research, Bengaluru, India, [4] 3B's Research Group, University of Minho, Guimarães, Portugal, [5] Amity Medical School, Amity Stem Cell Institute, Amity University Haryana, Gurugram, India

Keywords: omics, transcriptomics, proteomics, genomics, microRNA, targeted therapy, signaling

Editorial on the Research Topic

Multi-Omics Approaches to Study Signaling Pathways

With the advent of omics technologies, tremendous progress has been made in understanding the signaling pathways in normal and disease states across different species. Multi-omics approaches can be categorized into two groups: molecular profiling (MPro) and molecular perturbation (MPer) (Yao et al., 2015). The MPro grouping includes the profiling of genomic, transcriptomic, proteomic, post-translational modifications, and interactome; and MPer includes genetic and functional perturbations.

Omics approaches such as genomics, transcriptomics, miRNAomics, proteomics, and metabolomics have changed the landscape of different diseases including stroke, diabetes, and cancer. Genomic approaches such as genome wide association studies have led to the identification of 30 loci, which were used in swaying body mass index and the risk of obesity (McCarthy, 2010). At mRNA levels, transcriptomic profiling using cDNA microarrays has helped not only in detecting the downregulation of significant tumor suppressors in breast cancer metastasis (Zheng et al., 2017) but also enabled medical practitioners to discriminate patients with activated B-like diffuse large B-cell lymphoma (DLBCL) from those with germinal center B-like DLBCL (Alizadeh et al., 2000). High-throughput studies focused on microRNAs (miRNAs or miRs) in early stage breast cancer have led to the identification of unique predictive miR signatures specific to ER, PR, and HER2 status (Lowery et al., 2009). At the protein level, an *in vivo* labeling technique like stable isotope labeling with amino acids in animal cell culture, coupled with a mass-spectrometry based proteomic approach, has allowed for the comparison of different mutations in lung adenocarcinoma cell lines in relation to EGFR signaling (Guha et al., 2008).

For this special issue, we present a collection of 12 articles, which provide a comprehensive overview of the different biological pathways within the MPro and MPer approaches.

WGS approaches have been extensively used to unravel the different types of genomic alterations in cancer and facilitate understanding of the mutational landscape of cancer genomes. Using WGS, Dr. Bandapalli's group identified candidate predisposing genes in families with a reported recurrence of Hodgkin-lymphoma (HL), a lymphoproliferative malignancy of B-cell origin. These variants were prioritized using an in-house pipeline "FCVPPv2." The authors used this pipeline along with gene/variant panels based on cancer predisposing genes and variants prioritized in the largest familial HL cohort study reported to date, to identify high penetrance germline

OPEN ACCESS

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

***Correspondence:**
Jyoti Sharma
jyoti@ibioinformatics.org
Manoj Kumar Kashyap
mkkashyap@ggn.amity.edu

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

Received: *24 June 2020*
Accepted: *29 June 2020*
Published: *04 September 2020*

Citation:
Sharma J, Balakrishnan L, Kaushik S and Kashyap MK (2020) Editorial: Multi-Omics Approaches to Study Signaling Pathways. Front. Bioeng. Biotechnol. 8:829. doi: 10.3389/fbioe.2020.00829

variants in the HL families. Furthermore, pathway and network analyses of these variants have provided additional molecular cues relevant to the molecular pathogenesis of HL that may aid the development of targeted therapy and the screening of individuals who are at risk of developing HL (Srivastava et al.).

MicroRNAs play a key role(s) in regulating gene expression via either degradation of a transcript or the inhibition of its translation. Dr. Skoblov's group presents initial insights into the complexity of human microRNA-mRNA interactions. They performed a comprehensive computational analysis on HEK293 and Huh7.5 datasets and reported interesting features of human mRNA interactome, along with more than 46,000 experimentally confirmed mRNA-miRNA duplex regions. As part of this study, they also developed a web-based tool, publicly available at http://score.generesearch.ru/services/mirna/ (Plotnikova et al.).

In addition, Dr. Fang's group performed a comprehensive analysis of hsa-let-7i-5p miRNA in normal and pathogenic fibroblasts and studied its role ranging from controlling cellular phenotype to molecular signaling particularly TGF-β signaling (Zhang et al.).

In a transcriptome study, Dr. Evelo's group performed computational analysis on transcriptomic data derived from cattle breeds with different intramuscular fat deposition to identify pathways that define marbling in beef cattle. A total of 17 pathways were significantly dysregulated between well-marbled vs. lean-marbled beef including MAPK and insulin signaling, and immune response associated pathways (Roudbari et al.).

Dr. Guan's group carried out a genome-wide analysis in gastric cancer and identified 548 and 2,399 differentially methylated sites and lncRNAs, respectively. The lncRNAs were able to discriminate between normal vs. cancerous samples of gastric origin (Song et al.).

Dr. Jolly's group carried out a comprehensive and comparative analysis of methods utilizing different transcriptomics signatures to quantify the status of EMT—a cell biological process involved in cancer metastasis and chemoresistance. They showed that these methods exhibited a concordance among themselves in quantifying the extent of EMT in a given sample and that tumor cells can undergo varying degrees of EMT across tumor types. While any of the three methods can capture the generic trend in the EMT status of a given cell (or population), the multinomial logistic regression EMT scoring method has an additional advantage of being able to predict from the transcriptomic signature of a population, whether it is comprised of "pure" single hybrid E/M cells at the single-cell level, or an ensemble of E and M cell subpopulations (Chakraborty et al.).

Another study, by Dr. Imhof's group in the context of hemolytic disorders, described a new ontology and knowledge graph "HemeKG," which is publicly available at https://github.com/hemekg/hemekg. This resource assembles heme-specific terms to better categorize, organize, and analyze data on the effects of heme on cell biological and biochemical pathways (Humayun et al.).

In a study on the brain disorder Schizophrenia, Dr. Lane's group used an ensemble boosting predictive framework along with random undersampling, to assess the status of schizophrenia in the population of Taiwan by examining the levels of D-amino acid oxidase protein and its interaction partner, the D-amino acid oxidase activator, in the N-methyl-D-Aspartate receptor pathway, as well as by using melatonin levels in the tryptophan catabolic pathway. They also evaluated the performance of the ensemble boosting algorithm and compared it with other widely used machine learning algorithms, including support vector machine, and multi-layer feedforward neural networks. Notably, they showed that it performs better in distinguishing schizophrenia patients from healthy controls (Lin et al.).

Using an integrative approach, Dr. Domingo-Fernandez and his group demonstrated that the choice of pathway database could impact the results of statistical enrichment analysis and predictive modeling. They also developed an integrative pathway resource called "MPath" which showed that using multiple pathway databases or integrated resources could provide more biologically consistent results and improved prediction performances, as opposed to using equivalent pathways from different databases (Mubeen et al.).

Dr. Zayed's group employed an integrative and systematic bioinformatics in treating ovarian cancer, and identified not only the DEGs involved in the cell cycle, but also the hub genes including core genes (*FZD6, FZD8, CDK2*, and *RBBP8*) strongly linked to OC. A large majority of Frizzled receptors including *FZD6* and *FZD8* were involved in the ß-catenin canonical signaling pathway (Udhaya Kumar, Kumar, Siva, Doss, Zayed et al.).

In a study on the autoimmune disease SLE, Dr. Zayed's group used a high-throughput transcriptomics platform to identify dysregulated signaling pathways. They found that four genes including *EGR1, CD38, CAV1,* and *AKT1* were strongly associated with pathways in SLE (Udhaya Kumar, Kumar, Siva, Doss, Younes et al.).

In a study on *M. tuberculosis,* Dr. Gupta's group studied Rv1915/ICL2a protein as there has been difficulty in harvesting the soluble protein. They overcome this by expression of C-terminal truncated Rv1915/ICL2a in the heterologous host *E. coli* BL21 (DE3) (Antil et al.).

This collective effort brings together studies covering models from prokaryotic to eukaryotic organisms using different omics approaches to delineate either signaling or molecules directly or indirectly related to signaling in cancer and/or other diseases and aberrant conditions.

Abbreviations: DEGs, differentially regulated genes; DLBCL, diffuse large B-cell lymphoma; EMT, epithelial-Mesenchymal transition; FCVPPv2, familial cancer variant prioritization pipeline; GWAS, genomewide association studies; HL, Hodgkin-lymphoma; miRNA or miR, microRNA; MPro, molecular profiling; MPer, molecular perturbation; NMDAR, N-methyl-D-Aspartate Receptor; SLE, systemic lupus erythematosus; WGS, whole genome sequencing.

AUTHOR CONTRIBUTIONS

JS coordinated the Research Topic. MKK coordinated the editorial. JS, LB, SK, and MKK contributed to the development of the Research Topic, suggested and invited the participants, and

helped with the peer review process. All authors have approved the final version of the editorial.

FUNDING

JS is a recipient of Bio-CARe Women Scientists award by the Department of Biotechnology (DBT), Government of India (Grant # BT/PR19924/BIC/101/568/2016). LB and MKK are the recipients of the National-Post doctoral fellowship (Grant # PDF/2017/002992) & TARE fellowship (Grant # TAR/2018/001054), respectively from the Science and Engineering Research Board (SERB), Department of Science and Technology, Government of India.

ACKNOWLEDGMENTS

We acknowledge the contributions of all the participating authors for this Research Topic.

REFERENCES

Alizadeh, A. A., Eisen, M. B., Davis, R. E., Ma, C., Lossos, I. S., Rosenwald, A., et al. (2000). Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* 403, 503–511. doi: 10.1038/35000501

Guha, U., Chaerkady, R., Marimuthu, A., Patterson, A. S., Kashyap, M. K., Harsha, H. C., et al. (2008). Comparisons of tyrosine phosphorylated proteins in cells expressing lung cancer-specific alleles of EGFR and KRAS. *Proc. Natl. Acad. Sci. U.S.A.* 105, 14112–14117. doi: 10.1073/pnas.0806158105

Lowery, A. J., Miller, N., Devaney, A., McNeill, R. E., Davoren, P. A., Lemetre, C., et al. (2009). MicroRNA signatures predict oestrogen receptor, progesterone receptor and HER2/neu receptor status in breast cancer. *Breast Cancer Res.* 11:R27. doi: 10.1186/bcr2257

McCarthy, M. I. (2010). Genomics, type 2 diabetes, and obesity. *N. Engl. J. Med.* 363, 2339–2350. doi: 10.1056/NEJMra0906948

Yao, Z., Petschnigg, J., Ketteler, R., and Stagljar, I. (2015). Application guide for omics approaches to cell signaling. *Nat. Chem. Biol.* 11, 387–397. doi: 10.1038/nchembio.1809

Zheng, T., Wang, A., Hu, D., and Wang, Y. (2017). Molecular mechanisms of breast cancer metastasis by gene expression profile analysis. *Mol. Med. Rep.* 16, 4671–4677. doi: 10.3892/mmr.2017.7157

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: 08 October 2019
doi: 10.3389/fgene.2019.00933

Comprehensive Analysis of Human microRNA–mRNA Interactome

Olga Plotnikova[1,2]*, *Ancha Baranova*[2,3] *and Mikhail Skoblov*[2]

[1] Laboratory of Functional Genome Analysis, Moscow Institute of Physics and Technology, Moscow, Russia, [2] Laboratory of Functional Genomics, Research Centre for Medical Genetics, Moscow, Russia, [3] School of Systems Biology, George Mason University, Fairfax, VA, United States

MicroRNAs play a key role in the regulation of gene expression. A majority of microRNA–mRNA interactions remain unidentified. Despite extensive research, our ability to predict human microRNA-mRNA interactions using computational algorithms remains limited by a complexity of the models for non-canonical interactions, and an abundance of false-positive results. Here, we present the landscape of human microRNA–mRNA interactions derived from comprehensive analysis of HEK293 and Huh7.5 datasets, along with publicly available microRNA and mRNA expression data. We show that, while only 1–2% of human genes were the most regulated by microRNAs, few cell line–specific RNAs, including EEF1A1 and HSPA1B in HEK293 and AFP, APOB, and MALAT1 genes in Huh7.5, display substantial "sponge-like" properties. We revealed a group of microRNAs that are expressed at a very high level, while interacting with only a few mRNAs, which, indeed, serve as their specific expression regulators. In order to establish reliable microRNA-binding regions, we collected and systematically analyzed the data from 79 CLIP datasets of microRNA-binding sites. We report 46,805 experimentally confirmed mRNA–miRNA duplex regions. Resulting dataset is available at http://score.generesearch.ru/services/mirna/. Our study provides initial insight into the complexity of human microRNA–mRNA interactions.

Keywords: microRNA, regulation of gene expression, microRNA–mRNA interactions, microRNA-binding sites, miRNA-target RNA duplexes, web tool for searching microRNA-binding regions

OPEN ACCESS

Edited by:
Sandeep Kaushik,
University of Minho, Portugal

Reviewed by:
Shaveta Kanoria,
Wadsworth Center, United States
David L. Corcoran,
Duke University, United States

***Correspondence:**
Olga Plotnikova
plotnikova@phystech.edu

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

Received: *21 May 2019*
Accepted: *05 September 2019*
Published: *08 October 2019*

Citation:
Plotnikova O, Baranova A and Skoblov M (2019) Comprehensive Analysis of Human microRNA–mRNA Interactome. Front. Genet. 10:933. doi: 10.3389/fgene.2019.00933

INTRODUCTION

MicroRNAs are small noncoding RNAs that associate with Argonaute (*AGO*) protein to form a silencing complex, which then regulates a gene expression (Jonas and Izaurralde, 2015). MicroRNAs accomplish essential post-transcriptional regulatory step of gene expression regulation through either the degradation of a transcript or the inhibition of translation and are involved in key cellular processes, such as apoptosis, proliferation, or differentiation (He and Hannon, 2004). Hence, dysregulation of microRNAs may result in the development of a disease or in a malignant transformation (Weiss and Ito, 2017). According to some estimates, nearly all mature sequences

Abbreviations: AGO, Argonaute; CDS, coding DNA sequence; CLASH, cross-linking, ligation, and sequencing of hybrids technique; CLEAR-CLIP, covalent ligation of endogenous Argonaute-bound RNA-CLIP technique; CLIP, UV cross-linking and immunoprecipitation technique; Exp-MiBRs, experimentally confirmed microRNA-binding regions; HITS-CLIP, high-throughput sequencing of RNA isolated by cross-linking immunoprecipitation; iCLIP, individual-nucleotide resolution cross-linking and immunoprecipitation; PAR-CLIP, photoactivatable-ribonucleoside-enhanced immunoprecipitation; UTR, untranslated region.

of coding transcripts contain potential sites for microRNA regulation (Bartel, 2004; Friedman et al., 2009).

Human genome encodes approximately 2,600 mature microRNAs (miRBase v.22) and, according to GENCODE data (v.29), more than 200,000 of transcripts, including isoforms with slight variations. A particular microRNA may target many different mRNAs (Selbach et al., 2008); a particular messenger RNA may bind to a variety of microRNAs, either simultaneously or in context-dependent fashion (Uhlmann et al., 2012). Notably, the target regions for particular microRNAs commonly cluster together, thus resulting in the cooperative repression effect (Grimson et al., 2007; Sætrom et al., 2007). The mapping of microRNA–mRNA interactions is far from being complete due to the recognized challenge of computational prediction of mRNA–microRNA interactions.

In our previous study, we showed that the outputs generated by commonly used microRNA–mRNA interactions predicting software differ substantially, while failing to pinpoint experimentally confirmed microRNA-binding regions correctly (Plotnikova and Skoblov, 2018). Nowadays, many tools for the prediction microRNA–mRNA interactions are in development, all with different underlying algorithms (Agarwal et al., 2015; Gumienny and Zavolan, 2015; Lu and Leslie, 2016; Riffo-Campos et al., 2016). Among most advanced algorithms, we should highlight the ones taking into account expression levels of both the microRNAs and their targets. Notably, the changes in expression of microRNA may also affect expression levels of other, non-target mRNAs—for example, due miRNA targeting of their upstream regulators. Consequently, newer, more comprehensive approaches—for example, MiRImpact (Artcibasova et al., 2016), PanMiRa (Li and Zhang, 2014), and ProMISe (Li et al., 2014), aim at explaining complex phenotypes by performing analysis of each microRNA along with its direct and indirect targets.

Experimental identification of direct microRNA targets remains a crucial step in attaining reliable prediction results. There are two main groups of the experimental approaches for a direct identification of microRNA–mRNA interactions. The first approach relies on a construction of reporter gene assays and one-by-one evaluation of possible interactions between the microRNA and its cognate mRNA region of interest through measuring the activity of the reporter (Steinkraus et al., 2016). Another group of techniques comprises involves a coupling of a cross-linking with immunoprecipitation (CLIP); this group represented by variety of the protocols including PAR-CLIP, iCLIP, HITS-CLIP, and others (Licatalosi et al., 2008; Steinkraus et al., 2016). CLIP group of methods identifies the microRNA-binding regions in target mRNAs only, while information about pairing of a particular microRNA with a particular mRNA region remains obscure.

Two modifications of AGO-CLIP based technology were developed specifically for identifying microRNAs ligated to their endogenous mRNA targets as part of chimeric molecules. To date, AGO-CLIP-based evaluations of microRNA–mRNA interactomes were executed only in two human cell lines. Helwak and colleagues applied so-called cross-linking ligation and sequencing of hybrids, or CLASH, to HEK293 cell line, retrieving more than 18,000 high-confidence microRNA–mRNA interactions (Helwak et al., 2013). Later, Moore and colleagues used another variety of AGO-CLIP termed CLEAR (covalent ligation of endogenous Argonaute-bound RNAs)-CLIP for the study of microRNA interactome in Huh7.5 cell (Moore et al., 2015). CLASH and CLEAR-CLIP techniques closely resemble each other, with the only difference that CLASH protocol employs HEK293 cell line over-expressed AGO1, while CLEAR-CLIP targets endogenous AGO allowing experimenting with any cell line. Thus, CLEAR-CLIP does not require full denaturation of AGO and involves a single purification step. It is of note that both publications cited above concentrated on the development of the experimental protocol and subsequent evaluation of the technical aspects of analytic procedure, rather than on extracting biological insights from the data collected.

A flowchart at **Figure 1A** represents the methodology for analysis of microRNA–mRNA human interactome employed in this study. We aggregated various experimental data on human miRNA–mRNA interactions and then investigated how expression levels of each studied microRNA and each of its cognate mRNAs correlate, and whether the behavior of miRNA–mRNA pairs depends on a cell line context. In order to do this, we analyzed together (i) sequences and abundance of microRNA and their target mRNAs in CLASH dataset for HEK293 cell line and in CLEAR-CLIP dataset for Huh7.5 cell line, and (ii) expression level of microRNAs and target mRNAs in HEK293 and in Huh7.5 cell lines. Second, we performed systematic extraction of credible, experimentally confirmed

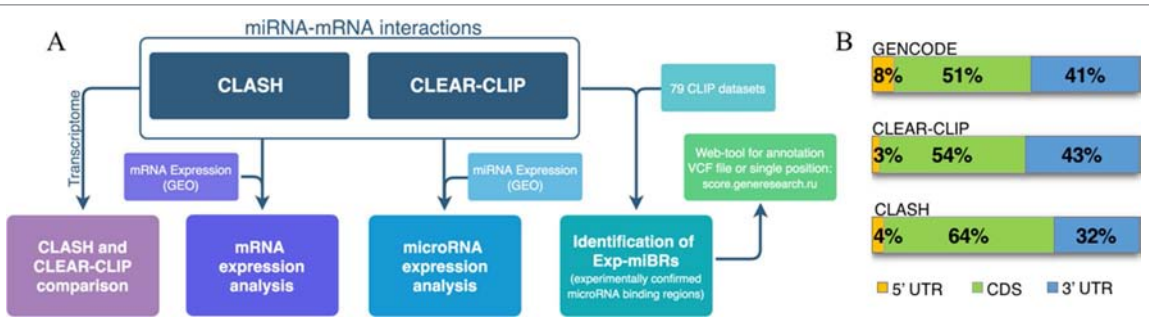


FIGURE 1 | Analysis of human microRNA–mRNA interactome. **(A)** Flowchart describing main steps of human microRNA–mRNA interactome data analysis. **(B)** Distribution of the summarized lengths of 3'UTR, CDS, or 5'UTR mRNA regions in CLEAR-CLIP, CLASH, and GENCODE, respectively.

microRNA-binding regions across CLASH/CLEAR-CLIP datasets and in 79 additional CLIP datasets and present them here as a collection.

MATERIALS AND METHODS

microRNA–mRNA Interactions

microRNA–mRNA interactome data were extracted from previously published CLASH (Helwak et al., 2013) and CLEAR-CLIP (Moore et al., 2015) datasets. CLASH data provide transcriptome coordinates for 18,514 miRNA–mRNA interactions, while CLEAR-CLIP dataset include genome coordinates (version hg18) for 32,712 interactions. Using Ensembl API (https://rest.ensembl.org/, Yates et al., 2014), the coordinates of CLASH microRNA–mRNA-interacting regions were transformed into genome coordinates. For 36 interactions, the transforming of their coordinates failed and, in total, we revealed 18,478 microRNA–mRNA interactions in 22,030 genome regions (all interactions were located in mRNA regions, with 19% being divided between two exons and 36 interactions of three exons). We used LiftOver (https://genome.ucsc.edu/cgi-bin/hgLiftOver, Kuhn et al., 2012) to transform CLEAR-CLIP interactome data from hg18 genome version into hg19. A total of two interactions failed to transform. Hence, resultant amount of interactions equaled 32,710. Genomic regions (CDS, 3'UTR, 5'UTR, intronic, intergenic, etc.) were annotated by wAnnovar (Wang et al., 2010; Yang and Wang, 2015).

To compare CLASH and CLEAR-CLIP data, CLEAR-CLIP dataset was reduced to microRNA–mRNA interactions mapped to the expressed transcriptome (n = 10,032). For each of the sets of genomic regions (3'UTR, CDS, and 5'UTR) found in miRNA bound regions present in CLASH and CLEAR-CLIP, their average length (mean) was comparable to that calculated for all protein-coding transcripts (N = 59,900) downloaded from GENCODE, version 24 (Frankish et al., 2018).

To calculate expected overlap between CLASH and CLEAR-CLIP datasets, five independent CLASH-like and CLEAR-CLIP-like datasets were generated. For each simulation, binding regions were randomly selected from CLASH/CLEAR-CLIP transcripts in amounts equal to detected amount of interactions.

CLASH and CLEAR-CLIP datasets were utilized to evaluate the amount of interactions for each of the genes as a sum of all interactions between microRNAs and mRNA encoded by each gene.

mRNA Expression

Publicly available RNAseq datasets GSE68611 (Murakawa et al., 2015) and GSE64677 (Luna et al., 2015) were used for extracting and examining gene sets expressed in HEK293 and Huh7.5 cell lines. Each of these datasets includes two biological replicates. Initial quality control of sequencing outputs was performed using FastQC (www.bioinformatics.babraham.ac.uk/projects/fastqc). Next, we used kallisto (Bray et al., 2016) to map raw reads to the human reference transcript sequences (GENCODE, 28 version).

First, in each experiment, we calculated the gene expression levels as the sum of expression levels for individual gene transcripts. Second, we took the mean value for each gene between two processed datasets in each of the two cell lines. Finally, we kept only genes that had expression more or equal to 1 tpm as total value and that had expression level of at the level at least 1 tpm in one of the two experiments.

In order to compare only genes reliably expressed both in HEK293 and Huh7.5 cells, only the genes expressed at levels of more than 10tpm or higher were included.

Gene functions were interpreted using PANTHER toolkit Version 12.0 (http://www.pantherdb.org/tools). We used InteractiVenn tool (Heberle et al., 2015) to create Venn diagrams in our analysis.

microRNA Expression

We downloaded microRNA expression data from the GEO database: two experimental replicates for HEK293 cell line (GSE75136 (Wissink et al., 2016)) and three experimental replicates for Huh7.5 cell line (GSE74014 (Bandiera et al., 2016)). The correlations of experimental results obtained in two cell lines were calculated by the Spearman's procedure. We used the R package "DeSeq2" (Love et al., 2014) to normalize microRNA expression. Particular microRNA was considered as expressed if its expression levels were of three or more counts.

CLASH and CLEAR-CLIP datasets were used to calculate the amount of interactions for each microRNA. The correlation of the amounts of interactions formed by microRNAs and their expression levels were estimated using the Spearman correlation coefficient.

In order to calculate conservative phyloP scores, for all microRNAs, we downloaded the coordinates of the mature microRNAs from miRBase (Kozomara and Griffiths-Jones, 2014) (release 22, coordinates corresponded to the GRCh38 human reference genome). Next, we used UCSC table browser (Karolchik et al., 2004) to obtain phyloP conservative values across 20 vertebrates for all mature microRNAs. For each group of microRNAs, the mean value between the phyloP scores was calculated.

CLIP Data

We collected 79 CLIP datasets (**Supplementary Table 1**) from the POSTAR database (Hu et al., 2016). Raw data of these CLIP datasets were initially pre-processed by FASTX-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit) and then were processed by specialized tools for different CLIP-seq technologies: PARalyzer (Corcoran et al., 2011) for PAR-CLIP datasets (N = 18) and CIMS (Moore et al., 2014) for HITS-CLIP datasets (N = 61). We used python to analyze all microRNA-binding regions from CLIP datasets together with microRNA–mRNA interactions from CLASH and CLEAR-CLIP. In total, all regions were merged in six million nucleotides, and each position was characterized by the following parameters: list of supported experiments (GEO GSM ID), their corresponding cell lines

and list of interacted microRNAs (if accessible). We used wAnnovar to annotate genes and their parts (CDS, 3'UTR, 5'UTR, intronic, etc.).

microRNA-Binding Regions

Our analysis of CLIPs, CLASH, and CLEAR-CLIP revealed 156,000 regions. We used a custom python script to select experimentally confirmed microRNA-binding regions (Exp-MiBR). Exp-MiBR was defined as a region that had a subsequence of length L = 10, whereas each nucleotide (position) in this subsequence had been supported by at least n = 2 different datasets or chimeras. We estimated the amount of Exp-MiBRs for all combination of length and amount of supported datasets/chimeras in ranges: L = 1–25 and n = 1–10 (**Supplementary Table 2**).

Exp-MiBR Application

We characterized each Exp-MiBR (total amount = 46,805) by the following parameters: gene information, amount and list of supported experiments (GEO GSM ID) and their corresponding cell lines, and list of interacted microRNAs (if accessible).

Besides that all the Exp-MiBRs with the corresponded information are available as **Supplementary Table 3**, we also provide an open-access web tool *via* http://score.generesearch.ru/services/mirna. As input, the tool requires any VCF file (v4.0 or 4.1), no more than 20MB or a single (point) genome coordinate. The file or coordinate could be recorded in human genome assembly version 38 or 19.

Web Tool for Searching Exp-MiBRs

All microRNA-binding regions identified as experimentally confirmed (Exp-MiBR) and reported in this paper (**Supplementary Table 3**) may be searched by a web tool available online: http://score.generesearch.ru/services/mirna/.

RESULTS

Comparison of High-Throughput microRNA–mRNA Interactions From CLASH and CLEAR-CLIP Datasets

First, the sets of microRNA–mRNA interactions retrieved in HEK293 and in Huh7.5 by CLASH (Helwak et al., 2013) and CLEAR-CLIP (Moore et al., 2015) protocols were compared, respectively, to hg19 genome references. Although CLASH and CLEAR-CLIP techniques are somewhat similar, CLEAR-CLIP study (N = 32,710) revealed almost two times more interactions than CLASH study (N = 18,478). One of the reasons for this may be due to the differences in the data processing procedures. While CLASH sequences were aligned to the mature transcriptome, CLEAR-CLIP data have been mapped to human genome. Because of that, CLEAR-CLIP technique was capable to highlight additional interaction sites located in the introns and the intergenic regions (~70% of all interactions).

To enable the comparison, we focused our analysis on miRNA-binding regions residing within the mature transcriptome (**Supplementary Table 4**). Because of that, CLEAR-CLIP dataset was limited to about one-third of its entries (n = 10,032). Further analysis estimated that approximately 2–3% of the total length of all expressed protein-coding transcripts serve as a target for one or another microRNAs in either CLASH or CLEAR-CLIP datasets. In addition, in both datasets, the microRNA-binding regions had similar distribution by mRNA regions (3'UTR, CDS, 5'UTR), and to the distribution of the mRNA parts present in GENCODE (**Figure 1B**). Thus, we conclude that the datasets generated by CLASH and CLEAR-CLIP techniques are comparable.

In the experimentally obtained CLASH and CLEAR-CLIP datasets, we detected 1,153 common miRNA–mRNA interactions, which were built upon combinations of 933 interactions in CLASH and 944 interactions in CLEAR-CLIP. Average length of experimentally obtained interaction was at 37.2 nt +/− 19.4 nt. Eight hundred and sixty-seven interactions which were common for both datasets had the length of overlap of more than 20 nt, with an average length of 45.8 nt +/− 13.9 nt. To evaluate if this overlap reflects biological phenomenon rather than statistical fluke, we performed computational simulation of CLASH and CLEAR-CLIP interactions in transcripts expressed in HEK293 (N = 7,299) and Huh7.5 (N = 4,977), respectively. For these cell lines, a common set of expressed mRNAs (n = 3,044) was reduced to a set of randomly selected nucleotide fragments with the size distribution matching that for nucleotide fragments of CLASH and CLEAR-CLIP; then, we analyzed these sets of sequences for overlap. After five independent runs with randomly selected fragments of matching size distribution, we detected, on average, 7.4 +/− 1.3 interactions with an average length of overlapped segments at 14 nt +/− 6.7 nt. Among these interactions, only a fraction had the length of overlap of more than 20 nt (5.0 +/− 2.5). Therefore, the characteristics of experimentally detected patterns of miRNA–mRNA interactions differ from that of interactions generated by simulation of random events ($P < 0.0001$).

To investigate whether the low degree of the overlap between miRNA–mRNA interactions registered in CLASH and CLEAR-CLIP datasets could be due to low degree of the overlap between HEK293 and Huh7.5 transcriptomes, expression data collected from these two cell lines were downloaded from GEO repository and analyzed. While about half of expressed microRNAs were found in both these cell lines, an overall difference of HEK293 and Huh7.5 specific sets of highly expressed genes was evident (**Supplementary Figures 1A, B**). To find out if cell-specific differences in microRNA–mRNA interactomes are due to cell-specific environment, the relationships between the levels of expression for individual miRNAs and their targets as well as the patterns of interactions for each mRNA and miRNA in the both cell lines were investigated in details.

Expression Analysis of microRNA–mRNA Interactome

mRNA Expression Analysis

To investigate the degree to which cell-specific levels of transcripts depend on respective microRNAs, we compared expression levels of each gene in HEK293 and Huh7.5 cell lines

then cross compared them to sets of experimentally detected microRNA interactions. HEK293 and Huh7.5 cell lines express a total of 15,8k and 14,5k genes, respectively. In each of these two cell lines, approximately 6.9k genes interacted with one or more microRNAs (**Supplementary Figure 1C**). Our analysis highlighted 1–2% of mRNAs with confirmed interactions and no expression detected in respective cell line. We found that only few of these mRNAs had more than 10 interactions each. A majority of them were found to have highly conservative paralogs, which may erroneously align with miRNAs or mRNAs and affect the results of miRNA mapping. A majority of non-expressed mRNAs (about 70%) had only one interaction. It is possible that these mRNAs have been detected as chimeric reads resulting from their protection by AGO protein from ribonucleases. Below, we will describe a few microRNAs that were detected only as a part of chimeras.

In each of these cell lines, a majority of expressed mRNAs (57–59%) did not interact with any microRNA (**Figures 2A, B**). In CLASH and CLEAR-CLIP datasets, there were 215 and 333 high-interacting mRNAs, respectively, with nine or more miRNA interactions for each.

Cell line–specific pie charts built for the miRNA–mRNA interactions per each mRNA were similar. Nevertheless, comparison of the most regulated sets of genes with nine or more interactions each revealed that these sets were cell-line-specific, with only 18 genes in common. These common 18 genes formed in average of 15.7+/−3.2 and 14.1+/− 2.4 interactions with microRNAs in the HEK293 and Huh7.5 cell lines, respectively. Surprisingly, cell line–specific sets of microRNA regulators for each of these genes were completely different. By PANTHER analysis of the common set of genes, we detected enrichment in only one Gene Ontology

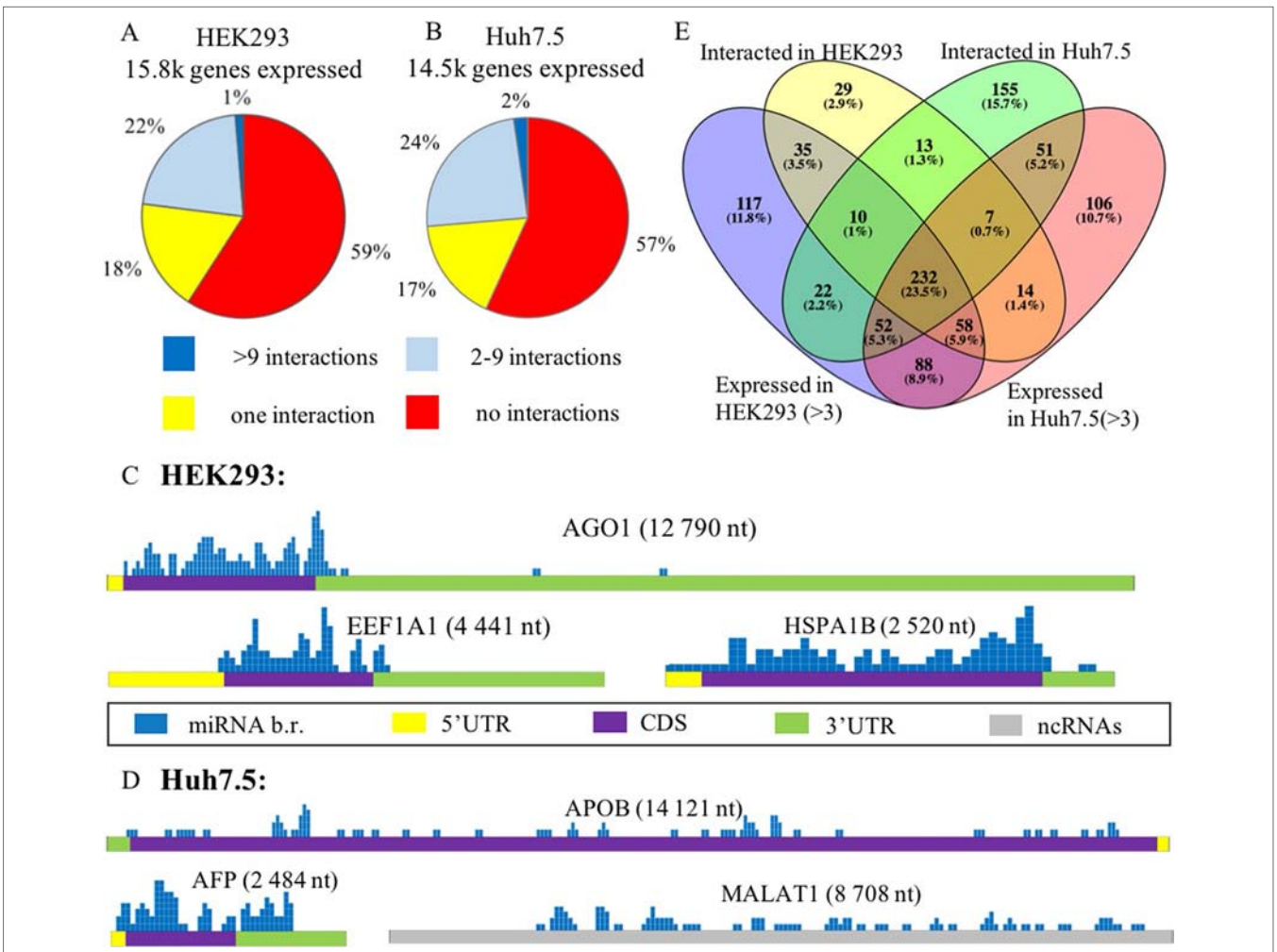


FIGURE 2 | Expression analysis of microRNA and mRNA in HEK293 and Huh7.5 cell lines**. (A)** and **(B)**: Analysis of expressed genes according to amounts of their interactions with microRNAs in HEK293 **(A)** and Huh7.5 **(B)** cell lines. **(C)** and **(D)**: Overview of the microRNA-binding regions locations in sponge-like RNAs expressed in HEK293/CLASH **(C)** and Huh7.5/CLEAR-CLIP **(D)** datasets. After segmenting each of the presented RNAs into 50-nt pieces, the segments that interacted with microRNAs were marked blue on the mRNA map. The height represents the number of interactions detected in each of the segment. For each of the sponge-like RNAs, both name and length are placed above the gene schematics. Colored parts of RNAs are labeled as follows: 5'UTR—yellow, coding region—violet, 3'UTR—green, noncoding region—gray. **(E)** The overlaps between expressed and interacting microRNAs in HEK293 and Huh7.5 cell lines.

(GO) category—a molecular function of RNA binding (**Supplementary Table 5**).

Further, we identified a set of mRNAs capable of interaction with many different types of microRNA molecules, with no preference to a particular miRNA. Such behavior of ambiguous interactions with many microRNAs is similar to that of circular RNAs and lncRNAs with "sponge" properties. Among "sponge-like" mRNAs with 50 or more interactions detected in HEK293/CLASH were those encoded by *AGO1*, *EEF1A1*, and *HSPA1B* genes. Peculiarly, in Huh7.5/CLEAR-CLIP, same property has been attributed to different set of mRNA, namely, *APOB, AFP, MALAT1*, and *XIST*. In mRNAs with "sponge-like" property, microRNA interaction sites were located predominantly in the protein-coding part of the transcript (**Figures 2C, D**).

Remarkably, in HEK293 cells, the most interacting mRNA was the one for AGO1 protein, which had been overexpressed on purpose, as part of CLASH protocol. In this experiment, AGO1-encoding mRNA yielded 88 interactions with a total of 50 different microRNAs. Mean expression levels for AGO1-binding miRNAs were similar to that for all other miRNAs, at 7,279.36 counts *vs*. 7,183.92 counts, respectively. In addition to *AGO1* mRNA, HEK293 cell line expressed two other mRNAs displaying non-specific "sponge-like" effect, *HSPA1B* with 77 interactions to 41 different microRNAs and *EEF1A1* with 50 interactions to 42 microRNAs. Similar to artificially over-expressed AGO1 mRNA, *EEF1A1* also highly expressed in HEK293 cell line (>19K tpm), while another "sponge-like" mRNA *HSPA1B* had expression level equals to 775 tpm.

The set of "sponge-like" mRNAs expressed in Huh7.5 cell line was entirely different. The set of "sponge-like" mRNAs expressed in Huh7.5 cell line was entirely different. We revealed two protein-coding "sponge-like" mRNAS: AFP that formed 47 interactions with 32 different microRNAs, and APOB that also formed 47 interactions with 32 different microRNAs. In set of Huh7.5 "sponge-like" RNAs, two well-described noncoding RNAs were detected: MALAT1 (47 interactions to 27 microRNAs) and XIST (55 interactions to 31 microRNAs). In coherence to expression levels of "sponge-like" mRNAs in HEK293 cell line, we observed difference in expression levels for these mRNAs: *AFP*—more than 19K tpm, *APOB*—358 tpm, *XIST*—202 tpm, and *MALAT1*—80 tpm, while the averages for a gene expressed in Huh7.5 were at 69 tpm.

Comparative Analysis of microRNA Expression Levels and Their mRNA-Interacting Properties

To assess the role of microRNAs in the regulation of their target mRNAs, we studied two HEK293 and three Huh7.5 miRNA profiles retrieved from RNAseq datasets deposited in GEO (GSE75136 and GSE74014). For each cell line, only high-quality datasets with very high correlation of miRNA-specific expression levels were selected (Pearson's correlation $r > > 0.99$). For each miRNA, we analyzed their cell-line specific levels of expression by R package "DeSeq2" in order to normalize miRNA expression and compared these levels to the sets of experimentally detected microRNA–mRNA interactions retrieved from HEK293/CLASH, and Huh7.5/CLEAR-CLIP datasets microRNA was considered as expressed if it had expression levels of more than three counts (see Methods). Less than a quarter (23.5%) of 989 detected miRNAs was present in both cell lines (**Figure 2E**, **Supplementary Table 6**). Notably, many microRNAs expressed in the HEK293 (N = 205) and Huh7.5 (N = 194) cell lines then failed experimental detection as mRNA-interacting molecules in CLASH or CLEAR-CLIP, respectively.

On the other hand, both CLASH and CLEAR-CLIP datasets included 4–17% of mRNA-interacting microRNAs not detected in respective RNAseq datasets at all. On average, these microRNAs had relatively small amounts of interactions: 2.2+/−0.6 interacting partners for 197 microRNAs present in CLASH dataset but absent in HEK293-based RNAseq and 5.1+/−2.2 interacting partners for 168 miRNAs present in CLEAR-CLIP dataset but absent in Huh7.5-based RNAseq. For comparison, mean amounts of detected interactions across all microRNAs were at 55.8 +/−12.7 for 398 miRNAs of HEK293/CLASH and at 143.5 +/− 28.5 for 542 miRNAs in Huh7.5/CLEAR-CLIP. We could expect that these miRNAs could possibly have a low expression level and, therefore, had not reached a detection cut-off in RNAseq. Alternatively, these miRNAs may be somehow protected from degradation by RISC.

Next, for each cell line, we kept only expressed and interacted microRNAs and evaluated their cell-specific expression level and the amount of interactions in this cell line (**Supplementary Figure 2**). For each cell line, Spearman correlation levels were quite low, at 0.18 and 0.29 in HEK293 (N = 335) and Huh7.5 (N = 342), respectively. For each miRNA, we calculated the cell line–specific ratios (R) of its expression level to amount of detected interactions (**Supplementary Table 6**).

Detailed analysis of this data allowed us to highlight two interesting types of miRNA. Type 1 comprised microRNAs with high expression level and relatively small amount of interactions with respective mRNAs. When the cut-offs for both R and expression levels were set as ranking at 90th percentile or higher, only 16 miRNAs for HEK293 (expression > 4,418 and ratio > 252) and 12 miRNAs in Huh7.5 (expression > 6,941 and ratio > 209) were classified as type 1. Notably, eight type 1 miRNAs were present in both cell lines examined.

Type 2 microRNAs were characterized by a low R, and many detected interactions with mRNAs. When the cut-off for R was set as ranking at 10th percentile or lower, and amounts of interactions at 90th percentile or higher, only 11 and 6 miRNAs for HEK293 (amount of interactions > 150 and ratio < 0.9) and Huh7.5 (amount of interactions > 165 and ratio < 2.5), respectively, were classified as type 2. Unlike the type 1 microRNAs, type 2–specific sets from HEK293 and Huh7.5 did not overlap.

In order to evaluate whether these types of microRNAs are evolutionarily constrained, for all mature microRNAs from miRBase, we calculated the mean of the phyloP conservative values in 20 vertebrates. The average cell line–specific phyloP scores for the type 1 and type 2 microRNAs were similar, at 0.99 and 0.95, respectively. Notably, these scores were higher than the average score value calculated for all known microRNAs (0.24), and the score values for all microRNAs that were identified as expressed or interacted in HEK293 or Huh7.5 cell lines (0.74 and 0.71, respectively). Notably, 80% of top 100 miRBase microRNAs with the highest conservative phyloP scores were seen either

as expressed or interacted (or both) in at least one of these two cell lines. On average, in HEK293 and Huh7.5 cells, these most conservative microRNAs had two times higher expression levels than less conservative expressed microRNAs (**Supplementary Table 6**). Overall, higher than average conservativeness of type 1 and type 2 microRNAs may point at the relative importance of their functions.

Comparing Cellular Contexts for microRNA's Interactions

As expected, a majority of microRNAs were concordant in two cell lines: their expression levels and amounts of mRNA interactions were similar in both cellular contexts (**Supplementary Figure 3A**). Nevertheless, some miRNAs have demonstrated remarkable cell specificity in their ratios R (**Supplementary Figures 3B, C**).

For 30 microRNAs, we detected high concordance between their expression level and amount of experimentally detected interactions. Eighteen of these miRNAs had higher expression and mRNA-binding activity in Huh7.5 cell line, while for 12 remaining microRNA, both mRNA-binding activity and expression level were higher in HEK293 cells (**Supplementary Figure 3B**). As an example, in Huh7.5 cell line, expression levels of MAPK1-repressing hsa-miR-194-5p (Kong et al., 2018) were 89 times higher than that in HEK293 cells; in Huh7.5 cells, this microRNA displayed 336 interactions, while in HEK293, it formed only 7 interactions. On the other hand, in HEK293, expression levels of lanosterol synthase suppressing miRNA hsa-miR-10a-5p (Kim et al., 2018) were 450 times higher than that in Huh7.5 cells. In HEK293 cells, this microRNA displayed 267 interactions, while in Huh7.5, it formed only 8 interactions. Such observations were expectable: microRNAs with higher expression level may be capable of the binding to a larger repertoire of targets.

Peculiarly, a total of four microRNAs have performed in exactly opposite way: in cells with higher expression levels, these microRNAs displayed lesser amounts of interactions with their mRNA targets (**Supplementary Figure 3C**). For example, in Huh7.5 cell line, expression levels of hsa-miR-331-3p and hsa-miR-100-5p were at 1,030 and 916 counts, respectively, while in HEK293, these miRNAs had 65 and 41 expression counts, respectively. However, in both cases, amounts of interactions in Huh7.5 cell line were lesser than that in HEK293 cell line, 47 *versus* 342 partners for hsa-miR-331-3p, and 1 *versus* 30 partners for hsa-miR-100-5p. To investigate if this phenomenon is due to the difference in the cell-specific expression levels of target genes, we performed an analysis of all these targets. This was not the case as well. As an example, only 21 out of 318 individual miRNA targets of hsa-miR-331-3p were active in HEK293 cell line but not detected in Huh7.5.

Analysis of Expanded Set of Experimentally Confirmed microRNA-Binding Regions

Experimentally identified microRNA-binding regions form a promising basis for further queries into the basics of the gene expression regulation and lead to uncovering novel disease-causing mechanisms. To enhance a set of microRNA–mRNA interactions retrieved from CLASH and CLEAR-CLIP studies, we performed the database integration of the data collected in cross-linking with immunoprecipitation (CLIP) experiments that provide information about microRNA-binding regions of target genes but unable to identify mRNA–microRNA pairings.

For this purpose, we collected data from 79 CLIP experiments, comprising 61 HITS-CLIP and 18 PAR-CLIP datasets covering 9 different cell lines, with a majority of these data obtained either in HEK293 (N = 34 datasets) or Huh7.5 (N = 19 datasets) (**Supplementary Table 1**). After combining CLIP datasets with the data of previously mentioned CLASH and CLEAR-CLIP studies, approximately 156,000 unique microRNA-binding regions were catalogued within their respective mRNA targets.

At the next stage, the set of microRNA-binding regions was cleaned up to include only these satisfying following criteria: (i) every position in this microRNA-binding subsequence is supported by evidence from at least two different datasets or two different chimeric sequences and (ii) the length of at least 10 nt (**Figure 3A**, **Supplementary Table 3**). MiRNA-binding subsequences of this kind (N = 46,805) formed a dataset of experimentally confirmed microRNA-binding regions (Exp-MiBR). In this dataset, each Exp-MiBR record includes following attributes: genomic coordinates, gene name, type of mRNA part, and list of GEO GSM IDs for experiments which support this microRNA interaction, cellular context, and the list of interacting microRNAs (if accessible). The criteria for inclusion of individual microRNA-binding regions in Exp-MiBR database are justified by analysis presented in **Supplementary Table 2.**

Exp-MiBR subsequences (N = 46,805) were mapped to approximately 15,000 human genes. About one-half of Exp-MiBRs (48%) were located in 3'UTRs, 24% in a coding part of the gene, 10% in introns, and 6% in intergenic parts. Remaining 10% of the Exp-MiBRs were mapped to non-coding RNAs, being matched to either exonic or intronic regions of these loci. For 8,000 of Exp-MIBRs, at least 1 bound microRNA was confirmed by either CLASH or CLEAR-CLIP data (**Figure 3C**).

Approximately 68% of Exp-MiBRs were 20–40 nt in size, closely matching the mean length (33 nt) for all miRNA-binding regions extracted from CLIPs, CLASH, and CLEAR-CLIP data (**Figure 3B**). The second peak in size distribution of Exp-MiBRs was at 75 to 80 nt, being predominantly comprised (86%) of miRNA-interacting region extracted from CLEAR-CLIP dataset. While the sizes of 99% of these Exp-MiBRs were smaller than 150 nt, a few Exp-MiBRs were much longer than that, while remaining supported by many experiments. The longest Exp-MiBR of 631 nt was formed by the regions confirmed as microRNA-interacting in 54 different experiments in nine different cell lines. In addition, there were a few Exp-MiBRs located closely to each other. Such clusters of Exp-MiBRs with many interacting microRNAs do not display a tendency to any particular region of mRNA, as they may be present in CDS, 3'UTR, 5'UTR, or intergenic regions. As an example, chromosome 2 contains a cluster of Exp-MiBRs covering an area of approximately 1.5 kb in size, which is located between the loci of RNA5-8SP5 and MIR663B genes. According to CLASH and CLEAR-CLIP studies, this cluster of Exp-MiBRs

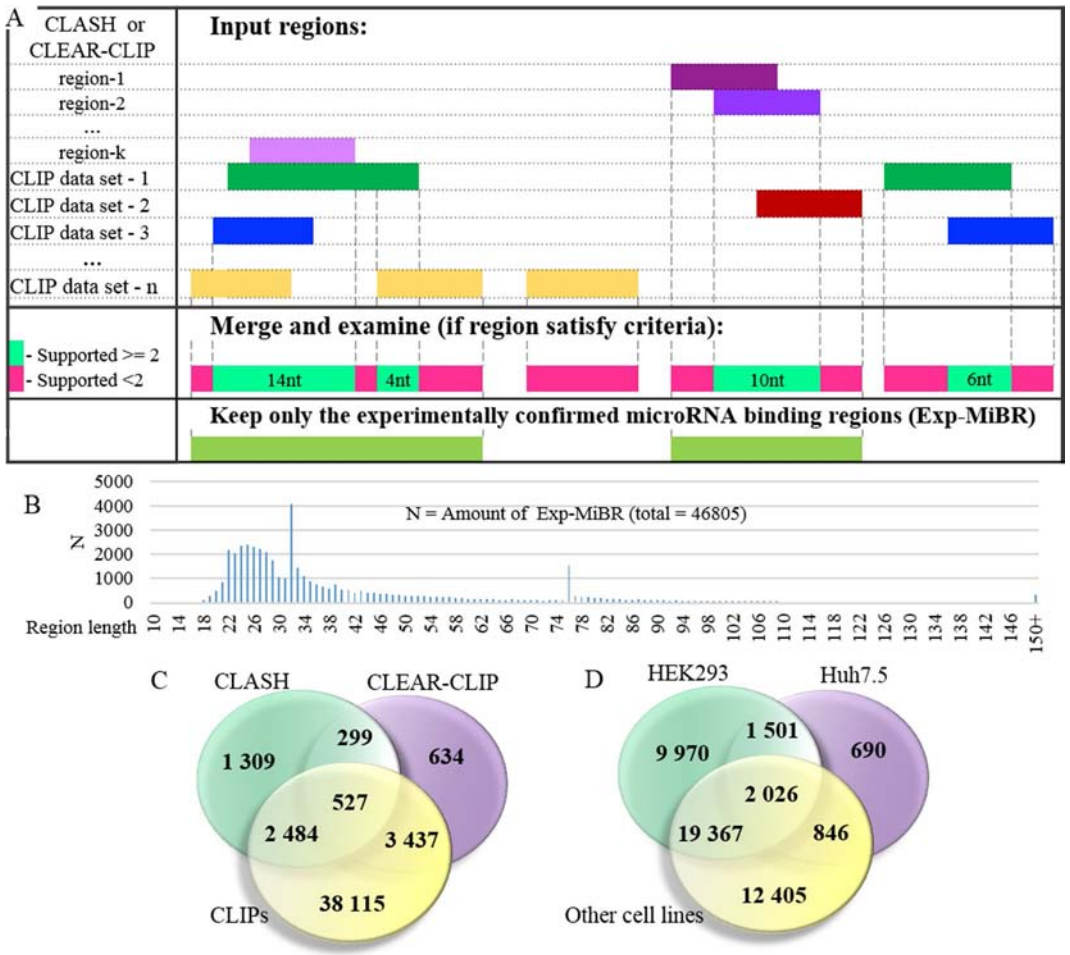


FIGURE 3 | Detailed analysis of experimentally confirmed microRNA-binding regions (Exp-MiBRs). **(A)** Validation of the Exp-MiBR by their independent occurrence in two or more datasets, or in two or more chimeric sequences from one dataset. **(B)** Exp-MiBRs: distribution of the lengths. On horizontal axis—the length of the Exp-MiBRs subsequence; on vertical axis—amounts of the detected Exp-MiBRs (N). **(C)**. Venn diagram depicting Exp-MiBRs detected in experiments employing three different types of identification techniques. **(D)** Venn diagram depicting tissue specificity of Exp-MiBRs detected in HEK293, Huh7.5, and all other cell lines.

interacts with 52 different miRNAs (**Supplementary Figure 4**, **Supplementary Table 7**).

Tissue-Specific and Housekeeping microRNA-Binding Regions

To characterize Exp-MiBRs further, we analyzed their tissue specificity. Most CLIP experiments were performed either in HEK293 (43%) or in Huh7.5 (24%) cells, while the rest of the CLIP data were collected in HeLa, HFF, BC-1, BC-3, EF3D, LCL35, or LCL cells. In HEK293 cells, we found approximately 9,900 unique MiBRs, while analysis of Huh7.5 cells yielded 690 tissue-specific interacting regions (**Figure 3D**). Larger amounts of Exp-MiBRs in HEK293 as compared to that Huh7.5 cells may be explained either by better coverage of HEK293 transcriptome by various CLIPs (**Supplementary Table 1**), or by intrinsic cell-specific features of miRNA interactomes.

Interestingly, some Exp-MiBRs were observed in a majority of studied cells, possibly reflecting a housekeeping function of these interactions. Approximately 1% of all Exp-MiBRs were found in seven or more cell lines. The functional roles of 351 ubiquitous Exp-MiBRs were investigated using Panther software. The GO analysis showed enrichment of genes participating in cellular process of cell cycle (FC 3.17; p-value 1e10−8) and in molecular function of nucleic acid binding (FC 1.75; p-value 5e10−4).

Mitochondrial Regulation by microRNA

An analysis of Exp-MiBRs revealed that these microRNA interacting sequences cover 86% of the mitochondrial genome, including 35 out of 37 mitochondrial genes. Mitochondrial Exp-MiBRs (N = 37) were found in all nine investigated cell lines, with each Exp-MiBR discovered, on average, in 11 independent experiments. In total, we identified 182 miRNAs that bind to various mitochondrial RNAs, with two mitochondrial regions binding 107 out of 182 miRNAs.

DISCUSSION

Experimental identification of microRNA-binding regions is an important prerequisite for querying into the basics of the gene expression regulation, and for uncovering novel disease-causing mechanisms. To date, only two sequencing-based experimental datasets describing full miRNA–mRNA interactomes of human cells, CLASH and CLEAR-CLIP, are available. In both studies, the primary goal was to develop and optimize the experimental protocol itself, while identifying miRNA–mRNA interactions in a particular cell line grown under different conditions. Although these techniques provide a unique window into miRNA targeting, they are not free of limitations, which preclude thorough mapping of entire miRNA–mRNA interactome. Nevertheless, intersecting CLASH and CLEAR-CLIP datasets allowed us to detect much larger set of validated interactions than the overlap of two randomly generated datasets in all five replications. Surprisingly, in both CLASH and CLEAR-CLIP datasets, the distributions of miRNA-binding regions were similar to that in GENCODE transcripts, and more or less even across all types of mRNA regions (3'UTR, CDS or 5'UTR), with no enrichment in miRNA-binding sites within 3'UTRs. Thus, our analysis supports observations of Ragan et al. (Ragan et al., 2009), rather than the model of Grimson et al. (Grimson et al., 2007).

Typically, miRNA–mRNA interaction networks are built *in silico,* with an aid of one or another miRNA prediction tool, and include thousands of mRNA targets. In our study, we attempted painting a holistic picture of human miRNA–mRNA interactome by comparing the entries from experimentally collected datasets describing miRNA-binding activity to the gene expression data. Interestingly, we found that more than half of mRNA transcripts do not bind to any miRNAs present in the same cellular environment. On the other hand, from 1 to 2% of human transcripts interact with nine or more miRNAs, thus, displaying sponge-like activity (Thomson and Dinger, 2016). It was surprising to find that more than half of mRNA transcripts do not bind to any miRNAs present in the same cellular environment. On the other hand, we observed that from 1 to 2% of human transcripts interact with nine or more miRNAs each, thus, displaying sponge-like activity (Thomson and Dinger, 2016). These observations suggest that one can figure out whether some mRNAs may possess such property by analyzing the number of its interactions and the level of its expression: some genes are expressed at a high level but have much fewer interactions than other expressed at same tpm range. This means that each mRNA differs in their miRNA-binding capacities, and some of them do it in more efficient manner than others. Remarkably, observed miRNA–mRNA sponge-like interactions were cell-line-specific, with very little overlap identified. In HEK293 cells, the most prominent sponge-like mRNA, with 77 different miRNA interactions detected, was one encoding for AGO1. In settings of this particular experiment, this mRNA had been overexpressed artificially, as part of CLASH protocol. Two other HEK293-specific "sponge-like" mRNAs, HSPA1B and EEF1A1, formed 77 and 50 interactions, respectively.

For each of these mRNAs, amounts of detected interactions were comparable to that of a well-known circular RNA with sponge properties, Cdr1as (74 predicted sites) (Xu et al., 2015). In Huh7.5 cells, the set of RNAs with "sponge-like" activities included many well-described noncoding RNAs—for example, MALAT1 and XIST. It is peculiar that some Huh7.5-specific sponge-like RNAs, including those for alpha-fetoprotein (AFP) (Parpart et al., 2014) and APOB (Bi et al., 2014), were previously described as biomarkers of liver carcinoma, a tissue of origin for Huh7.5 cell line. In any case, presented set of experimentally identified miRNA–mRNA interactions allows finding a set of endogenous RNAs competing for any particular miRNA.

Some miRNAs expressed at relatively high levels were not among RNA interactors at all. About a hundred of such non-interacting miRNAs were present in both studied cell lines. There is a possibility that the natural targets for these microRNAs are either not expressed in studied cellular contexts, or that they have no targets at all. In total, only 232 microRNAs had at least one interaction in each of studied cell lines.

For individual miRNAs, levels of their expression have no bearing on amounts of interactions they display, possibly reflecting difference in their functions depending on the cellular context. As an example, we revealed that, in Huh7.5 cell line, miR-423-3p is abundant but displays only a few interactions, while in HEK293 cell line, the same miRNA forms more than 200 interactions and expressed at the quite low level. These observations complement previous findings of Mullokandov and colleagues (Mullokandov et al., 2012), who have shown that the binding activity of some highly expressed miRNAs may be weakened by either high target-to-miRNA ratio or the relocation of this miRNA to the nucleus. Further studies are required in order to investigate how RNA binding properties of individual miRNAs may change in response to context-dependent regulation by extrinsic or intrinsic factors.

Augmenting CLASH and CLEAR-CLIP datasets with additional 79 CLIP datasets provided us with information about microRNA footprints of many thousands of experimentally confirmed microRNA-binding regions (Exp-MiBRs) distributed through both coding and noncoding regions of RNA loci. At least some Exp-MiBRs are tissue-specific, in agreement with Clark and colleagues, who revealed the differences in the microRNA targetomes across tissues (Clark et al., 2014).

In addition to chromosomes, many Exp-MiBRs map to mitochondrial DNA, where they are quite abundant. Previous studies showed four mitochondrial regions with high degree of homology to microRNAs, namely, hsa-miR-4461 (chrM: 10,690–10,712), hsa-miR-4463 (chrM: 13,050–13,068), hsa-miR-4484 (chrM: 5,749–5,766), and hsa-miR-4485 (chrM: 2,562–2,582) (Sripada et al., 2012). Two of these regions encode mitochondrial *ND4L* and 16S rRNA genes and correspond to highly interacting Exp-MiBRs, with 70 and 63 cognate miRNAs, respectively, all confirmed in nine different cell lines. In both cases, previously identified cognate miRNAs hsa-miR-4461 and hsa-miR-4485 were among confirmed interactors. Our study expands the coverage of mitochondrial genome by various miRNA-interacting regions to 86% of its lengths. Altogether, these findings support the notion that miRNA–mRNA interactions take place in a variety of cellular compartments, including mitochondria (Ni and Leng, 2015).

The landscape of microRNA–mRNA human interactions, which we derived from both direct microRNA–mRNA interactions experimentally defined in HEK293 and Huh7.5 cell lines, when analyzed along with microRNA and mRNA expression data, highlights enormous complexity of human microRNA–mRNA interactome. For individual miRNAs, levels of their expression have no bearing on amounts of interactions they display, possibly reflecting context depending difference in their functions. In this article, we show that, while only 1–2% of human genes are highly regulated by microRNAs, a few cell-specific RNAs display sponge-like effects, including *EEF1A1* and *HSPA1B* in HEK293 and *AFP*, *APOB*, and *MALAT1* genes in Huh7.5 cell lines. Some miRNAs might be expressed at relatively low levels and interact with many mRNAs. On the other hand, there is a set of microRNAs expressed at a very high level and interacting with only a few mRNAs, thus, indeed, regulating expression of their targets in a specific manner. Notably, microRNAs are capable of switching between these two modes of action, depending on cellular context. The question of the biological significance of these two miRNA groups remains open. CLASH and/or CLEAR-CLIP coverage of additional cell lines is warranted. It is notable, however, that the presence of miRNA groups, one with a low expression level and a high number of interactions, and one with opposite characteristics, was independently detected in both cell lines profiled.

We have also established a collection of reliable microRNA-binding regions that we systematically extracted in course of an analysis of 79 CLIP datasets. This collection is available at http://score.generesearch.ru/services/mirna/. The promise of microRNAs as potential diagnostic mean and therapeutic target got expanded with a number of pathogenic loss-of-function and, recently, gain-of-function mutations described (Grigelioniene et al., 2019). Hence, our efforts in mapping the human miRNA–mRNA interactome may be aided in untangling molecular underpinnings of hereditary and acquired diseases.

DATA AVAILABILITY STATEMENT

microRNA–mRNA interactome data were extracted from published CLASH (Helwak et al., 2013) and CLEAR-CLIP (Moore et al., 2015) studies. Publicly available datasets of RNA and microRNA expression were from GEO accessions "GSE68611" (Murakawa et al., 2015), "GSE64677" (Luna et al., 2015), "GSE75136" (Wissink et al., 2016), "GSE74014" (Bandiera et al., 2016). GEO IDs' of open-accessed Raw CLIP datasets are listed as **Supplementary Table 3**. All data generated during this study are included in this published article and its supplementary information files.

AUTHOR CONTRIBUTIONS

MS and OP designed the study and carried out the research. AB contributed to the discussion of the results. OP and AB wrote the paper. All authors read and approved the final manuscript.

FUNDING

This project has been funded in part by the Laboratory of functional genomics of the Research Centre for Medical Genetics and by the Laboratory of functional genome analysis of the Moscow Institute of Physics and Technology.

ACKNOWLEDGMENTS

We thank Andrey Marakhonov and members of the Skoblov laboratory for helpful discussions.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00933/full#supplementary-material

REFERENCES

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

Artcibasova, A. V., Korzinkin, M. B., Sorokin, M. I., Shegay, P. V., Zhavoronkov, A. A., Gaifullin, N., et al. (2016). MiRImpact, a new bioinformatic method using complete microRNA expression profiles to assess their overall influence on the activity of intracellular molecular pathways. *Cell Cycle* 15 (5), 689–698. doi: 10.1080/15384101.2016.1147633

Bandiera, S., Pernot, S., El Saghire, H., Durand, S. C., Thumann, C., Crouchet, E., et al. (2016). Hepatitis C virus-induced upregulation of microRNA miR-146a-5p in hepatocytes promotes viral infection and deregulates metabolic pathways associated with liver disease pathogenesis. *J. Virol.* 90 (14), 6387–6400. doi: 10.1128/JVI.00619-16

Bartel, D. P. (2004). MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* 116 (2), 281–297. doi: 10.1016/S0092-8674(04)00045-5

Bi, Y., He, Y., Huang, J., Su, Y., Zhu, G. H., Wang, Y., et al. (2014). Functional characteristics of reversibly immortalized hepatic progenitor cells derived from mouse embryonic liver. *Cell. Physiol. Biochem.* 34 (4), 1318–1338. doi: 10.1159/000366340

Bray, N. L., Pimentel, H., Melsted, P., and Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* 34 (5), 525–527. doi: 10.1038/nbt.3519

Clark, P. M., Loher, P., Quann, K., Brody, J., Londin, E. R., and Rigoutsos, I. (2014). Argonaute CLIP-Seq reveals miRNA targetome diversity across tissue types. *Sci. Rep.* 4, 5947. doi: 10.1038/srep05947

Corcoran, D. L., Georgiev, S., Mukherjee, N., Gottwein, E., Skalsky, R. L., Keene, J. D., et al. (2011). PARalyzer: definition of RNA binding sites from PAR-CLIP short-read sequence data. *Genome Biol.* 12 (8), R79. doi: 10.1186/gb-2011-12-8-r79

Frankish, A., Diekhans, M., Ferreira, A. M., Johnson, R., Jungreis, I., Loveland, J., et al. (2018). GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res.* 47 (D1), D766–D773. doi: 10.1093/nar/gky955

Friedman, R. C., Farh, K. K. H., Burge, C. B., and Bartel, D. P. (2009). Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res.* 19 (1), 92–105. doi: 10.1101/gr.082701.108

Grigelioniene, G., Suzuki, H. I., Taylan, F., Mirzamohammadi, F., Borochowitz, Z. U., Ayturk, U. M., et al. (2019). Gain-of-function mutation of microRNA-140 in human skeletal dysplasia. *Nat. Med.* 1, 583. doi: 10.1038/s41591-019-0353-2

Grimson, A., Farh, K. K. H., Johnston, W. K., Garrett-Engele, P., Lim, L. P., and Bartel, D. P. (2007). MicroRNA targeting specificity in mammals: determinants beyond seed pairing. *Mol. Cell.* 27 (1), 91–105. doi: 10.1016/j.molcel.2007.06.017

Gumienny, R., and Zavolan, M. (2015). Accurate transcriptome-wide prediction of microRNA targets and small interfering RNA off-targets with MIRZA-G. *Nucleic Acids Res.* 43 (3), 1380–1391. doi: 10.1093/nar/gkv050

He, L., and Hannon, G. J. (2004). MicroRNAs: small RNAs with a big role in gene regulation. *Nat. Rev. Genet.* 5 (7), 522–531. doi: 10.1038/nrg1379

Heberle, H., Meirelles, G. V., da Silva, F. R., Telles, G. P., and Minghim, R. (2015). InteractiVenn: a web-based tool for the analysis of sets through Venn diagrams. *BMC Bioinformatics* 16, 169. doi: 10.1186/s12859-015-0611-3

Helwak, A., Kudla, G., Dudnakova, T., and Tollervey, D. (2013). Mapping the human miRNA interactome by CLASH reveals frequent noncanonical binding. *Cell* 153 (3), 654–665. doi: 10.1016/j.cell.2013.03.043

Hu, B., Yang, Y. C. T., Huang, Y., Zhu, Y., and Lu, Z. J. (2016). POSTAR: a platform for exploring post-transcriptional regulation coordinated by RNA-binding proteins. *Nucleic Acids Res.* 45 (D1), D104–D114. doi: 10.1093/nar/gkw888

Jonas, S., and Izaurralde, E. (2015). Towards a molecular understanding of microRNA-mediated gene silencing. *Nat. Rev. Genet.* 16 (7), 421–433. doi: 10.1038/nrg3965

Karolchik, D., Hinrichs, A. S., Furey, T. S., Roskin, K. M., Sugnet, C. W., Haussler, D., et al. (2004). The UCSC Table Browser data retrieval tool. *Nucleic Acids Res.* 32 (Database issue), D493–D496. doi: 10.1093/nar/gkh103

Kim, J. E., Hong, J. W., Lee, H. S., Kim, W., Lim, J., Cho, Y. S., et al. (2018). Hsa-miR-10a-5p downregulation in mutant UQCRB-expressing cells promotes the cholesterol biosynthesis pathway. *Sci. Rep.* 8 (1), 12407. doi: 10.1038/s41598-018-30530-6

Kong, Q., Zhang, S., Liang, C., Zhang, Y., Kong, Q., Chen, S., et al. (2018). LncRNA XIST functions as a molecular sponge of miR-194-5p to regulate MAPK1 expression in hepatocellular carcinoma cell. *J. Cell. Biochem.* 119 (6), 4458–4468. doi: 10.1002/jcb.26540

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

Kuhn, R. M., Haussler, D., and Kent, W. J. (2012). The UCSC genome browser and associated tools. *Brief. Bioinformatics* 14 (2), 144–161. doi: 10.1093/bib/bbs038

Li, Y., and Zhang, Z. (2014). Potential microRNA-mediated oncogenic intercellular communication revealed by pan-cancer analysis. *Sci. Rep.* 4, 7097. doi: 10.1038/srep07097

Li, Y., Liang, C., Wong, K. C., Jin, K., and Zhang, Z. (2014). Inferring probabilistic miRNA–mRNA interaction signatures in cancers: a role-switch approach. *Nucleic Acids Res.* 42 (9), e76–e76. doi: 10.1093/nar/gku182

Licatalosi, D. D., Mele, A., Fak, J. J., Ule, J., Kayikci, M., Chi, S. W., et al. (2008). HITS-CLIP yields genome-wide insights into brain alternative RNA processing. *Nature* 456 (7221), 464–469. doi: 10.1038/nature07488

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 15 (12), 550. doi: 10.1186/s13059-014-0550-8

Lu, Y., and Leslie, C. S. (2016). Learning to predict miRNA-mRNA interactions from AGO CLIP sequencing and CLASH data. *PLoS Comput. Biol.* 12 (7), e1005026. doi: 10.1371/journal.pcbi.1005026

Luna, J. M., Scheel, T. K., Danino, T., Shaw, K. S., Mele, A., Fak, J. J., et al. (2015). Hepatitis C virus RNA functionally sequesters miR-122. *Cell* 160 (6), 1099–1110. doi: 10.1016/j.cell.2015.02.025

Moore, M. J., Scheel, T. K., Luna, J. M., Park, C. Y., Fak, J. J., Nishiuchi, E., et al. (2015). miRNA-target chimeras reveal miRNA 3'-end pairing as a major determinant of argonaute target specificity. *Nat. Commun.* 6, 8864. doi: 10.1038/ncomms9864

Moore, M. J., Zhang, C., Gantman, E. C., Mele, A., Darnell, J. C., and Darnell, R. B. (2014). Mapping argonaute and conventional RNA-binding protein interactions with RNA at single-nucleotide resolution using HITS-CLIP and CIMS analysis. *Nat. Protoc.* 9 (2), 263–293. doi: 10.1038/nprot.2014.012

Mullokandov, G., Baccarini, A., Ruzo, A., Jayaprakash, A. D., Tung, N., Israelow, B., et al. (2012). High-throughput assessment of microRNA activity and function using microRNA sensor and decoy libraries. *Nat. Methods.* 9 (8), 840–846. doi: 10.1038/nmeth.2078

Murakawa, Y., Hinz, M., Mothes, J., Schuetz, A., Uhl, M., Wyler, E., et al. (2015). RC3H1 post-transcriptionally regulates A20 mRNA and modulates the activity of the IKK/NF-κB pathway. *Nat. Commun.* 6, 7367. doi: 10.1038/ncomms8367

Ni, W. J., and Leng, X. M. (2015). Dynamic miRNA-mRNA paradigms: new faces of miRNAs. *Biochem. Biophys. Rep.* 4, 337–341. doi: 10.1016/j.bbrep.2015.10.011

Parpart, S., Roessler, S., Dong, F., Rao, V., Takai, A., Ji, J., et al. (2014). Modulation of miR-29 expression by α-fetoprotein is linked to the hepatocellular carcinoma epigenome. *Hepatology* 60 (3), 872–883. doi: 10.1002/hep.27200

Plotnikova, O. M., and Skoblov, M. Y. (2018). Efficiency of the miRNA-mRNA interaction prediction programs. *Mol. Biol. (Mosk.)* 52 (3), 543–554. doi: 10.7868/S0026898418030187

Ragan, C., Cloonan, N., Grimmond, S. M., Zuker, M., and Ragan, M. A. (2009). Transcriptome-wide prediction of miRNA targets in human and mouse using FASTH. *PLoS One* 4 (5), e5745. doi: 10.1371/annotation/e0842765-3cae-4737-8b5b-96aeb12d7fb5

Riffo-Campos, Á., Riquelme, I., and Brebi-Mieville, P. (2016). Tools for sequence-based miRNA target prediction: what to choose? *Int. J. Mol. Sci.* 17 (12), 1987. doi: 10.3390/ijms17121987

Sætrom, P., Heale, B. S., Snøve, O., Jr., Aagaard, L., Alluin, J., and Rossi, J. J. (2007). Distance constraints between microRNA target sites dictate efficacy and cooperativity. *Nucleic Acids Res.* 35 (7), 2333–2342. doi: 10.1093/nar/gkm133

Selbach, M., Schwanhäusser, B., Thierfelder, N., Fang, Z., Khanin, R., and Rajewsky, N. (2008). Widespread changes in protein synthesis induced by microRNAs. *Nature* 455 (7209), 58–63. doi: 10.1038/nature07228

Sripada, L., Tomar, D., Prajapati, P., Singh, R., Singh, A. K., and Singh, R. (2012). Systematic analysis of small RNAs associated with human mitochondria by deep sequencing: detailed analysis of mitochondrial associated miRNA. *PLoS One* 7 (9), e44873. doi: 10.1371/journal.pone.0044873

Steinkraus, B. R., Toegel, M., and Fulga, T. A. (2016). Tiny giants of gene regulation: experimental strategies for microRNA functional studies. *Wiley Interdiscip. Rev. Dev. Biol.* 5 (3), 311–362. doi: 10.1002/wdev.223

Thomson, D. W., and Dinger, M. E. (2016). Endogenous microRNA sponges: evidence and controversy. *Nat. Rev. Genet.* 17 (5), 272–283. doi: 10.1038/nrg.2016.20

Uhlmann, S., Mannsperger, H., Zhang, J. D., Horvat, E. Á., Schmidt, C., Küblbeck, M., et al. (2012). Global microRNA level regulation of EGFR-driven cell-cycle protein network in breast cancer. *Mol. Syst. Biol.* 8, 570. doi: 10.1038/msb.2011.100

Wang, K., Li, M., and Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* 38 (16), e164. doi: 10.1093/nar/gkq603

Weiss, C. N., and Ito, K. (2017). A macro view of microRNAs: the discovery of microRNAs and their role in hematopoiesis and hematologic disease. *Int. Rev. Cell Mol. Biol.* 334, 99–175. doi: 10.1016/bs.ircmb.2017.03.007

Wissink, E. M., Fogarty, E. A., and Grimson, A. (2016). High-throughput discovery of post-transcriptional cis-regulatory elements. *BMC Genomics* 17, 177. doi: 10.1186/s12864-016-2479-7

Xu, H., Guo, S., Li, W., and Yu, P. (2015). The circular RNA Cdr1as, *via* miR-7 and its targets, regulates insulin transcription and secretion in islet cells. *Sci. Rep.* 5, 12453. doi: 10.1038/srep12453

Yang, H., and Wang, K. (2015). Genomic variant annotation and prioritization with ANNOVAR and wANNOVAR. *Nat. Protoc.* 10 (10), 1556–1566. doi: 10.1038/nprot.2015.105

Yates, A., Beal, K., Keenan, S., McLaren, W., Pignatelli, M., Ritchie, G. R., et al. (2014). The Ensembl REST API: Ensembl data for any language. *Bioinformatics* 31 (1), 143–145. doi: 10.1093/bioinformatics/btu613

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

frontiers in Genetics

ORIGINAL RESEARCH
published: 22 November 2019
doi: 10.3389/fgene.2019.01203

The Impact of Pathway Database Choice on Statistical Enrichment Analysis and Predictive Modeling

*Sarah Mubeen[1,2], Charles Tapley Hoyt[1,2†], André Gemünd[1], Martin Hofmann-Apitius[1,2], Holger Fröhlich[2] and Daniel Domingo-Fernández[1,2]**

[1] Department of Bioinformatics, Fraunhofer Institute for Algorithms and Scientific Computing (SCAI), Sankt Augustin, Germany, [2] Bonn-Aachen International Center for IT, Rheinische Friedrich-Wilhelms-Universität Bonn, Bonn, Germany

OPEN ACCESS

Edited by:
Lavanya Balakrishnan, Mazumdar Shaw Medical Centre, India

Reviewed by:
George C. Tseng, University of Pittsburgh, United States
Inyoung Kim, Virginia Tech, United States

***Correspondence:**
Daniel Domingo-Fernández
daniel.domingo.fernandez@scai.fraunhofer.de

†ORCID:
Charles Tapley Hoyt
orcid.org/0000-0003-4423-4370

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

Received: *23 August 2019*
Accepted: *30 October 2019*
Published: *22 November 2019*

Citation:
Mubeen S, Hoyt CT, Gemünd A, Hofmann-Apitius M, Fröhlich H and Domingo-Fernández D (2019) The Impact of Pathway Database Choice on Statistical Enrichment Analysis and Predictive Modeling. Front. Genet. 10:1203. doi: 10.3389/fgene.2019.01203

Pathway-centric approaches are widely used to interpret and contextualize *-omics* data. However, databases contain different representations of the same biological pathway, which may lead to different results of statistical enrichment analysis and predictive models in the context of precision medicine. We have performed an in-depth benchmarking of the impact of pathway database choice on statistical enrichment analysis and predictive modeling. We analyzed five cancer datasets using three major pathway databases and developed an approach to merge several databases into a single integrative one: MPath. Our results show that equivalent pathways from different databases yield disparate results in statistical enrichment analysis. Moreover, we observed a significant dataset-dependent impact on the performance of machine learning models on different prediction tasks. In some cases, MPath significantly improved prediction performance and also reduced the variance of prediction performances. Furthermore, MPath yielded more consistent and biologically plausible results in statistical enrichment analyses. In summary, this benchmarking study demonstrates that pathway database choice can influence the results of statistical enrichment analysis and predictive modeling. Therefore, we recommend the use of multiple pathway databases or integrative ones.

Keywords: pathway enrichment, benchmarking, databases, machine learning, statistical hypothesis testing

INTRODUCTION

As fundamental interactions within complex biological systems have been discovered in experimental biology labs, they have often been assembled into computable pathway representations. Because they have proven immensely useful in the analysis and interpretation of *-omics* data when coupled with algorithmic approaches (e.g., gene set enrichment analysis, GSEA), academic and commercial groups have generated and maintained a comprehensive set of databases during the last 15 years (Bader et al., 2006). Examples include KEGG, Reactome, WikiPathways, NCIPathways, and Pathway Commons (Schaefer et al., 2008; Cerami et al., 2011; Kanehisa et al., 2016; Slenter et al., 2017; Fabregat et al., 2018).

However, these databases tend to differ in the average number of pathways they contain, the average number of proteins per pathway, the types of biochemical interactions they incorporate, and the subcategories of pathways that they provide (e.g., signal transduction, genetic interaction, and metabolic) (Kirouac et al., 2012; Türei et al., 2016). Pathways are often also described at varying levels of detail, with diverse data types and with loosely defined boundaries (Domingo-Fernández et al.,

2018). Nonetheless, most pathway analyses are still conducted exclusively by employing a single database, often chosen in part by researchers' preferences or previous experiences (e.g., bias towards a database previously yielding good results and ease of use of a particular database) (**Table 1**). Notably, the selection of a suitable pathway database depends on the actual biological context that is investigated, yet KEGG remains severely overrepresented in published -omics studies. This raises concerns and motivates the consideration of multiple pathway databases or, preferably, an integration over several pathways resources.

Several integrative resources have been developed, including meta-databases [e.g., Pathway Commons (Cerami et al., 2011), MSigDB (Liberzon et al., 2015), and ConsensusPathDB (Kamburov et al., 2008)] that enable pathway exploration in their corresponding web applications and integrative software tools [e.g., graphite (Sales et al., 2018), PathMe (Domingo-Fernandez et al., 2019), and OmniPath (Türei et al., 2016)] designed to enable bioinformatics analyses. By consolidating pathway databases, these resources have attempted to summarize major reference points in the existing knowledge and demonstrate how data contained in one resource can be complemented by data contained in others. Thus, through their usage, the biomedical community has benefitted from comprehensive overviews of pathway landscapes which can then make for more robust resources highly suited for analytic usage.

The typical approach to combine pathway information with -omics data is *via* statistical enrichment analysis, also known as pathway enrichment. The task of navigating through the continuously developing variants of enrichment methods has been undertaken by several recent studies which benchmarked the performance of these techniques (Bayerlová et al., 2015; Ihnatova et al., 2018; Lim et al., 2018) and guide users on the choice for their analyses (Fabris et al., 2019; Reimand et al., 2019). While Bateman et al. (2014) examined the impact of choice of different subsets of MSigDB on GSEA, it remains unclear what broader impact an integrative pathway meta-database would have for statistical enrichment analysis. Additionally, the overlap of pathways within the same integrative database can induce biases (Liberzon et al., 2015), specifically when conducting multiple testing correction *via* the popular Benjamini–Hochberg method (Benjamini and Hochberg, 1995) that supposes independence of statistical tests. This issue is of particular concern for large-scale meta-databases such as MSigDB.

TABLE 1 | Number of publications citing major pathway resources for pathway enrichment in PubMed Central (PMC), 2019. To develop an estimate on the number of publications using several pathway databases for pathway enrichment, SCAIView (http://academia.scaiview.com/academia; indexed on 01/03/2019) was used to conduct the following query using the PMC corpus: "<pathway resource>" AND "pathway enrichment".

Type	Pathway resource	Publications
Primary	**KEGG**	27,713
	Reactome	3,765
	WikiPathways	651
Integrative	**MSigDB**	2,892
	ConsensusPathDB	339
	Pathway Commons	1,640

The aim of this work is to systematically investigate the influence of alternative representations of the same biological pathway (e.g., in KEGG, Reactome, and WikiPathways) on the results of statistical enrichment analysis *via* three common methods: the hypergeometric test, GSEA, and signaling pathway impact analysis (SPIA) (Fisher, 1992; Subramanian et al., 2005; Tarca et al., 2008) using five The Cancer Genome Atlas (TCGA) datasets (Weinstein et al., 2013). In addition, we also show that pathway activity-based patient classification and survival analysis *via* single-sample GSEA (ssGSEA; Barbie et al., 2009) can be impacted by the choice of pathway resource in some cases. As a solution, we propose to integrate different pathway resources *via* a method where semantically analogous pathways across databases (e.g., "Notch signaling pathway" in KEGG and "Signaling by NOTCH" pathway in Reactome) are combined. This approach exploits the pathway mappings and harmonized pathway representations described in our previous work (Domingo-Fernández et al., 2018; Domingo-Fernandez et al., 2019). We demonstrate that when aided by our integrative pathway database, it is possible to better capture expected disease biology than with individual resources, and to sometimes obtain better predictions of clinical endpoints. Our entire analytic pipeline is implemented in a reusable Python package (pathway_forte; see *Materials and Methods*) to facilitate reproducing the results with other databases or datasets in the future.

MATERIALS AND METHODS

In the first two subsections, we describe the pathway resources and the clinical and genomic datasets we used in benchmarking. The following sections then outline the statistical enrichment analysis and predictive modeling conducted in this study. Finally, in the last two subsections, we describe the statistical methods and the software implemented to conduct the benchmarking.

Pathway Databases

Selection Criteria

Numerous viable pathway databases have been made available to infer biologically relevant pathway activity (Bader et al., 2006). In this work, we systematically compared three major ones (i.e., KEGG, Reactome, and WikiPathways) as the subset of databases to benchmark. The rationale for the inclusion of these databases was twofold: firstly, these databases are open-sourced, well-established, and highly cited in studies investigating pathways associated with variable gene expression patterns in different sets of conditions (**Table 1**). Secondly, we expected distinctions between these databases to be strong enough to observe variable results of enrichment analysis and patient classification, yet these databases also contain a reasonable number of equivalent pathways such that objective comparisons could be made, as outlined in our previous work (Domingo-Fernández et al., 2018).

Data Retrieval and Processing

In order to systematically compare results yielded by different databases, we retrieved the contents of KEGG, Reactome, and WikiPathways using ComPath (Domingo-Fernández et al., 2018)

and converted it into the Gene Matrix Transposed (GMT) file format. Generated networks encoded in Biological Expression Language (BEL; Slater, 2014) were retrieved using PathMe (Domingo-Fernández et al., 2019).

To test the potential utility of an integrative pathway resource, we used equivalent pathways across the three databases that were manually curated in our previous work (Domingo-Fernández et al., 2018; see our earlier publication for further details). In the following, we call these "pathways analogs" or "equivalent pathways" (**Figure 1A**), while we call a pathway found as analogous across all KEGG, Reactome, as well as WikiPathways a "super pathway".

In a second step, we merged equivalent pathways by taking the graph union with respect to contained genes and interactions (**Figures 1B, C**). We have also described this step in more detail in our earlier work (Domingo-Fernandez et al., 2019).

The set union of KEGG, Reactome, and WikiPathways, while taking into account pathway equivalence, gave rise to an integrative resource to which we refer as *MPath* (**Figure 1D**). By merging equivalent pathways, MPath contains a fewer number of pathways than the sum of all pathways from all primary resources. In total, MPath contains 2,896 pathways, of which 238 are derived from KEGG, 2,119 from Reactome, and 409 from

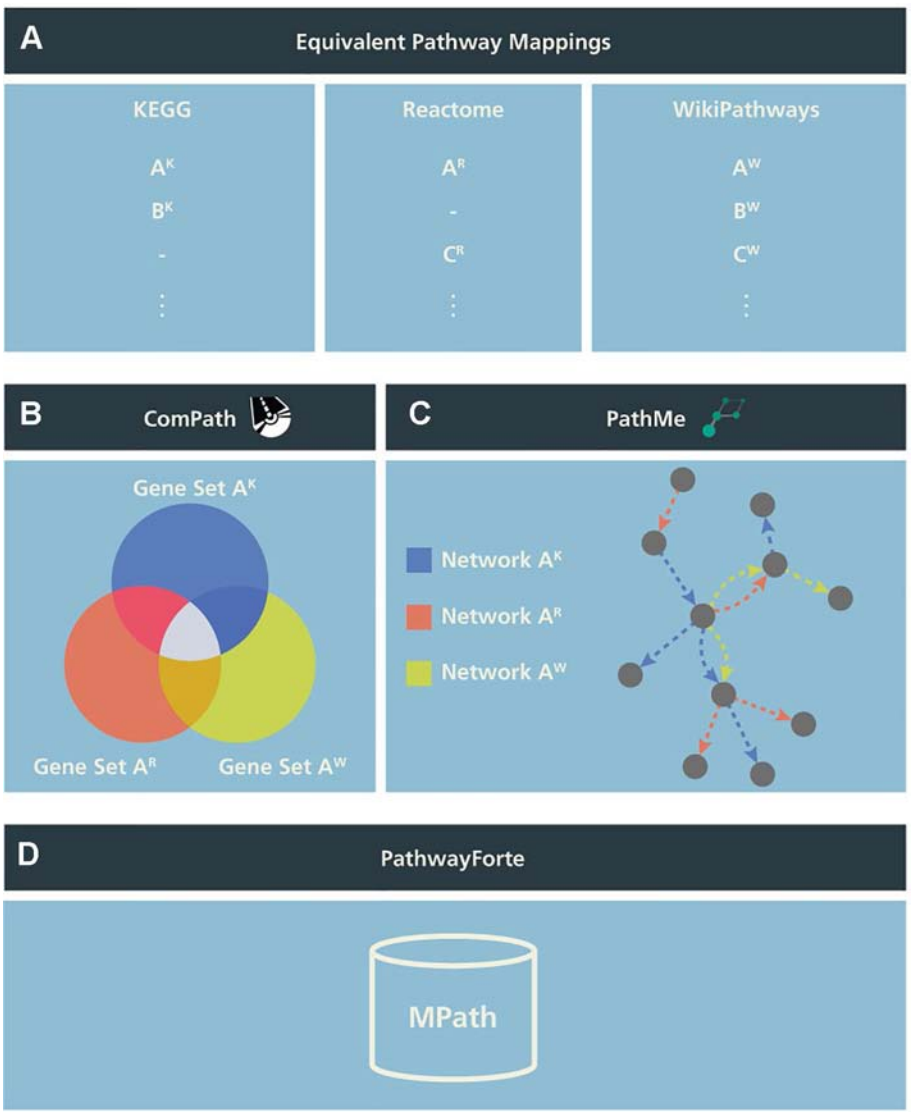


FIGURE 1 | Schema illustrating the generation of MPath. The curated pathway mapping catalog is depicted in **(A)**, which links equivalent pathways from different resources. Pathways that are shared across two resources are referred to as pathway analogs (i.e., Pathway A in Reactome and Pathway A′ in KEGG) and pathways that are shared across all three resources are referred to as "super pathways" (i.e., Pathway A in KEGG, Pathway A′ in Reactome, and Pathway A″ in WikiPathways). **(B)** Using these mappings, gene sets of equivalent pathways from different resources can be combined, ensuring key molecular players from the different resources are included. **(C)** Similarly, network representations of the pathways can be overlaid to generate more comprehensive pathways. **(D)** Finally, both the combined gene sets and networks representations are included in MPath. Note that pathways that are exclusive to a single database are included in MPath unchanged.

WikiPathways, while another 129 pathways are pathway analogs and 26 are super pathways.

We next compared the latest versions of pathway gene sets from KEGG, Reactome, WikiPathways, and MPath with pathway gene sets from MSigDB, a highly cited integrative pathway database containing older versions of the KEGG and Reactome gene sets (Liberzon et al., 2015). We downloaded KEGG and Reactome gene sets from the curated gene set (C2) collection of MSigDB (http://software.broadinstitute.org/gsea/msigdb/collections.jsp #C2; version6.2; July 2018). Detailed statistics on the number of pathways from each resource are presented in **Table S1**.

Clinical and Genomic Data

We used five widely used datasets acquired from TCGA (Weinstein et al., 2013), a cancer genomics project that has catalogued molecular and clinical information for normal and tumor samples (**Table 2**). TCGA data were retrieved through the Genomic Data Commons (GDC; https://gdc.cancer.gov) portal and cBioportal (https://www.cbioportal.org) on 14-03-2019. RNA-seq gene expression data subjected to an mRNA quantification analysis pipeline for BRCA, KIRC, LIHC, OV, and PRAD TCGA datasets were queried, downloaded, and prepared from the GDC through the R/Bioconductor package, TCGAbiolinks (R version: 3.5.2; TCGAbiolinks version: 2.10.3) (Colaprico et al., 2015). The data were preprocessed as follows: gene expression was quantified by the number of reads aligned to each gene and read counts were measured using HTSeq and normalized using fragments per kilobase of transcript per million mapped reads upper quartile (FPKM-UQ). HTSeq raw read counts also subject to the GDC pipeline were similarly queried, downloaded, and prepared with TCGAbiolinks. Read count data downloaded for the BRCA, KIRC, LIHC, and PRAD datasets were processed to remove identical entries, while unique measurements of identical genes were averaged. The differential gene expression analysis of cancer versus normal samples was performed using the R/Bioconductor package, DESeq2 (version 1.22.2). Genes with adjusted p value < 5% were considered significantly dysregulated. For all downloaded data, gene identifiers were mapped to HGNC gene symbols (Povey et al., 2001), where possible. To obtain additional information on the survival status and time to death, or censored survival times of patients, patient identifiers in the TCGA datasets were mapped to their equivalent identifiers in cBioPortal. Additionally, cancer subtype classifications or the PRAD and BRCA datasets were retrieved from the GDC. We would like to note that although there are other cohorts available (e.g., COAD and STAD) containing all of these modalities, we did not include them in this analysis because of the limited number of samples they contain (i.e., less than 300 patients). Detailed statistics of all five datasets are presented in **Table 2**.

Pathway Enrichment Methods

In this subsection, we describe three different classes of pathway enrichment methods that we tested: 1) statistical overrepresentation analysis (ORA); 2) functional class scoring (FCS); and 3) pathway topology (PT)-based enrichment (**Figure 2**) (Khatri et al., 2012; García-Campos et al., 2015; Fabris et al., 2019).

Overrepresentation Analysis

We conducted pathway enrichment using genes that exhibited a q value <0.05 using a one-sided Fisher's exact test (Fisher, 1992) for each of the pathways in all pathway databases. We consider a pathway to be significantly enriched if its q value is smaller than 0.05 after applying multiple hypothesis testing correction with the Benjamini–Yekutieli method under dependency (Benjamini and Yekutieli, 2001).

Functional Class Scoring Methods

We selected GSEA, one of the most commonly used FCS methods (Subramanian et al., 2005). We performed GSEA with the Python package, GSEApy (version 0.9.12; https://github.com/zqfang/gseapy), using normalized RNA-seq expression quantifications (FPKM-UQ) obtained for the BRCA, KIRC, LIHC, and PRAD datasets containing both normal and tumor samples (**Table 2**). All genes were ranked by their differential expression based on their $\log_2$ fold changes. Query gene sets for GSEA included pathways from KEGG, Reactome, WikiPathways, and MPath. GSEA results were filtered to include pathway gene sets with p values below 0.05 and a minimum gene set size of 10 or a maximum gene size of 3,000. Similarly, GSEApy was used to perform ssGSEA (Barbie et al., 2009) (**Table S2**) to acquire sample-wise pathway scores using FPKM-UQ for BRCA, KIRC, LIHC, OV, and PRAD datasets, irrespective of phenotype labels (Barbie et al., 2009). Datasets were filtered to only include normalized expression data for genes found in the pathway gene sets of KEGG, Reactome, WikiPathways, and MPath and then used for ssGSEA. Expression data were ranked and sample-wise normalized enrichment scores were obtained.

TABLE 2 | Statistics of the five TCGA cancer datasets used in this work.

Cancer type	TCGA abbreviation	Tumor samples	Normal samples	Surviving patients	Deceased patients
Breast invasive carcinoma	BRCA	1,102	113	946	153
Kidney renal clear cell carcinoma	KIRC	538	72	365	173
Liver hepatocellular carcinoma	LIHC	371	50	240	130
Prostate adenocarcinoma	PRAD	498	52	498	10
Ovarian cancer	OV	374	0	143	229

The statistics correspond to those retrieved from the GDC portal and cBioportal on 14-03-2019. Longitudinal statistics of survival data are presented in Figure S1.

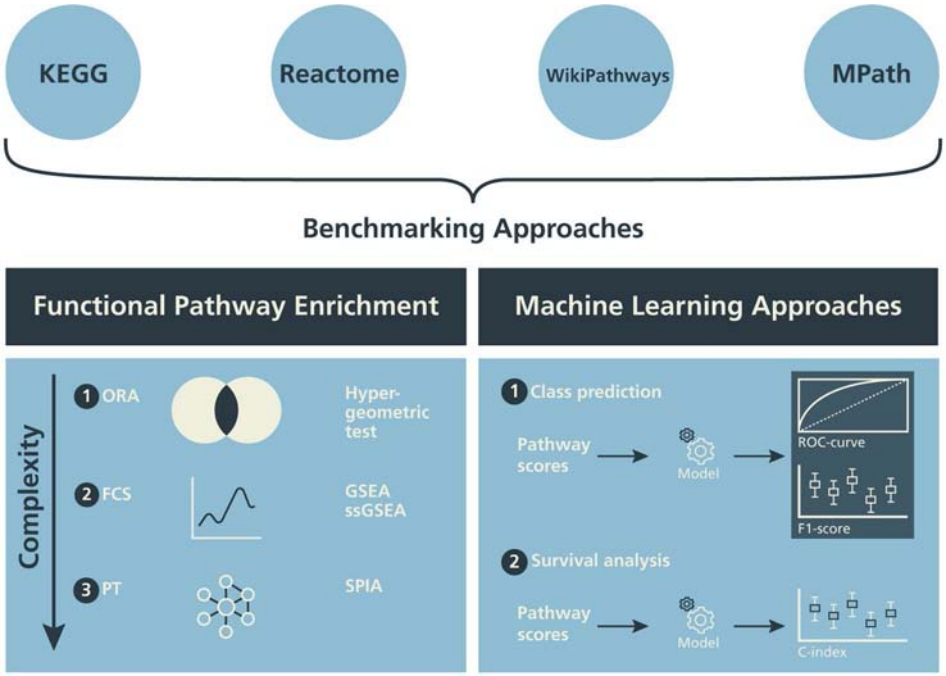


FIGURE 2 | Design of the benchmarking schema. The influence of alternative pathway databases on the results of statistical pathway enrichment (left) and machine learning classification tasks (right) are compared.

Pathway Topology-Based Enrichment

To evaluate PT-based methods, we selected the well-known and highly cited SPIA method (Tarca et al., 2008) for two main reasons: firstly, the guidelines outlined by a comparative study on topology-based methods (Ihnatova et al., 2018) recommend the use of SPIA for datasets with properties similar to TCGA (i.e., possessing two well-defined classes, full expression profiles, many samples, and numerous differentially expressed genes). Secondly, SPIA has been reported to have a high specificity while preserving dependency on topological information (Ihnatova et al., 2018). Because the R/Bioconductor's SPIA package only contains KEGG pathways, we converted the pathway topologies from the three databases used in this work to a custom format in a similar fashion as graphite (Sales et al., 2018) (**Supplementary Text**). We declared significance for SPIA-based pathway enrichment, if the Bonferroni corrected p value was <5%.

Evaluation Based on Enrichment of Pathway Analogs

To better understand the impact of database choice, we compared the raw p value rankings (i.e., before multiple testing correction) of pathway analogs across each possible pair of databases (i.e., in KEGG and Reactome, Reactome and WikiPathways, and WikiPathways and KEGG) and in each statistical enrichment analysis (i.e., hypergeometric test, GSEA, and SPIA) with the Wilcoxon signed-rank test. It assessed the average rank difference of the pathway analogs and reported how significantly different the results were for each database pair. Importantly, we only tested statistical enrichment of the analogous pathways in order to avoid statistical biases due to differences in the size of pathway databases.

Machine Learning

ssGSEA was conducted to summarize the gene expression profile mapping to a particular pathway of interest within a given patient sample, hence resulting in a pathway activity profile for each patient. We then evaluated the different pathway resources with respect to three machine learning tasks:

1. Prediction of tumor vs. normal
2. Prediction of known tumor subtype
3. Prediction of overall survival

Prediction of Tumor vs. Normal

The first task was to train and evaluate binary classifiers to predict normal versus tumor sample labels. This task was conducted for four of the five TCGA datasets (i.e., BRCA, KIRC, LIHC, and PRAD), while OV, which only contains tumor samples, was omitted. We performed this classification using a commonly used elastic net penalized logistic regression model (Zou and Trevor, 2005). Prediction performance was evaluated *via* a 10 times repeated 10-fold stratified cross-validation. Importantly, tuning of elastic net hyper-parameters (l_1, l_2 regularization parameters) was conducted within the cross-validation loop to avoid over-optimism (Molinaro et al., 2005).

Prediction of Tumor Subtype

The second task was to train and evaluate multi-label classifiers to predict tumor subtypes using sample-wise pathway activity scores generated from ssGSEA. This task was only conducted for the BRCA and PRAD datasets, similar to the work done by Lim et al. (2018), because the remaining three datasets included

in this work lacked subtype information. From the five breast cancer subtypes present in the BRCA dataset by the PAM50 classification method (Sorlie et al., 2001), we included four subtypes (i.e., 194 Basal samples, 82 Her2 samples, 567 LumA samples, and 207 LumB samples). These four were selected as they constitute the agreed-upon intrinsic breast cancer subtypes according to the 2015 St. Gallen Consensus Conference (Coates et al., 2015) and are also recommended by the ESMO Clinical Practice Guidelines (Senkus et al., 2015). For the PRAD dataset, evaluated subtypes included 151 ERG samples, 27 ETV1 samples, 14 ETV4 samples, 38 SPOP samples, and 87 samples classified as other (Cancer Genome Atlas Research Network, 2014). Similar to the approach by Graudenzi et al. (2017), support vector machines (SVMs) (Cortes and Vapnik, 1995) were used for subtype classification by implementing a one-versus-one strategy in which a single classifier is fit for each pair of class labels. This strategy transforms a multi-class classification problem into a set of binary classification problems. We again used a 10 times repeated 10-fold cross-validation scheme, and the soft margin parameter of the linear SVM was tuned within the cross-validation loop *via* a grid search. We assessed the multi-class classifier performance in terms of accuracy, precision, and recall.

Prediction of Overall Survival

The third task was to train and evaluate machine learning models to predict overall survival of cancer patients. For this purpose, a Cox proportional hazards model with elastic net penalty was used (Tibshirani, 1997; Friedman et al., 2010). Prediction performance was evaluated on the basis of five TCGA datasets (i.e., BRCA, LIHC, KIRC, OV, and PRAD) (**Table 2**) using the same 10 times repeated 10-fold nested cross-validation procedure as described before. The performance of the model was assessed by Harrell's concordance index (c-index; Harrell et al., 1982), which is an extension of the well-known area under receiver operating characteristic (ROC) curve for right censored time-to-event (here: death) data.

Statistical Assessment of Database Impact on Prediction Performance

To understand the degree to which the observed variability of area under the ROC curve (AUC) values, accuracies, and c-indices could be explained by the actually used pathway resource, we conducted a two-way analysis of variance (ANOVA). The ANOVA model had the following form:

$$\textbf{performance} \sim \textbf{database} + \textbf{dataset} + \textbf{database} \times \textbf{dataset}$$

We then tested the significance of the database factor *via* an F test. In addition, we performed Wilcoxon tests analysis to understand specific differences between databases in a dataset-dependent manner.

Software Implementation

The workflow presented in this article consists of three major components: 1) the acquisition and preprocessing of gene set and pathway databases; 2) the acquisition and preprocessing of experimental datasets; and 3) the re-implementation or adaptation of existing analytical pipelines for benchmarking. We implemented these components in the pathway_forte Python package to facilitate the reproducibility of this work, the inclusion of additional gene set and pathway databases, and to include additional experimental datasets.

The acquisition of KEGG, MSigDB, Reactome, and WikiPathways was mediated by their corresponding Bio2BEL Python packages (Hoyt et al., 2019; https://github.com/bio2bel) in order to provide uniform access to the underlying databases and to enable the reproduction of this work as they are updated. Each Bio2BEL package uses Python's *entry points* to integrate in the previously mentioned ComPath framework in order to support uniform preprocessing and enable the integration of further pathway databases in the future, without changing any underlying code in the pathway_forte package. The network preprocessing defers to PathMe (Domingo-Fernandez et al., 2019; https://github.com/pathwaymerger). Because it is based on PyBEL (Hoyt et al., 2018; https://github.com/pybel), it is extensible to the growing ecosystem of BEL-aware software.

While the acquisition and preprocessing of experimental datasets is currently limited to a subset of TCGA, it is extensible to further cancer-specific and other condition-specific datasets. We implemented independent preprocessing pipelines for several previously mentioned datasets using extensive manual curation, preparation, and processing with the pandas Python package (McKinney, 2010; https://github.com/pandas-dev/pandas). Unlike the pathway databases, which were amenable to standardization, the preprocessing of each new dataset must be bespoke.

The re-implementation and adaptation of existing analytical methods for functional enrichment and prediction involved wrapping several existing analytical packages (**Table S3**) in order to make their application programming interfaces more user-friendly and to make the business logic of the benchmarking more elegantly reflected in the source code of pathway_forte. Each is independent and can be used with any combination of pathway database and dataset. Finally, all figures presented in this paper and complementary analyses can be generated and reproduced with the Jupyter notebooks located at https://github.com/pathwayforte/results/.

Ultimately, we wrapped each of these components in a command line interface (CLI) such that the results presented in each section of this work can be generated with a corresponding command following the guidelines described by Grüning et al. (2019). The scripts for generating the figures in this manuscript are not included in the main pathway_forte, but rather in their own repository within Jupyter notebooks at https://github.com/PathwayForte/results.

The source code of the pathway_forte Python package is available at https://github.com/PathwayForte/pathway-forte, its latest documentation can be found at https://pathwayforte.readthedocs.io, and its distributions can be found on PyPI at https://pypi.org/project/pathway-forte.

The pathway_forte Python package has a tool chain consisting of pytest (https://github.com/pytest-dev/pytest) as a testing

framework, coverage (https://github.com/nedbat/coveragepy) to assess testing coverage, sphinx (https://github.com/sphinx-doc/sphinx) to build documentation, flake8 (https://github.com/PyCQA/flake8) to enforce code and documentation quality, setuptools (https://github.com/pypa/setuptools) to build distributions, pyroma (https://github.com/regebro/pyroma) to enforce package metadata standards, and tox (https://github.com/tox-dev/tox) as a build tool to facilitate the usage of each of these tools in a reproducible way. It leverages community and open-source resources to improve its usability by using Travis-CI (https://travis-ci.com) as a continuous integration service, monitoring testing coverage with Codecov (https://codecov.io), and hosting its documentation on Read the Docs (https://readthedocs.org).

Hardware

Computations for each of the tasks were performed on a symmetric multiprocessing (SMP) node with four Intel Xeon Platinum 8160 processors per node with 24 cores/48 threads each (96 cores/192 threads per node in total) and 2.1-GHz base/3.7-GHz Turbo Frequency with 1,536-GB/1.5-TB RAM (DDR4 ECC Reg). The network was 100 GBit/s Intel OmniPath, storage was 2× Intel P4600 1.6-TB U.2 PCIe NVMe for local intermediate data and BeeGFS parallel file system for Home directories. **Table 3** provides a qualitative description of the memory and time requirements for each task.

RESULTS

The results of the benchmarking study have been divided into two subsections for each of the pathway methods described above. We first compared the effects of database selection on the results of functional pathway enrichment methods. In the following subsection, we benchmarked the performance of the pathway resources on the various machine learning classification tasks conducted.

Benchmarking the Impact on Enrichment Methods

Overrepresentation Analysis

As illustrated by our results, pathway analogs from different pathway databases in several cases showed clearly significant rank differences (**Figure 3**). These differences were most pronounced between Reactome and WikiPathways. For example, while the "Thyroxine Biosynthesis" pathway was highly statistically significant (q value <0.01) in the LIHC dataset for Reactome, its analogs in WikiPathways (i.e., "Thyroxine (Thyroid Hormone) Production") and KEGG (i.e., "Thyroid Hormone Synthesis") were not. However, the pathway was found to be significantly enriched in MPath. Such differences were similarly observed for the "Notch signaling" pathway in the PRAD dataset, in which the pathway was highly statistically significant (q value <0.01) for Reactome and MPath, but showed no statistical significance for KEGG and WikiPathways. Similar cases were systematically observed for additional pathway analogs and super pathways, demonstrating that marked differences in rankings can arise depending on the database used.

TABLE 3 | A qualitative description of the computational costs of the analyses performed.

Task	Relative memory usage	Timescale
ORA	Low	Seconds
GSEA	Medium	Minutes
ssGSEA	Very high	Hours
Prediction of tumor vs. normal	Medium	Minutes
Prediction of known tumor subtype	Medium	Minutes
Prediction of overall survival	Medium	Hours

Performing ssGSEA required on the scale of 100 GB of RAM for some dataset/database combinations, while the other tasks could be run on a modern laptop with no issues.

Gene Set Enrichment Analysis

Similar to ORA, GSEA showed significant differences between pathway analogs across databases in several cases (**Figure 3**). These differences were most pronounced between KEGG and WikiPathways in the KIRC and LIHC datasets and between KEGG and Reactome in the BRCA and PRAD datasets. Since GSEA calculates the observed direction of regulation (e.g., over/underexpressed) of each pathway, we also examined whether super pathways or pathway analogs exhibited opposite signs in their normalized enrichment scores (NES) (e.g., one pathway is overexpressed while its equivalent pair is underexpressed). As an illustration, GSEA results of the LIHC dataset revealed the contradiction that the "DNA replication" pathway, one of 26 super pathways, was overexpressed according to Reactome and underexpressed according to KEGG and WikiPathways, though the pathway was not statistically significant for any of these databases. However, the merged "DNA replication" pathway in MPath appeared as significantly underexpressed. Similarly, in the BRCA dataset, the WikiPathways definition of the "Notch signaling" and "Hedgehog signaling" pathways were significantly overexpressed, while the KEGG and Reactome definitions were insignificantly overexpressed. Interestingly, both the merged "Notch signaling" and merged "Hedgehog signaling" pathways appeared as significantly underexpressed ($q < 0.05$) in MPath.

Signaling Pathway Impact Analysis

The final of the three statistical enrichment analyses conducted revealed further differences between pathway analogs across databases. As expected, differences in the results of analogous pathways were exacerbated on topology-based methods compared with ORA and GSEA, as these latter methods do not consider pathway topology (i.e., incorporation of pathway topology introduces one extra level of complexity, leading to higher variability) (**Figure 3**). Beyond a cursory inspection of the statistical results, we also investigated the concordance of the direction of change of pathway activity (i.e., activation or inhibition) for equivalent pathways. We found that for two database (i.e., LIHC and KIRC), the direction of change was inconsistently reported for the "TGF beta signaling" pathway, depending on the database used (i.e., the KEGG representation

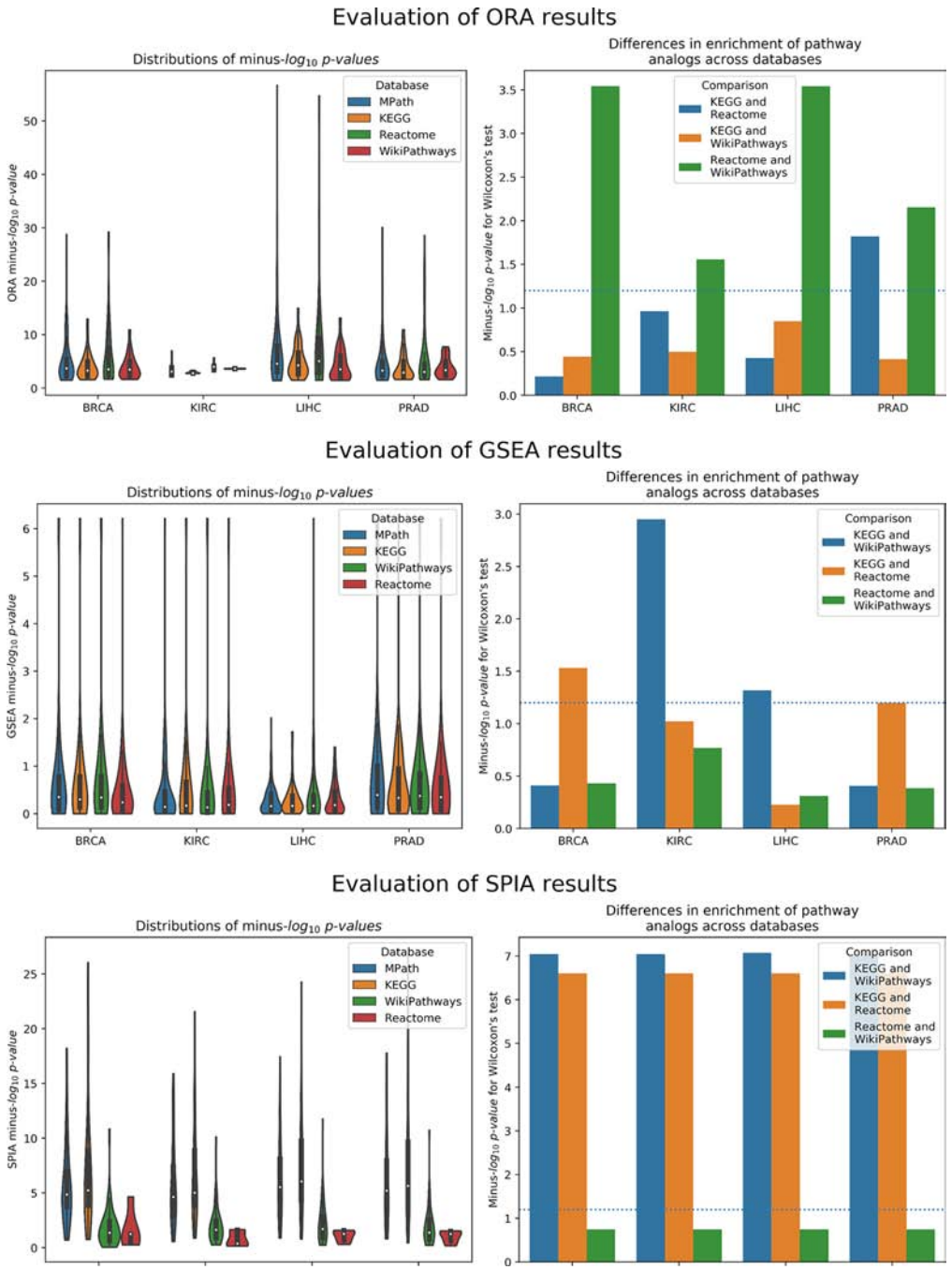


FIGURE 3 | *Left* Distribution of raw *p* values of pathway analogs across databases [*top to bottom*: overrepresentation analysis (ORA), gene set enrichment analysis (GSEA), and signaling pathway impact analysis (SPIA)]. *Right* Significance of average rank differences of pathway analogs across pairwise database comparisons for the given method.

was activated and the WikiPathways one inhibited). A similar effect was observed in the "Estrogen signaling pathway," found to be inhibited in KEGG and activated in WikiPathways in the LIHC dataset. The merging of equivalent pathway networks resulted in the observation of inhibition for both the "TGF beta signaling" and "Estrogen signaling" pathways in MPath results.

Benchmarking the Impact on Predictive Modeling

Prediction of Tumor vs. Normal

We compared the prediction performance of an elastic net penalized logistic regression classifier to discriminate normal from cancer samples based on their pathway activity profiles. The cross-validated prediction performance was measured

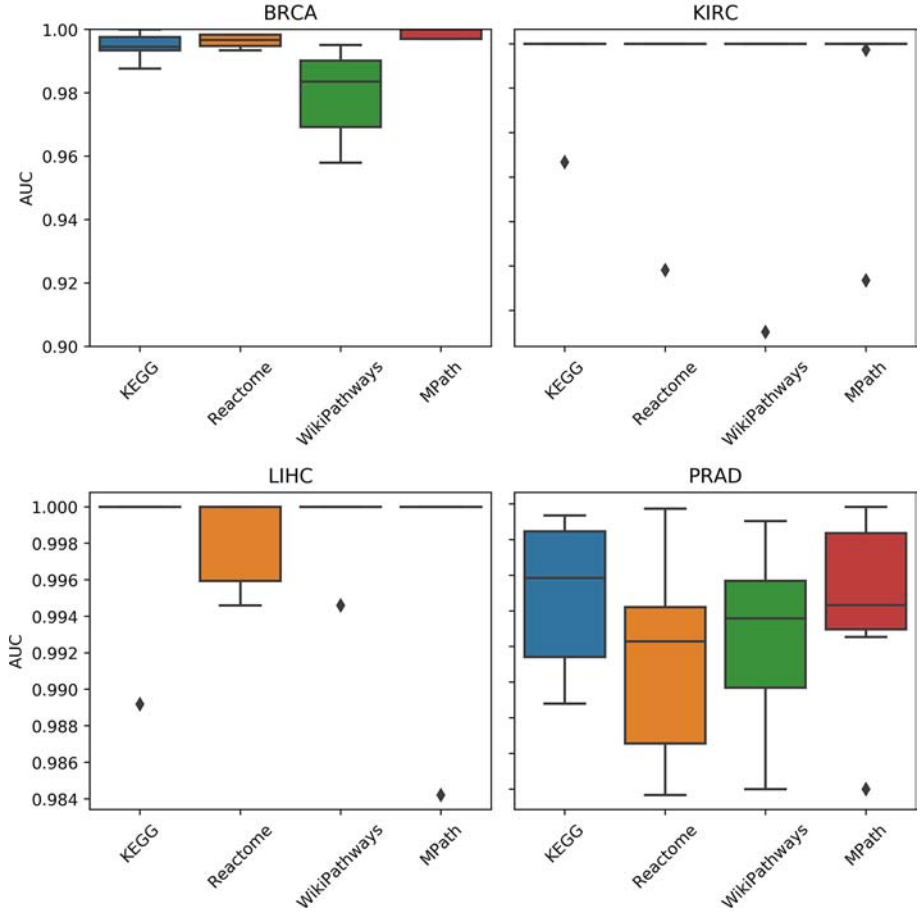


FIGURE 4 | Comparison of prediction performance of an elastic net classifier (tumor vs. normal) using single-sample gene set enrichment analysis (ssGSEA)-based pathway activity profiles computed from different resources. Each box plot shows the distribution of the area under the ROC curves (AUCs) over 10 repeats of the 10-fold cross-validation procedure.

via the AUC and precision-recall curve (see the corresponding *Materials and Methods* section). The AUC indicated no overall significant effect of the choice of pathway database on model prediction performance (p = 0.5, ANOVA F test; **Figure 4**). Similarly, the results of the precision-recall curve did not show a significant effect of the database selected on the model's predictive performance. Finally, these results were not surprising due to the relative ease of the classification task (i.e., all AUC values were close to 1).

Prediction of Tumor Subtype

We next compared the prediction performances of a multi-class classifier predicting known tumor subtypes of BRCA and PRAD using ssGSEA-based pathway activity profiles. **Figure 5** demonstrated no overall significant effect of the choice of pathway database (p = 0.16, ANOVA F test). We used Wilcoxon tests to investigate if each pair of distributions of the accuracies based on each database were different, but did not achieve statistical significance ($q < 0.01$) after Benjamini–Hochberg correction for multiple hypothesis testing. While the lack of significance is probably due to the limited amount of datasets (only two contained subtype information) and measurements, we would like to note that MPath showed the best classification metrics (similar to the previous classification task).

Prediction of Overall Survival

As a next step, we compared the prediction performance of an elastic net penalized Cox regression model for overall survival using ssGSEA-based pathway activity profiles derived from different resources. As indicated in **Figure 6**, no overall significant effect of the actually used pathway database could be observed (p = 0.28, ANOVA F test). A limiting factor of this analysis is the fact that overall survival can generally only be predicted slightly above chance level (c-indices range between 55% and 60%) based on gene expression alone, which is in agreement with the

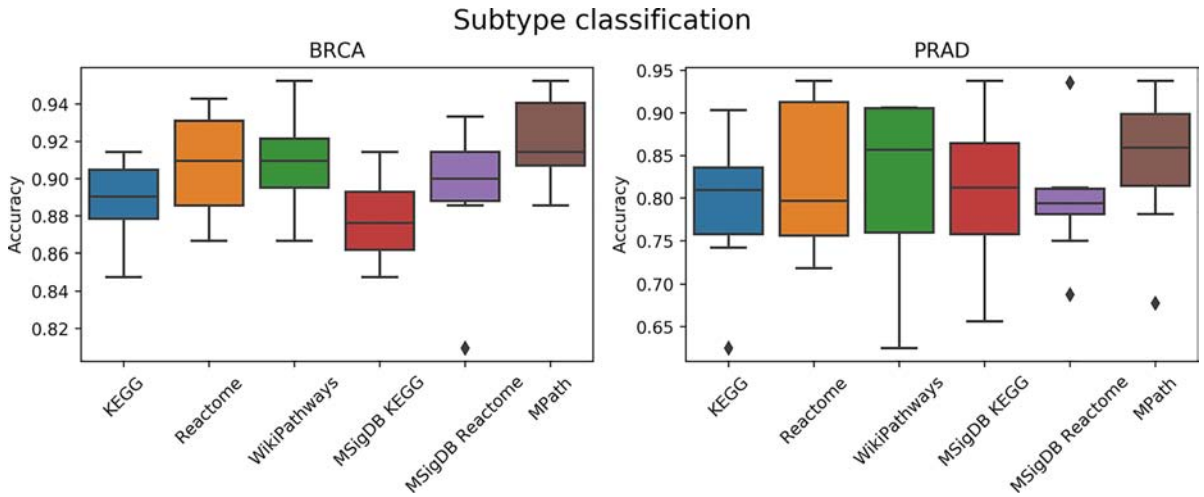


FIGURE 5 | Comparison of prediction performance of an elastic net classifier (BRCA and PRAD subtypes) using single-sample gene set enrichment analysis (ssGSEA)-based pathway activity profiles computed from different resources. Each box plot shows the distribution of the area under the ROC curves (AUCs) over 10 repeats of the 10-fold cross-validation procedure.

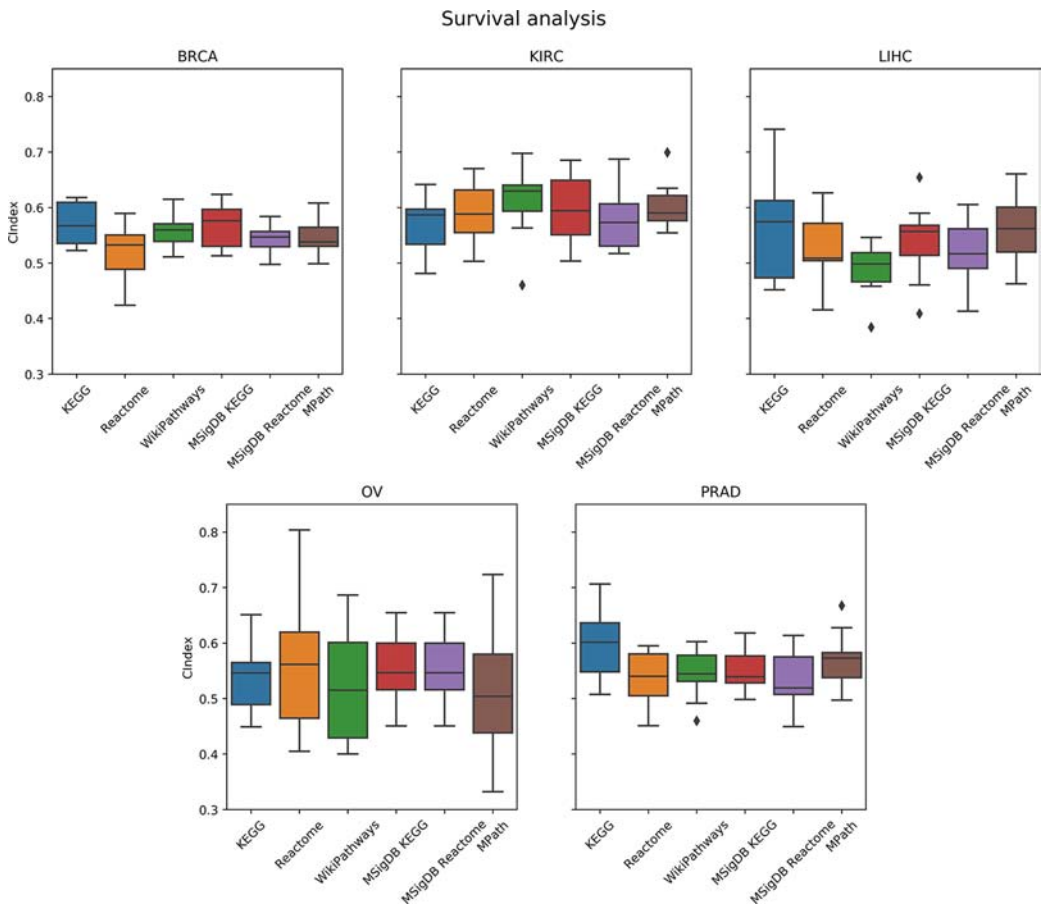


FIGURE 6 | Comparison of prediction performance of an elastic net penalized Cox regression model (overall survival) using single-sample gene set enrichment analysis (ssGSEA)-based pathway activity profiles computed from different resources. Each box plot shows the distribution of the area under the ROC curves (AUCs) over 10 repeats of the 10-fold cross-validation procedure.

literature (Van Wieringen et al., 2009; Fröhlich, 2014; Mayr and Schmid, 2014; Zhang et al., 2018).

DISCUSSION

In this work, we presented a comprehensive comparative study of pathway databases based on functional enrichment and predictive modeling. We have shown that the choice of pathway database can significantly influence the results of statistical enrichment, which raises concerns about the typical lack of consideration that is given to the choice of pathway resource in many gene expression studies. This finding was specifically pronounced for SPIA because this method is a topology-based enrichment approach and therefore expected to be most sensitive to the actual definition of a pathway. At the same time, we observed that an integrative pathway resource (MPath) led to more biologically consistent results and, in some cases, improved prediction performance.

Generating a merged dataset such as MPath is non-trivial. We purposely restricted this study to three major pathway databases because of the availability of inter-database pathway mappings and pathway networks from our previous work which enabled conducting objective database comparisons. The incorporation of additional pathway databases into MPath would first require the curation of pathway mappings prior to conducting the benchmarking study, which can be labor-intensive. Furthermore, performing the tasks described in this work comes with a high computational cost (**Table 1**).

Our strategy to build MPath is one of many possible approaches to integrate pathway knowledge from multiple databases. Although alternative meta-databases such as Pathway Commons and MSigDB do exist, the novelty of this work lies in the usage of mappings and harmonized pathway representations for generating a merged dataset. While we have presented MPath as one possible integrative approach, alternative meta-databases may be used, but would require that researchers ensure that the meta-databases' contents are continuously updated (Wadi et al., 2016).

Our developed mapping strategy between different graph representations of analogous pathways enabled us to objectively compare pathway enrichment results that otherwise would have been conducted manually and subjectively. Furthermore, they allowed us to generate super pathways inspired by previous approaches that have shown the benefit of merging similar pathway representations (Doderer et al., 2012; Vivar et al., 2013; Belinky et al., 2015; Stoney et al., 2018; Miller et al., 2019). In this case, this was made possible by the fully harmonized gene sets and networks generated by our previous work, ComPath and PathMe. A detailed description of the ComPath and PathMe publications, source code, and extensions to existing analyses (i.e., SPIA) to better suit the methods used in this work can be found in the **Supplementary Text**.

One of the limitations of this work is that we restricted the analysis to five cancer datasets from TCGA and we did not expand it to other conditions besides cancer. The use of this disease area was mainly driven by the availability of data and the corresponding possibilities to draw statistically valid conclusions. However, we acknowledge the fact that data from other disease areas may result in different findings. More specifically, we believe that a similar benchmarking study based on data from disease conditions with an unknown pathophysiology (e.g., neurological disorders) may yield even more pronounced differences between pathway resources. Additionally, further techniques for gene expression-based pathway activity scoring could be incorporated, such as Pathifier or SAS (Drier et al., 2013; Lim et al., 2016).

DATA AVAILABILITY STATEMENT

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

AUTHOR CONTRIBUTIONS

DD-F conceived and designed the study. SM and DD-F conducted the main analysis and implemented the Python package. HF supervised methodological aspects of the analysis. CH and AG assisted technically in the analysis of the results. MH-A acquired the funding. SM, HF, CH, MH-A, and DD-F wrote the paper.

FUNDING

This work was supported by the EU/EFPIA Innovative Medicines Initiative Joint Undertaking under AETIONOMY (grant number 115568), resources of which are composed of financial contribution from the European Union's Seventh Framework Programme (FP7/2007-2013) and EFPIA companies in kind contribution.

ACKNOWLEDGMENTS

The authors would like to thank Mohammad Asif Emon for his assistance in conducting SPIA and Jan-Eric Bökenkamp for his assistance in processing the TCGA datasets. Furthermore, we would like to thank Jonas Klees and Carina Steinborn for generating the visuals in this paper. Finally, we would like to thank the curators of KEGG, Reactome, and WikiPathways as well as the TCGA network for generating the pathway content and datasets used in this work, respectively.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.01203/full#supplementary-material

REFERENCES

Bader, G. D., Cary, M. P., and Sander, C. (2006). Pathguide: a pathway resource list. *Nucleic Acids Res.* 34 (suppl_1), D504–D506. doi: 10.1093/nar/gkj126

Barbie, D. A., Tamayo, P., Boehm, J. S., Kim, S. Y., Moody, S. E., Dunn, I. F., et al. (2009). Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. *Nature* 462 (7269), 108. doi: 10.1038/nature08460

Bateman, A. R., El-Hachem, N., Beck, A. H., Aerts, H. J., and Haibe-Kains, B. (2014). Importance of collection in gene set enrichment analysis of drug response in cancer cell lines. *Sci. Rep.* 4, 4092. doi: 10.1038/srep04092

Bayerlová, M., Jung, K., Kramer, F., Klemm, F., Bleckmann, A., and Beißbarth, T. (2015). Comparative study on gene set and pathway topology-based enrichment methods. *BMC Bioinf.* 16 (1), 334. doi: 10.1186/s12859-015-0751-5

Belinky, F., Nativ, N., Stelzer, G., Zimmerman, S., Iny Stein, T., Safran, M., et al. (2015). PathCards: multi-source consolidation of human biological pathways. *Database* 2015. doi: 10.1093/database/bav006

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. (Methodological)* 57 (1), 289–300. doi: 10.2307/2346101

Benjamini, Y., and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Ann. Stat.* 29 (4), 1165–1188. doi: 10.1214/aos/1013699998

Cancer Genome Atlas Research Network. (2014). Comprehensive molecular characterization of gastric adenocarcinoma. *Nature* 513 (7517), 202. doi: 10.1038/nature13480

Cerami, E. G., Gross, B. E., Demir, E., Rodchenkov, I., Babur, O., Anwar, N., et al. (2011). Pathway commons, a web resource for biological pathway data. *Nucleic Acids Res.* 39 (Suppl. 1), D685–D690. doi: 10.1093/nar/gkq1039

Coates, A. S., Winer, E. P., Goldhirsch, A., Gelber, R. D., Gnant, M., Piccart-Gebhart, M., et al. (2015). Tailoring therapies—improving the management of early breast cancer: St Gallen International Expert Consensus on the Primary Therapy of Early Breast Cancer 2015. *Ann. Oncol.* 26 (8), 1533–1546. doi: 10.1093/annonc/mdv221

Colaprico, A., Silva, T. C., Olsen, C., Garofano, L., Cava, C., Garolini, D., et al. (2015). TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Res.* 44 (8), e71–e71. doi: 10.1093/nar/gkv1507

Cortes, C., and Vapnik, V. (1995). Support-vector networks. *Mach. Lear.* 20 (3), 273–297. doi: 10.1007%2FBF00994018

Doderer, M. S., Anguiano, Z., Suresh, U., Dashnamoorthy, R., Bishop, A. J., and Chen, Y. (2012). Pathway Distiller-multisource biological pathway consolidation. *BMC Genom.* 13 (6), S18. doi: 10.1186/1471-2164-13-S6-S18

Domingo-Fernández, D., Hoyt, C. T., Bobis-Álvarez, C., Marin-Llao, J., and Hofmann-Apitius, M. (2018). ComPath: an ecosystem for exploring, analyzing, and curating mappings across pathway databases. *NPJ Syst. Biol. Appl.* 4 (1), 43. doi: 10.1038/s41540-018-0078-8

Domingo-Fernandez, D., Mubeen, S., Marin-Llao, J., Hoyt, C., and Hofmann-Apitius, M. (2019). PathMe: merging and exploring mechanistic pathway knowledge. *BMC Bioinf.* 20, 243. doi: 10.1186/s12859-019-2863-9

Drier, Y., Sheffer, M., and Domany, E. (2013). Pathway-based personalized analysis of cancer. *Proc. Nat. Acad. Sci.* 110 (16), 6388–6393. doi: 10.1073/pnas.1219651110

Fabregat, A., Jupe, S., Matthews, L., Sidiropoulos, K., Gillespie, M., Garapati, P., et al. (2018). The reactome pathway knowledgebase. *Nucleic Acids Res.* 46 (D1), D649–D655. doi: 10.1093/nar/gkx1132

Fabris, F., Palmer, D., de Magalhães, J. P., and Freitas, A. A. (2019). Comparing enrichment analysis and machine learning for identifying gene properties that discriminate between gene classes. *Briefings Bioinf.* doi: 10.1093/bib/bbz028

Fisher, R. A. (1992). Statistical methods for research workers in *Breakthroughs in Statistics* (New York, NY:Springer), 66–70.

Fröhlich, H. (2014). Including network knowledge into Cox regression models for biomarker signature discovery. *Biom. J.* 56 (2), 287–306. doi: 10.1002/bimj.201300035

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models *via* coordinate descent. *J. Stat. Software* 33 (1), 1. doi: 10.18637/jss.v033.i01

García-Campos, M. A., Espinal-Enríquez, J., and Hernández-Lemus, E. (2015). Pathway analysis: state of the art. *Front. Physiol.* 6, 383. doi: 10.3389/fphys.2015.00383

Grüning, B. A., Lampa, S., Vaudel, M., and Blankenberg, D. (2019). Software engineering for scientific big data analysis. *GigaScience* 8 (5), giz054. doi: 10.1093/gigascience/giz054

Graudenzi, A., et al. (2017). Pathway-based classification of breast cancer subtypes. *Front. Biosci., (Landmark Ed)* 22, 1697–1712. doi: 10.2741/4566

Harrell, F. E., Califf, R. M., Pryor, D. B., Lee, K. L., and Rosati, R. A. (1982). Evaluating the yield of medical tests. *JAMA* 247 (18), 2543–2546. doi: 10.1001/jama.1982.03320430047030

Hoyt, C. T., Konotopez, A., and Ebeling, C. (2018). PyBEL: a computational framework for Biological Expression Language. *Bioinformatics* 34 (4), 703–704. doi: 10.1093/bioinformatics/btx660

Hoyt, C. T., Domingo-Fernández, D., Mubeen, S., Llaó, J. M., Konotopez, A., Ebeling, C., et al. (2019). Integration of Structured Biological Data Sources using Biological Expression Language. *Biorxiv* 631812. doi: 10.1101/631812

Ihnatova, I., Popovici, V., and Budinska, E. (2018). A critical comparison of topology-based pathway analysis methods. *PloS One* 13 (1), e0191154. doi: 10.1371/journal.pone.0191154

Kamburov, A., Wierling, C., Lehrach, H., and Herwig, R. (2008). ConsensusPathDB—a database for integrating human functional interaction networks. *Nucleic Acids Res.* 37 (suppl_1), D623–D628. doi: 10.1093/nar/gkn698

Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., and Morishima, K. (2016). KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.* 45 (D1), D353–D361. doi: 10.1093/nar/gkw1092

Khatri, P., Sirota, M., and Butte, A. J. (2012). Ten years of pathway analysis: current approaches and outstanding challenges. *PloS Comput. Biol.* 8 (2), e1002375. doi: 10.1371/journal.pcbi.1002375

Kirouac, D. C., Saez-Rodriguez, J., Swantek, J., Burke, J. M., Lauffenburger, D. A., and Sorger, P. K. (2012). Creating and analyzing pathway and protein interaction compendia for modelling signal transduction networks. *BMC Syst. Biol.* 6 (1), 29. doi: 10.1186/1752-0509-6-29

Liberzon, A., Birger, C., Thorvaldsdóttir, H., Ghandi, M., Mesirov, J. P., and Tamayo, P. (2015). The molecular signatures database hallmark gene set collection. *Cell Syst.* 1 (6), 417–425. doi: 10.1016/j.cels.2015.12.004

Lim, S., Lee, S., Jung, I., Rhee, S., and Kim, S. (2018). Comprehensive and critical evaluation of individualized pathway activity measurement tools on pan-cancer data. *Briefings Bioinf.*

Lim, S., Park, Y., Hur, B., Kim, M., Han, W., and Kim, S. (2016). Protein interaction network (pin)-based breast cancer subsystem identification and activation measurement for prognostic modeling. *Methods* 110, 81–89. doi: 10.1016/j.ymeth.2016.06.015

Mayr, A., and Schmid, M. (2014). Boosting the concordance index for survival data–a unified framework to derive and evaluate biomarker combinations. *PloS One* 9 (1), e84483. doi: 10.1371/journal.pone.0084483

McKinney, W. (2010). Data Structures for Statistical Computing in Python in *Proceedings of the 9th Python in Science Conference*. Eds.van der Walt, S., and Millman, J., 51–56.

Miller, R. A., Ehrhart, F., Eijssen, L. M., Slenter, D. N., Curfs, L. M., Evelo, C. T., et al. (2019). Beyond pathway analysis: Identification of active subnetworks in Rett syndrome. *Front. Genet.* 10, 59. doi: 10.3389/fgene.2019.00059

Molinaro, A. M., Simon, R., and Pfeiffer, R. M. (2005). Prediction error estimation: a comparison of resampling methods. *Bioinformatics* 21 (15), 3301–3307. doi: 10.1093/bioinformatics/bti499

Povey, S., Lovering, R., Bruford, E., Wright, M., Lush, M., and Wain, H. (2001). The HUGO gene nomenclature committee (HGNC). *Hum. Genet.* 109 (6), 678–680. doi: 10.1007/s00439-001-0615-0

Reimand, J., Isserlin, R., Voisin, V., Kucera, M., Tannus-Lopes, C., Rostamianfar, A., et al. (2019). Pathway enrichment analysis and visualization of omics data using g: Profiler, GSEA, Cytoscape and EnrichmentMap. *Nat. Protoc.* 14 (2), 482–517. doi: 10.1038/s41596-018-0103-9

Sales, G., Calura, E., and Romualdi, C. (2018). meta Graphite–a new layer of pathway annotation to get metabolite networks. *Bioinformatics* 35 (7), 1258–1260. doi: 10.1093/bioinformatics/bty719

Schaefer, C. F., Anthony, K., Krupa, S., Buchoff, J., Day, M., Hannay, T., et al. (2008). PID: the pathway interaction database. *Nucleic Acids Res.* 37 (suppl_1), D674–D679. doi: 10.1093/nar/gkn653

Senkus, E., Kyriakides, S., Ohno, S., Penault-Llorca, F., Poortmans, P., Rutgers, E., et al. (2015). Primary breast cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. *Ann. Oncol.* 26 (suppl_5), v8–v30. doi: 10.1093/annonc/mdv298

Slater, T. (2014). Recent advances in modeling languages for pathway maps and computable biological networks. *Drug Discovery Today* 19 (2), 193–198. doi: 10.1016/j.drudis.2013.12.011

Slenter, D. N., Kutmon, M., Hanspers, K., Riutta, A., Windsor, J., Nunes, N., et al. (2017). WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research. *Nucleic Acids Res.* 46 (D1), D661–D667. doi: 10.1093/nar/gkx1064

Sorlie, T., Perou, C. M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., et al. (2001). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc. Natl. Acad. Sci.* 98 (19), 10869–10874. doi: 10.1073/pnas.191367098

Stoney, R. A., Schwartz, J. M., Robertson, D. L., and Nenadic, G. (2018). Using set theory to reduce redundancy in pathway sets. *BMC Bioinf.* 19 (1), 386. doi: 10.1186/s12859-018-2355-3

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Nat. Acad. Sci.* 102 (43), 15545–15550. doi: 10.1073/pnas.0506580102

Tarca, A. L., Draghici, S., Khatri, P., Hassan, S. S., Mittal, P., Kim, J. S., et al. (2008). A novel signaling pathway impact analysis. *Bioinformatics* 25 (1), 75–82. doi: 10.1093/bioinformatics/btn577

Tibshirani, R. (1997). The lasso method for variable selection in the Cox model. *Stat. Med.* 16 (4), 385–395. doi: 10.1002/(sici)1097-0258(19970228)16:4<385::aid-sim380>3.0.co;2-3

Türei, D., Korcsmáros, T., and Saez-Rodriguez, J. (2016). OmniPath: guidelines and gateway for literature-curated signaling pathway resources. *Nat. Methods* 13 (12), 966. doi:10.1038/nmeth.4077

Van Wieringen, W. N., Kun, D., Hampel, R., and Boulesteix, A. L. (2009).Survival prediction using gene expression data: a review and comparison. *Comput. Stat. Data Anal.* 53 (5), 1590–1603. doi: 10.1016/j.csda.2008.05.021

Vivar, J. C., Pemu, P., McPherson, R., and Ghosh, S. (2013). Redundancy control in pathway databases (ReCiPa): an application for improving gene-set enrichment analysis in Omics studies and "Big data" biology. *Omics: J. Integr. Biol.* 17 (8), 414–422. doi: 10.1089/omi.2012.0083

Wadi, L., Meyer, M., Weiser, J., Stein, L. D., and Reimand, J. (2016). Impact of outdated gene annotations on pathway enrichment analysis. *Nat. Methods* 13 (9), 705. doi: 10.1038/nmeth.3963

Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R. M., Ozenberger, B. A., Ellrott, et al. (2013). The cancer genome atlas pan-cancer analysis project. *Nat. Genet.* 45 (10), 1113. doi: 10.1038/ng.2764

Zhang, Y., Yang, W., Li, D., Yang, J. Y., Guan, R., and Yang, M. Q. (2018). Toward the precision breast cancer survival prediction utilizing combined whole genome-wide expression and somatic mutation analysis. *BMC Med. Genom.* 11 (5), 104. doi: 10.1109/BIBM.2017.8217762

Zou, H., and Trevor, H. (2005). Regularization and Variable Selection *via the* Elastic Net. *J. R. Stat. Soc Ser. B:* 67 (2), 301–320. doi: 10.1111/j.1467-9868.2005.00503.x

Conflict of Interest: HF received salaries from UCB Biosciences GmbH. UCB Biosciences GmbH had no influence on the content of this work.

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

frontiers in Genetics

CORRECTION
published: 29 April 2020
doi: 10.3389/fgene.2020.00436

Corrigendum: The Impact of Pathway Database Choice on Statistical Enrichment Analysis and Predictive Modeling

Sarah Mubeen[1,2], *Charles Tapley Hoyt*[1,2†], *André Gemünd*[1], *Martin Hofmann-Apitius*[1,2], *Holger Fröhlich*[2] *and Daniel Domingo-Fernández*[1,2*]

[1] Department of Bioinformatics, Fraunhofer Institute for Algorithms and Scientific Computing (SCAI), Sankt Augustin, Germany, [2] Bonn-Aachen International Center for IT, Rheinische Friedrich-Wilhelms-Universität Bonn, Bonn, Germany

Keywords: pathway enrichment, benchmarking, databases, machine learning, statistical hypothesis testing

OPEN ACCESS

Approved by:
Frontiers Editorial Office, Frontiers Media SA, Switzerland

***Correspondence:**
Daniel Domingo-Fernández daniel.domingo.fernandez@scai.fraunhofer.de

†ORCID:
Charles Tapley Hoyt orcid.org/0000-0003-4423-4370

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

Received: *30 March 2020*
Accepted: *08 April 2020*
Published: *29 April 2020*

Citation:
Mubeen S, Hoyt CT, Gemünd A, Hofmann-Apitius M, Fröhlich H and Domingo-Fernández D (2020) Corrigendum: The Impact of Pathway Database Choice on Statistical Enrichment Analysis and Predictive Modeling. Front. Genet. 11:436. doi: 10.3389/fgene.2020.00436

A Corrigendum on

The Impact of Pathway Database Choice on Statistical Enrichment Analysis and Predictive Modeling

by Mubeen, S., Hoyt, C. T., Gemünd, A., Hofmann-Apitius, M., Fröhlich, H., and Domingo-Fernández, D. (2019). Front. Genet. 10:1203. doi: 10.3389/fgene.2019.01203

In the original article, the correspondence email was incorrect. The correct one should be daniel.domingo.fernandez@scai.fraunhofer.de.

The authors apologize for the error and state that this does not change the scientific conclusions of the article in any way. The original article has been updated.

frontiers in Bioengineering and Biotechnology

ORIGINAL RESEARCH
published: 17 December 2019
doi: 10.3389/fbioe.2019.00391

Integrative Bioinformatics Approaches to Map Potential Novel Genes and Pathways Involved in Ovarian Cancer

S. Udhaya Kumar[1], D. Thirumal Kumar[1], R. Siva[1], C. George Priya Doss[1] and Hatem Zayed[2]**

[1] School of Biosciences and Technology, Vellore Institute of Technology, Vellore, India, [2] Department of Biomedical Sciences, College of Health and Sciences, Qatar University, Doha, Qatar

OPEN ACCESS

Edited by:
Manoj Kumar Kashyap, Amity University Gurgaon, India

Reviewed by:
Harrys Kishore Charles Jacob, University of Miami Hospital, United States
Hiren Karathia, CosmosID, Inc., United States

***Correspondence:**
C. George Priya Doss
georgepriyadoss@vit.ac.in
Hatem Zayed
hatem.zayed@qu.edu.qa

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Bioengineering and Biotechnology

Received: 05 August 2019
Accepted: 19 November 2019
Published: 17 December 2019

Citation:
Kumar SU, Kumar DT, Siva R, Doss CGP and Zayed H (2019) Integrative Bioinformatics Approaches to Map Potential Novel Genes and Pathways Involved in Ovarian Cancer. Front. Bioeng. Biotechnol. 7:391. doi: 10.3389/fbioe.2019.00391

Background and aims: Ovarian cancer (OC) is the seventh most commonly detected cancer among women. This study aimed to map the hub and core genes and potential pathways that might be involved in the molecular pathogenesis of OC.

Methods: In the present work, we analyzed a microarray dataset (GSE126519) from the Gene Expression Omnibus (GEO) database and used the GEO2R tool to screen OC cells and ovarian SINE-resistant cancer cells for differentially expressed genes (DEGs). For the functional annotation of the DEGs, we conducted Gene Ontology (GO) and pathway enrichment analyses (KEGG) using the DAVID v6.8 online server and GenoGo Metacore™, Cortellis Solution software. Protein–protein interaction (PPI) networks were constructed using the STRING database, and Cytoscape software was used for visualization. The survival analysis was performed using the online platform GEPIA2 to determine the prognostic value of the expression of hub genes in cell lines from OC patients.

Results: We identified a total of 809 upregulated and 700 downregulated DEGs. GO analysis revealed that the genes with statistically significant differences in expression were mainly associated with biological processes involved in the cell cycle, the mitotic cell cycle, mitotic nuclear division, organ morphogenesis, cell development, and cell morphogenesis. By using the Analyze Networks (AN) algorithm in GeneGo, we identified the most relevant biological networks involving DEGs that were mainly enriched in the cell cycle (in metaphase checkpoints) and revealed the role of APC in cell cycle regulation pathways. We found 10 hub genes and four core genes (*FZD6*, *FZD8*, *CDK2*, and *RBBP8*) that are strongly linked to OC.

Conclusion: This study sheds light on the molecular pathogenesis of OC and is expected to provide potential molecular biomarkers that are beneficial for the treatment and clinical molecular diagnosis of OC.

Keywords: ovarian cancer, protein–protein interactions, Metacore, biomarkers, functional enrichment analysis, expression profiling data, microarray

INTRODUCTION

Ovarian cancer (OC) is the seventh most frequently detected cancer among women worldwide (Reid et al., 2017). Epithelial cancers represent ∼90% of OC in patients with different ailments (Cancer Genome Atlas Research Network, 2011) comprising five distinct histological subtypes that have various distinguishable complications, sources of cells, molecular compositions, clinical signs, and symptoms and treatments (Matulonis et al., 2016). Matulonis et al. (2016) reported that OC is typically detected at the late stage, and no successful screening approaches have been found thus far. However, patients with an increased risk of OC with germline mutations in *BRCA1*, *BRCA2*, or additional genes can be identified (Pennington and Swisher, 2012; Younes and Zayed, 2019).

Proteins transported by exportin 1 (XPO1 or CMR1), such as IkB, p53, pRb, p21, p27, and FOXO, play significant roles as tumor suppressors. When restricted to the nucleus, they inhibit the growth of cells and cell survival unless they are transferred to the cytoplasm (Senapedis et al., 2014). A selective inhibitor of nuclear export (SINE) acts along with CMR1 to block its interaction with nuclear proteins intended to be exported to the cytoplasm; inhibitors of CMR1 are known as SINE compounds (Gerecitano, 2014). Recent work has also revealed that SINE compounds enhance the proteasomal deterioration of CMR1, increase the nuclear retention of FOXO and p53, and contribute to enhanced apoptosis in prostate cancer cell lines (Mendonca et al., 2014). As an outcome of resistance to treatment, the elevated annual mortality rate is due to a variety of diagnoses at advanced phases of the disease and recurrence of the disease. Additionally, OC comprises several subtypes with distinct etiologies and molecular profiles that result in considerable variations in the inherent sensitivity to treatment (Zyl et al., 2018). To overcome treatment resistance, there is a need to understand the complete set of molecular mechanisms underlying SINE resistance in OC cell lines. Therefore, the development of OC and associated phenomena need to be investigated, and there is an urgent need to find candidate early diagnostic biomarkers.

Microarray-based gene expression assessment is the most commonly used high-throughput and successful technique used to study complicated disease pathogenesis. However, studies performed that utilize human OC gene expression profiling are very uncommon. In the current research, we tried to explore the differentially expressed genes (DEGs), gene network, pathways, and protein interactions that are unique to OC. To detect the DEGs between OC and SINE-resistant OC cell lines (GSE126519), we utilized a bioinformatics approach to analyze DEG data retrieved from the Gene Expression Omnibus (GEO) database. For the screened DEGs, functional annotation assessment with Gene Ontology (GO) and pathway enrichment assessment with the Kyoto Encyclopedia of Genes and Genomes (KEGG) were carried out using the Database for Annotation, Visualization, and Integrated Discovery and GeneGo Metacore™ software. Ultimately, we found 10 potential hub genes and four core genes that were strongly linked to OC.

MATERIALS AND METHODS

Data Preprocessing and Screening of DEGs

The expression profiling was performed on the OC gene dataset GSE126519, which was retrieved from GEO (Gene Expression Omnibus database, https://www.ncbi.nlm.nih.gov/geo/) and includes gene expression datasets from RNA-seq, high-throughput hybridization array, DNAseq, ChIPs, and microarray (Barrett et al., 2013). "Ovarian cancer" AND "Homo sapiens" were the keywords used to search OC-related expression profiles within the GEO datasets. The GSE126519 expression profiling was conducted in arrays that included three human OC cell lines and three SINE (selective inhibitors of nuclear export)-resistant human OC cell lines. We utilized the GEO2R (http://www.ncbi.nlm.nih.gov/geo/geo2r/) statistical tool to recalculate and assess the genes that were expressed differently between the human OC cell lines and the SINE-resistant human OC cell lines (Ritchie et al., 2015). The Benjamini and Hochberg (false discovery rate) and *t*-test methods were utilized with the GEO2R tool to calculate the FDR and *p*-values, respectively, to identify the DEGs. We considered $p < 0.05$ and a $|\log(\text{fold change})| > 1$ to be statistically significant for the DEGs, and $\text{logFC} \geq 1$ and $\text{logFC} \leq -1$ were considered to indicate upregulated and downregulated DEGs, respectively (Aubert et al., 2004).

By using all of the DEGs identified in the OC cell lines, we constructed a volcano plot by using the Volcano Plot (https://paolo.shinyapps.io/ShinyVolcanoPlot/) online server, which is hosted on shinyapps.io by RStudio. The resultant DEG dataset was collected and used for further analysis. The initial ontology of gene (GO) and KEGG pathway enrichment analyses of the DEGs was annotated ($p < 0.05$) using the online bioinformatics tool DAVID v6.8 (https://david.ncifcrf.gov/) (Huang et al., 2009a,b).

PPI Network Construction

The online database STRING (v11.0, http://www.string-db.org/) was used to visualize the PPIs between the statistically significant DEG-encoded proteins in the resultant dataset (Szklarczyk et al., 2015). The dataset contained more than 10,000 DEGs. To avoid an inaccurate PPI network, we used a cutoff ≥ 0.9 (high-confidence interaction score) to obtain the significant PPIs. We used Cytoscape software v3.7.1 (http://www.cytoscape.org/) to visualize the PPI network obtained from the STRING database (Shannon et al., 2003). Based on the log fold change values, the PPI network was plotted for both the upregulated and downregulated DEGs. The interrelation analysis of the identified genes was performed by using the GeneMANIA online tool (Franz et al., 2018).

Analyzing the Backbone Network

The NetworkAnalyzer app in Cytoscape was utilized to explore the networks of both the upregulated and downregulated DEGs (Saito et al., 2012). NetworkAnalyzer computes the topological parameters and centrality measures such as the distribution of the node degree, the betweenness centrality, the topological coefficients, the shortest path length, and the closeness centrality for directed and undirected networks (Assenov et al., 2008).

TABLE 1 | Patients' information in GSE126519 derived from the GEO database.

Group	Accession	Patient no.	Organism	Disease state	Type
OC	GSM3602932	Patient 1	*Homo sapiens*	Ovarian cancer	Human ovarian cancer cell
	GSM3602933	Patient 2	*Homo sapiens*	Ovarian cancer	Human ovarian cancer cell
	GSM3602934	Patient 3	*Homo sapiens*	Ovarian cancer	Human ovarian cancer cell
SINE resistance OC	GSM3602935	Patient 4	*Homo sapiens*	Ovarian cancer	Human ovarian cancer cell
	GSM3602936	Patient 5	*Homo sapiens*	Ovarian cancer	Human ovarian cancer cell
	GSM3602937	Patient 6	*Homo sapiens*	Ovarian cancer	Human ovarian cancer cell

GEO, Gene Expression Omnibus; OC, ovarian cancer; SINE, selective inhibitors of nuclear export.

The distribution of the node degree indicates the number of nodes with a certain degree and is a comparative measure of the degree to which a node parameter shares neighbors with other nodes in terms of the topological coefficient. NetworkAnalyzer calculates the topological coefficients for all network nodes with more than one neighbor (Stelzl et al., 2005). The networks that do not have multiple edges have been determined according to the betweenness centrality, whereas the closeness centrality computes this for all nodes and plots it against the number of neighbors in terms of the closeness centrality (Brandes, 2001; Newman, 2005).

GeneGo Analysis

The statistically significant DEGs were further analyzed in Metacore and Cortellis Solution software (https://clarivate.com/products/metacore/, Clarivate Analytics, London, UK) to perform the GO function and pathway enrichment analyses. GeneGo enables the fast analysis of protein networks, metabolic pathways, and maps for the list of genes and proteins obtained from experimental high-throughput data (MetaCore Login|Clarivate Analytics[1]). We used the pathway maps tool to identify the enriched pathways involving DEGs in terms of the hypergeometric distribution, and the *p*-values were calculated by using the default database as the background (based on an FDR $p < 0.005$). Based on a significant p-value < 0.05, graphical depictions of the molecular interactions between the DEGs were generated.

Hub Gene Survival Analysis

A comprehensive online platform called Gene Expression Profiling Interactive Analysis (GEPIA2, http://gepia2.cancer-pku.cn/#index) provides fast and customized delivery of functionalities based on TCGA (The Cancer Genome Atlas) and genotype-tissue expression (GTEx) data. GEPIA2 evaluates the survival effect of differentially expressed genes in a given cancer sample. The overall survival effect of hub genes in OC was estimated by calculating the log-rank *p*-value and the HR (hazard ratio-95% confidence interval) using GEPIA2 Single Gene Analysis. Then, the validation of the expression of the core hub genes in OC and normal tissues was performed and visualized in a boxplot (Tang et al., 2017).

[1]MetaCore Login | Clarivate Analytics Available at: https://portal.genego.com/ (accessed June 22, 2019).

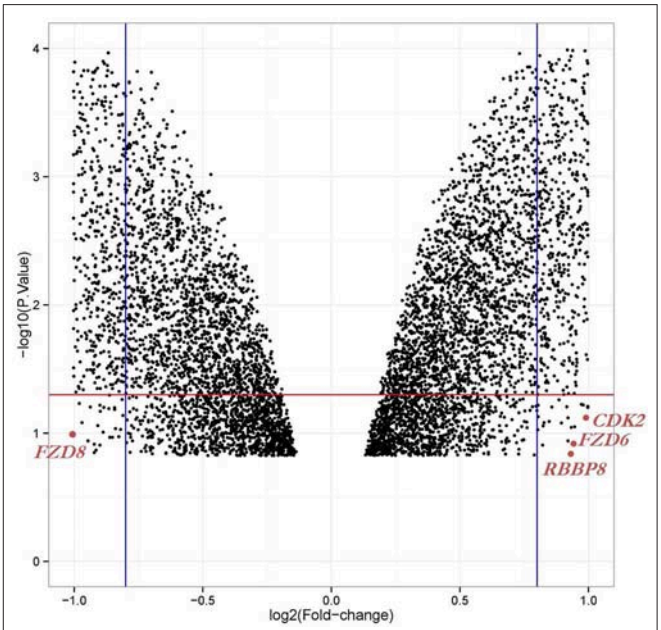


FIGURE 1 | Volcano plot of the DEGs in OC compared with those in SINE-resistant OC from the GSE126519 dataset. *x*-axis: log2FC, large-scale fold changes; *y*-axis: –log10 of the *p*-value showing the statistical significance. Each black point corresponds to one gene. The black points above the red line and beside the blue line (left side and right side) represent log2FC ≥ 1 and p-value < 0.05; all DEGs that were OC-related are shown below the red line, which represents log2FC < 1 or p-value > 0.05. The position of the core genes are named and marked in scarlet color.

RESULTS

DEG Identification

We obtained the gene expression profiles for GSE126519, "Analysis of RNA profiles in parent and selective inhibitors of nuclear export (SINE)-resistant OC cells" from the GEO datasets. (Miyake and Sood, 2019) submitted the GSE126519 dataset, which was generated on the GPL10558 platform (Illumina HumanHT-12 V4.0 expression bead chip). The GSE126519 dataset was obtained from three patient cell lines that comprised six samples, including three OC cell lines and three SINE-resistant OC cell lines (**Table 1**). To identify the DEGs from these two groups (OC and SINE-resistant OC), we conducted GEO2R web-server analysis (https://www.ncbi.nlm.nih.gov/geo/

geo2r/?acc=GSE126519) to calculate the *p*-values and |log2FC| values. The resulting genes that met the cutoff criteria (|log2FC| $\geq$ 1.0 and $p < 0.05$) were considered DEGs. Overall, 8,855 genes were identified from the GEO dataset (GSE126519) with $p > 0.05$ and $p < 0.05$ using the GEO2R tool and are shown in **Supplementary Table 1**. We constructed a volcano plot using the Shiny Volcano Plot online server by Rstudio to compare the two groups; a total of 2708 DEGs were identified from the GSE126519 dataset (**Figure 1**). Among them, 809 and 700 genes were upregulated and downregulated, respectively, between two groups according to their log2FC and *p*-values (**Supplementary Table 2**).

Construction of the PPI Network

To evaluate the PPIs between the DEGs, we used the STRING tool to identify the PPI networks for both the up- and downregulated genes. A combined score of $\geq$0.9 for the nodes was considered to indicate a significant PPI interaction. Then, we exported the resulting PPI network as a ".txt" file and imported it as a.csv file into Cytoscape v3.7.1 software for visualization.

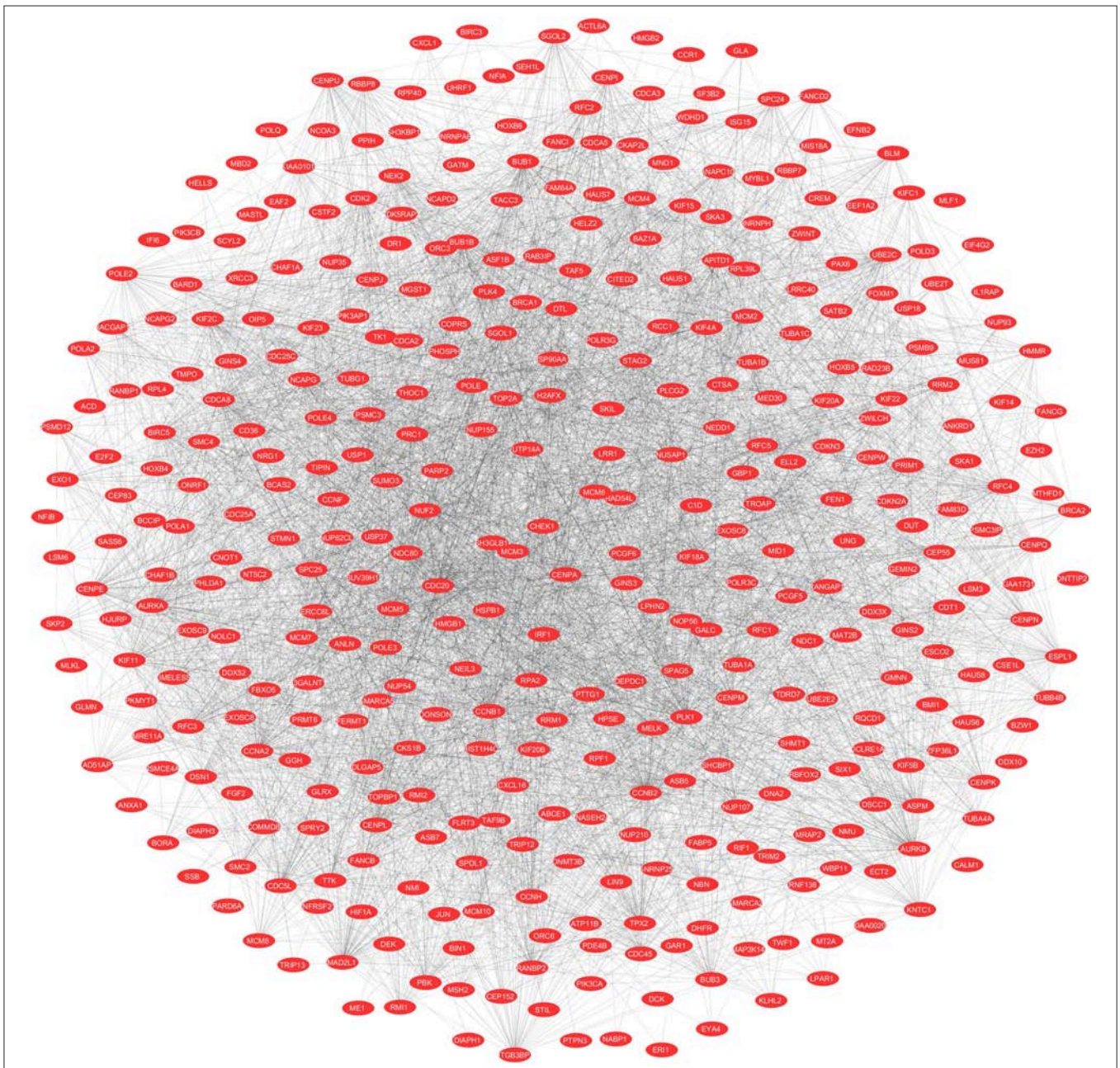


FIGURE 2 | The protein networks of the upregulated DEGs determined using Cytoscape are shown. The representation is as follows: spheres represent the nodes, and the edges are shown as lines.

The graphical representations of the PPI networks of the up- and downregulated DEGs are shown in **Figures 2**, **3**, respectively. The backbone network of the up- and downregulated genes consist of 794 nodes and 676 nodes with estimated clustering coefficients of 0.321 and 0.192, respectively. The Cytoscape plug-in Network Analyzer was used to analyze the networks for both the up- and downregulated DEGs. **Table 2** shows the topological parameters of the up- and downregulated PPI networks, and the topological components, including the distribution of the node degree, the topological coefficient, the shortest path length distribution, the betweenness centrality, and the closeness centrality for the individual PPI networks are shown in **Figures 4A,B**.

GO and Enrichment Analysis

To determine the potential GO classifications and KEGG pathway-enriched genes from the dataset, we imported all target DEGs into the online analysis tool DAVID to conduct the annotation process (**Supplementary Table 3**). The annotated

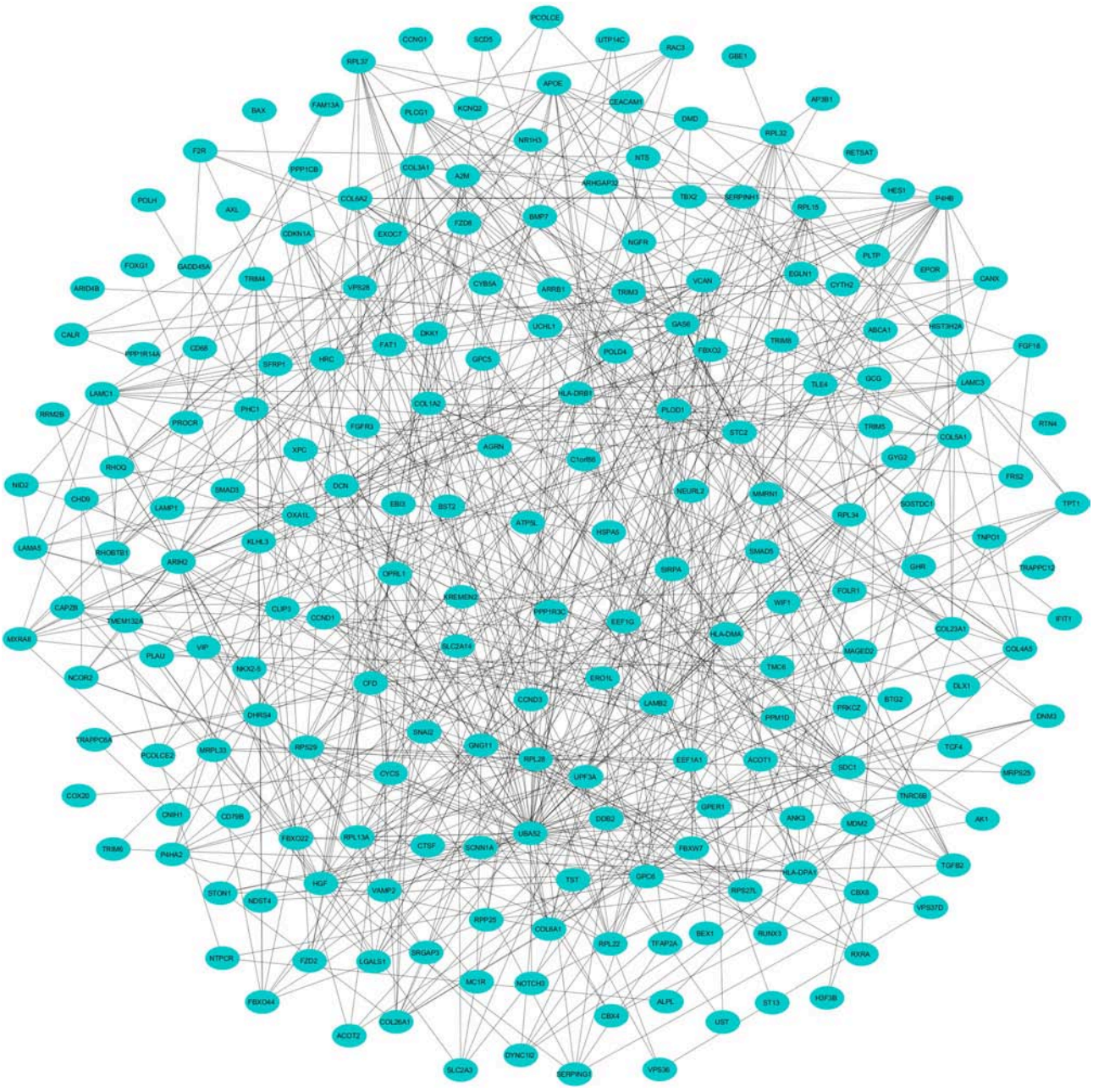


FIGURE 3 | The protein networks of the downregulated DEGs determined using Cytoscape are shown. The representation is as follows: spheres represent the nodes, and the edges are shown as lines.

TABLE 2 | Topological parameters for the upregulated and downregulated PPI network.

S. no.	Topological parameters	Comprehended values	
		Upregulated genes	Downregulated genes
(1)	Number of nodes	794	676
(2)	Clustering co-efficient	0.321	0.192
(3)	Network density	0.023	0.003
(4)	Network centralization	0.169	0.081
(5)	Network heterogeneity	1.748	1.893
(6)	Characteristic path length	3.295	4.509
(7)	Average number of neighbors	18.544	2.257

results for the GO terms were divided according to the MF (molecular function), BP (biological process), and CC (cell component) ontologies ($p < 0.05$, FDR < 0.05). The results of the GO biological process (BP) analysis revealed that the upregulated DEGs were mainly enriched in the cell cycle, mitotic cell cycle process, and mitotic nuclear division; the downregulated DEGs were mainly elevated in pathways related to organ morphogenesis, cell development, and cell morphogenesis, which are involved in differentiation, mesenchymal development, and cellular responses to UV. For the GO molecular function analysis, the upregulated DEGs were significantly enriched in nucleoside-triphosphatase activity and hydrolase activity, which acts on acid anhydrides and phosphorus-containing anhydrides, DNA-dependent ATPase activity, and pyrophosphatase activity, whereas the downregulated DEGs were largely enriched in beta-amyloid binding, carbonyl reductase (NADPH) activity, and collagen, amide, and calcium ion binding. Concerning the GO cell component analysis, the upregulated DEGs were mostly enriched in the chromosome and condensed chromosome, while the downregulated DEGs were enriched in membrane-bound vesicles, extracellular vesicles, and the extracellular region and organelles (**Supplementary Table 3**). Moreover, we used the DAVID online tool to categorize the DEGs involved in various signaling pathways according to the KEGG reference pathways ($p < 0.05$, FDR < 0.05). By examining the KEGG pathways, we noticed that the upregulated DEGs were enriched in DNA replication, the cell cycle, the nucleotide excision repair mechanism, the Fanconi anemia pathway, and DNA mismatch repair; the downregulated DEGs were mostly enriched in the ECM–receptor interaction, the PI3K–Akt signaling pathway, arginine and proline metabolism, Oligodontia-colorectal cancer syndrome, and Nevoid basal cell carcinoma syndrome (**Supplementary Table 4**).

Enrichment Analysis Using Metacore™ Software

To understand the map pathways and the genes that were differentially expressed in the OC patient cell lines, we used Metacore™ software (Calrivate Analytics) to perform enrichment analysis (EA) using a widely known database for protein–protein signaling. EA identified the gene IDs of the potential targets from the DEG sets with gene IDs via the functional ontology function in MetaCore. The possibility of a random intersection of a gene set and the corresponding ontological entities was determined according to the hypergeometric intersection p-value. A reduced p-value suggested that the object was more relevant to the dataset, which indicated that it had a higher rating. Comparative enrichment analysis of the DEG dataset identified the top 10 enriched pathways, and the maps, GO cellular processes, networks, and biomarkers (by disease) are shown in **Figures 5A–D**. These are the most statistically significant data for each category based on a very low p-value. The pictorial representation of the top-scored pathway map (lowest p-value) is based on the distribution of gene enrichment, as shown in **Figure 6A**; similarly, the second scored map (second-lowest p-value) is shown in **Figure 6B**. In **Figures 6A,B**, the well-characterized proteins or protein complexes are displayed as individual symbols; the data from all experiments are shown and linked on the maps as thermometer-like symbols. A red upward-facing thermometer indicates the upregulated genes, and a blue thermometer indicates the expression level of a downregulated gene. The AN algorithm in GeneGo was used to identify the most relevant biological networks by prioritizing the number of fragments of the canonical pathways in the network, as shown in **Table 3**. The top regulated network processes are presented in **Supplementary Figures 1A,B**, illustrating the two major pathways involving DEGs that were commonly affected in both OC groups. We identified several crucial hub genes, including *TCF4*, frizzled family proteins (*FZD2*, *FZD8*, and *FZD6*), *RUNX2*, CDC25 family protein (*CDC25A*), protein kinase family proteins (*CDK2*), *BRCA1*, *ATM*, and *RBBP8*. The selected hub genes were mainly involved in the regulation of the canonical Wnt signaling pathway, cell–cell signaling mediated by Wnt, cell cycle phase transition, and the positive regulation of the cell cycle (**Figure 6**, **Supplementary Figures 1A,B**).

Survival Analysis and Expression Levels of Hub Genes

GEPIA survival assessment was used to investigate the overall association with survival of 10 hub genes from both the low- and high-expression OC groups. As a result, we noticed that the high expression of FZD2 (HR = 0.93) (**Figure 7C**), FZD8 (HR = 0.88) (**Figure 7D**), CDC25A (HR = 0.83) (**Figure 7F**), CDK2 (HR = 0.86) (**Figure 7G**), and RBBP8 (HR = 0.95) (**Figure 7J**) were associated with improved overall survival in the OC cell line. However, the high expression of TCF4 (HR = 1) (**Figure 7A**), FZD6 (HR = 1) (**Figure 7B**), RUNX2 (HR = 1) (**Figure 7E**), BRCA1 (HR = 1.1) (**Figure 7H**), and ATM (HR = 1.2) (**Figure 7I**) were linked with worse overall survival in the OC cell line. Taken together, the results show that *FZD6*, *FZD8*, *CDK2*, and *RBBP8* function as core genes that have a close relationship with OC. Furthermore, the GEPIA box plot analysis investigated the level of expression of the core genes in 426 OC tissue samples and 88 normal tissue samples. The boxplot in **Figures 8A–D** shows a considerable increase in the level of core

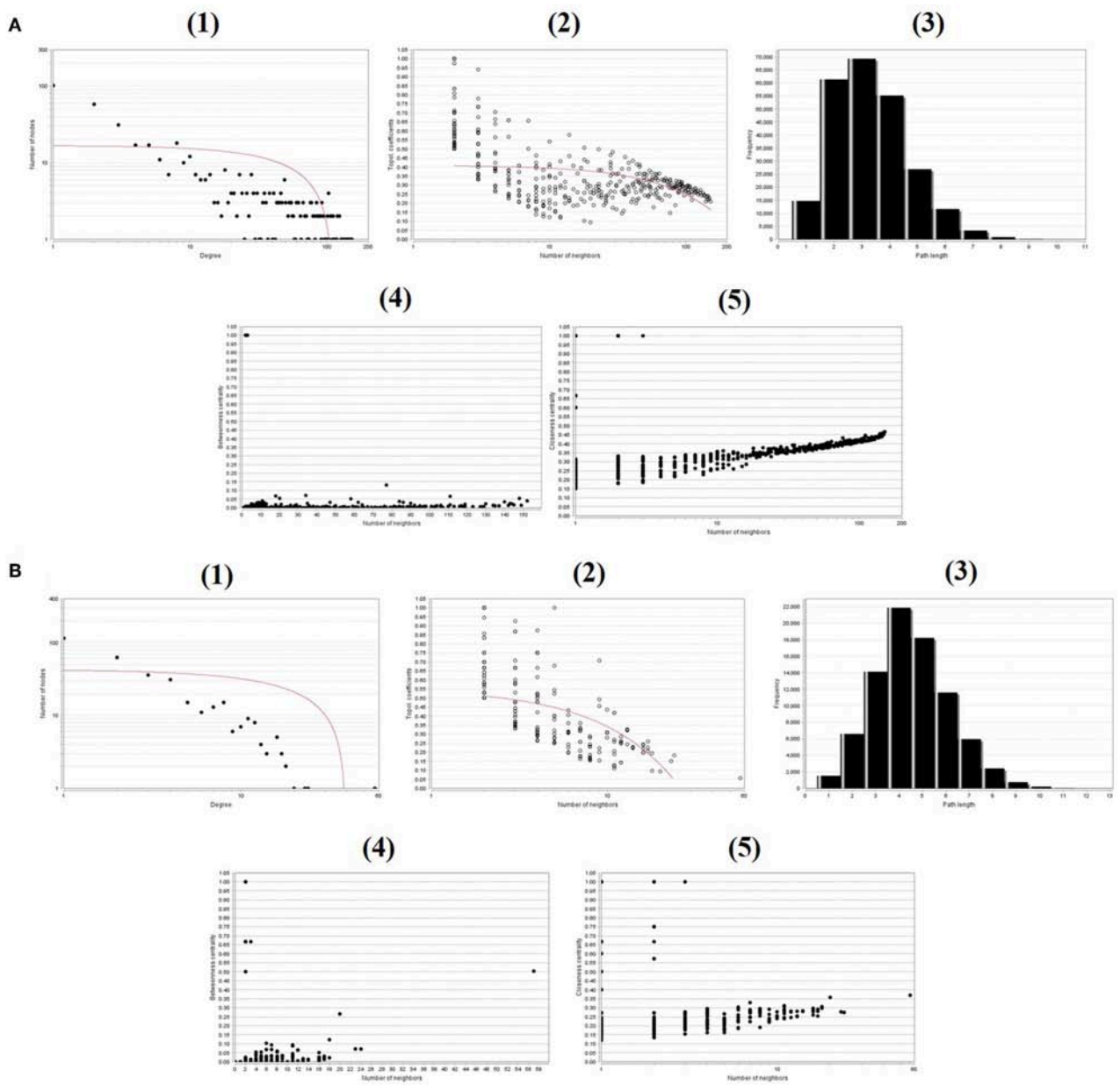


FIGURE 4 | Backbone network topology parameters. **(A)** The network of upregulated DEGs. **(B)** The network of downregulated DEGs. (1) Distribution of the node degree. (2) Topological coefficients. (3) Shortest path length distribution. (4) Betweenness centrality. (5) Closeness centrality.

gene expression (*FZD6*, *FZD8*, *CDK2*, and *RBBP8*) in the OC cell line.

DISCUSSION

Microarray technology is one of the most important approaches used by many researchers worldwide to explore the expression levels of genes involved in complex disorders (Russo et al., 2003; Babu, 2004; Perez-Diez et al., 2013). Therefore, studying the expression profiles of DEGs and predicting the target genes of OC is of the utmost importance. In this study, data from a total of three OC cell lines and three OC cell lines with SINE resistance were obtained from the GEO database (GSE126519). A total of 2708 DEGs were screened, including 809 upregulated and 700 downregulated genes. *In silico* methods have typically shown good efficiency, and networks have been demonstrated to be a reliable way to depict genomic data. The topological interpretation of upregulated and downregulated genes is required for large PPI networks and is thus substantially based on integrated local components, such as the distribution node of the degree, the topological coefficient, the shortest path length distribution, and the betweenness and closeness

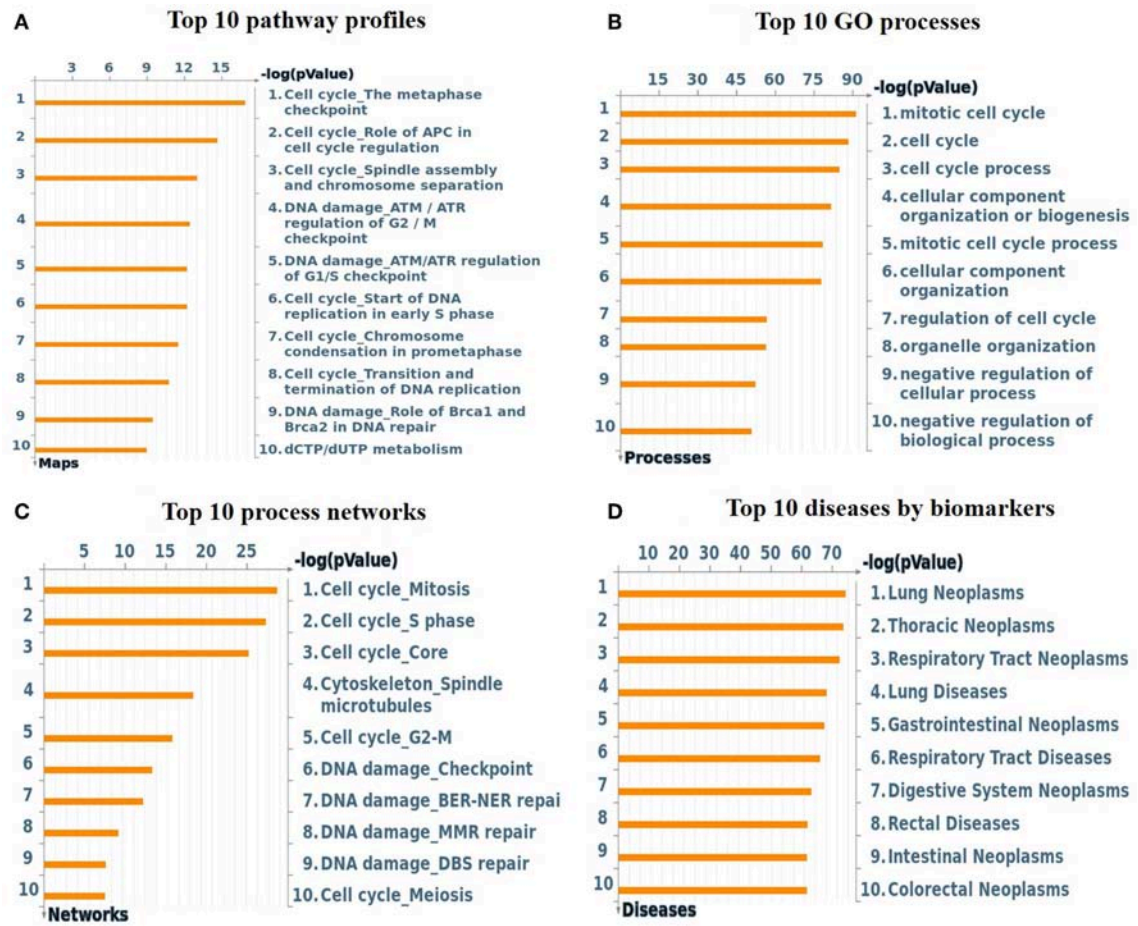


FIGURE 5 | (A) Top 10 pathway profiles; **(B)** top 10 GO processes; **(C)** top 10 process networks; **(D)** top 10 diseases according to biomarkers. GeneGo annotation of the top 10 pathway profiles, GO cellular processes, process networks, and diseases according to biomarkers for the DEG datasets. **(A)** The canonical pathway maps represent a set of signaling and metabolic maps comprehensively covering the relevant pathways in humans. **(B)** Gene Ontology (GO) cellular processes. As most GO processes have no gene/protein content, the "empty terms" were excluded from the *p*-value calculations. **(C)** The cellular and molecular processes were defined and annotated; each process represents a preset network of protein interactions characteristic of the process. **(D)** Disease folders (by biomarkers) were organized into a hierarchical tree.

centralities (Assenov et al., 2008). These parameters were used to analyze the nodes in individual PPI networks of the DEG dataset to deduce their significance in networks with different characteristics. Furthermore, we implemented GO and KEGG pathway analyses to determine MF, BP, CC, and pathways involving the DEGs using the DAVID online tool. The GO BP terms and KEGG assessment indicated that the upregulated DEGs were enriched primarily in the cell cycle, mitotic cell cycle process, mitotic nuclear division, DNA replication, cell cycle, nucleotide excision repair, DNA mismatch repair, and Fanconi anemia pathways. Interestingly, mutations in mismatch repair (MMR) and Fanconi anemia pathway-related genes in women have been shown to be one of the primary causes of hereditary OC (Norquist et al., 2016). Therefore, our observed results are consistent with the role of upregulated genes in pathways that cause OC. Similarly, the downregulated DEGs were mainly enriched in organ morphogenesis, cell development, cell morphogenesis, mesenchymal development and the interaction of the ECM receptor, PI3K-Akt signaling, and arginine and proline metabolism pathways. In line with this, a significant cause of cancer would appear to be the abnormal functioning of the cell cycle and mitosis (Kastan and Bartek, 2004; Malumbres and Barbacid, 2009).

The findings from the STRING, Cytoscape, GO, and KEGG analyses indicated that many pathways were primarily affected in OC. Several studies have used Cytoscape plugins such as MCODE, cytoHubba, CytoCluster, CytoKegg, and CytoNCA to elucidate the core interactions in PPI networks (Lan et al., 2015; Villaveces et al., 2015; Sriroopreddy and Sudandiradoss, 2018; Zhang et al., 2019). To delineate the molecular interactions and pathways identified from the STRING, GO, and KEGG analyses, we utilized GeneGo Metacore, which has a massive amount of information about regulatory and metabolic pathways and contains precisely curated biological networks. To obtain a

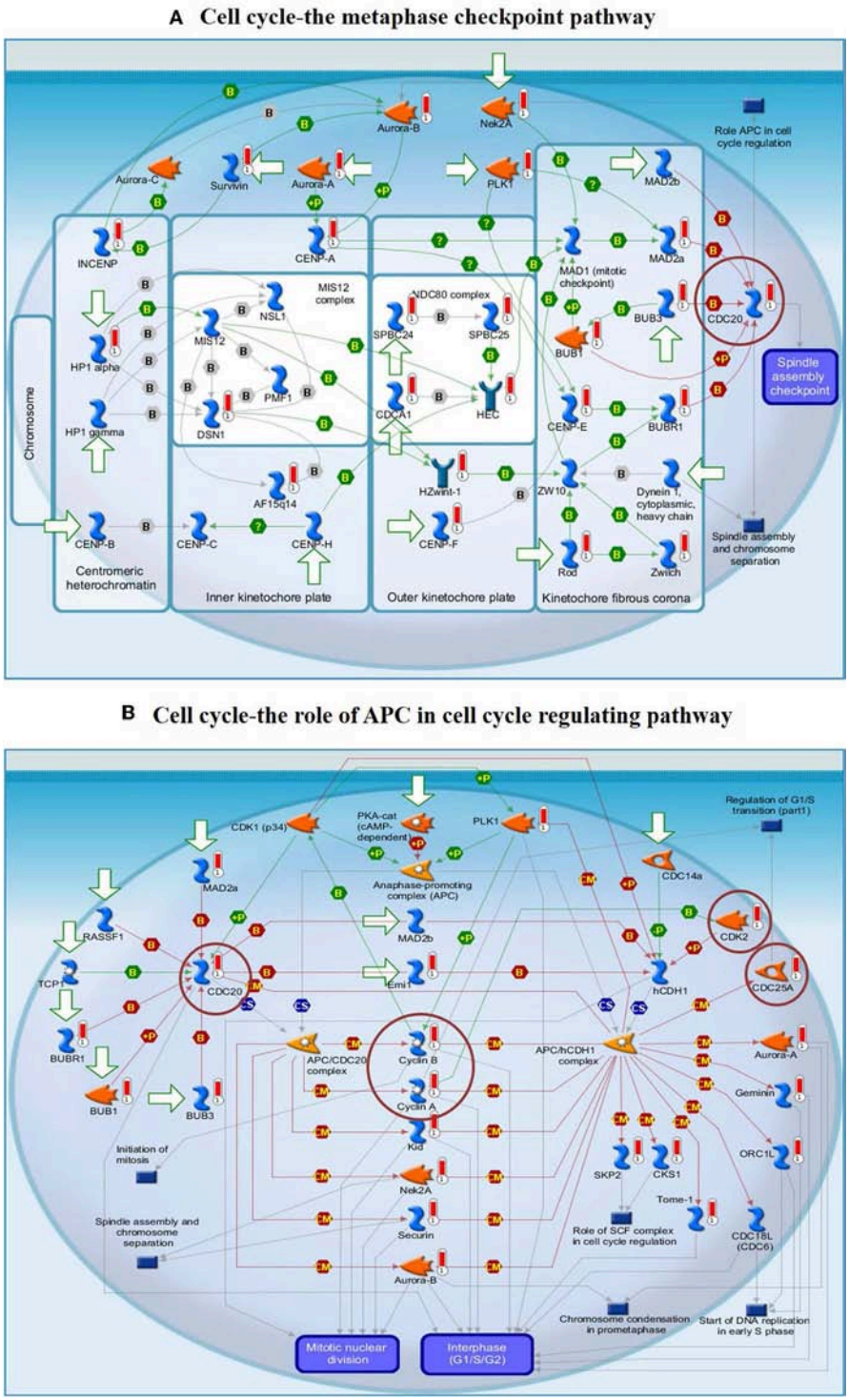


FIGURE 6 | (A) The cell cycle metaphase checkpoint pathway. **(B)** The APC cell cycle regulation pathway. The two top-scored regulated pathways were activated in the OC cell lines. The pathway images were generated by GeneGo Metacore™ enrichment analysis. Well-characterized proteins or protein complexes are shown as individual symbols within the image; experimental data from all the records are connected and depicted as thermometer-like figures on the maps. Upward-facing thermometers are shown in red and indicate upregulated gene transcripts. The linkage of proteins by arrows depicts the stimulatory and inhibitory effects or interaction of the encoded protein on the desired protein. The hub genes (protein families) that were involved in the canonical signaling pathways are marked in a circle (scarlet). Further explanations are provided at https://portal.genego.com/help/MC_legend.pdf.

TABLE 3 | Most relevant biological networks were generated using GeneGo Analyze Networks (AN) algorithm.

S. no.	Network name	Processes	Size	Target	Pathways	*p*-value	*z* score	*g* score
1	TCF7L2 (TCF4), Tcf(Lef), Frizzled, RUNX2, p21	Canonical Wnt signaling pathway (50.0%), regulation of Wnt signaling pathway (58.0%), cell-cell signaling by Wnt (56.0%), Wnt signaling pathway (56.0%), regulation of canonical Wnt signaling pathway (50.0%)	50	17	262	1.12e-20	21.47	348.97
2	CDC25A, CDK2, Brca1, ATM, RBBP8 (CtIP)	Mitotic cell cycle (57.1%), mitotic cell cycle process (53.1%), mitotic cell cycle phase transition (42.9%), cell cycle phase transition (42.9%), positive regulation of cell cycle (46.9%)	50	32	24	2.61e-49	41.10	71.10
3	KRMP1, FAM83D, LMO1, CBWD1, PSF1	cGMP catabolic process (6.4%), response to macrophage colony-stimulating factor (6.4%), cellular response to macrophage colony-stimulating factor stimulus (6.4%), purine ribonucleotide catabolic process (8.5%), ribonucleotide catabolic process (8.5%)	50	35	0	5.36e-56	45.03	45.03

In this workflow, the networks were prioritized based on the number of fragments of canonical pathways on the network.

comprehensive picture of the DEGs involved in OC, we used GeneGo Metacore™ software to identify the most significant genes and signaling pathways based on the calculated *p*-values. Among the top 10 enriched pathways, the cell cycle metaphase checkpoint, APC in cell cycle regulation, chromosome separation, spindle assembly, and DNA damage/ATR regulation of G2/M phase checkpoint/ATM were highly significant in the DEG datasets from both groups (**Figure 5A**). The GO cellular processes showed that the DEGs were enriched in a variety of cellular processes (**Figure 5B**), and these processes are mainly utilized in the enrichment analysis and to prioritize the genes in the constructed networks. The GO process networks were enriched in various groups. Among the top 10 process networks, we selected the four that were the most significant based on the calculated *p*-values, which included the mitotic cell cycle, the S phase in the cell cycle, and cytoskeleton-spindle microtubules (**Figure 5C**). The biomarkers of the diseases distinctly showed that the DEGs with the highest representation in the dataset were also known to contribute to other cancer types (**Figure 5D**). Additionally, there were two top-scored regulated pathways that were activated in the OC cell line that were involved in the cell cycle: the metaphase checkpoint and APC cell cycle regulation pathways (**Figure 6**). Furthermore, we analyzed the biological network of the upregulated DEGs in the signaling pathways by utilizing the AN algorithm in GeneGo. As a result, we determined the two most significant networks that were commonly affected in both of the OC groups. The components of these networks included several crucial hub genes, including *TCF4*, the frizzled family proteins (*FZD2*, *FZD8*, and *FZD6*), *RUNX2*, a CDC25 family protein (*CDC25A*), the protein kinase family proteins (*CDK2*), *BRCA1*, *ATM*, and *RBBP8*. Among them, the *TCF4, frizzled*, and *RUNX2* genes are primarily involved in the regulation of the canonical WNT signaling pathway (58%) and cell–cell signaling mediated by WNT (56%). Genes such as *CDC25A*, *CDK2*, *BRCA1*, *ATM*, and *RBBP8* are mostly involved in the mitotic cell cycle process (53.1%), mitotic cell cycle phase transition (42.9%), positive regulation of the cell cycle (46.9%), and cell cycle phase transition (42.9%). Finally, the GEPIA web server was used to assess the association between hub gene expression and OC prognosis. The overall survival analysis indicated that high expression of *FZD2*, *FZD8*, *CDC25A*, *CDK2*, and *RBBP8* were associated with better survival, and high expression of *TCF4*, *FZD6*, *RUNX2*, *BRCA1*, and *ATM* were associated with decreased survival in the OC cell line. Collectively, *FZD6*, *FZD8*, *CDK2*, and *RBBP8* were identified as core genes that were strongly associated with overall survival in OC. Therefore, these four genes could contribute to OC metastasis.

Supplementary Figure 1A shows that the frizzled family of proteins is involved in canonical Wnt signaling pathway regulation. FZD6, also known as frizzled class receptor 6, is a member of the "frizzled" gene family, which consists of 7-transmembrane domain proteins that are Wnt signaling protein receptors. Many studies have observed through mutagenesis experiments that several residues in the intracellular loops and the C-terminus of FZD play a prominent role in signaling (Cong et al., 2004; Wallingford and Habas, 2005). Kim et al. (2015) found that the expression of FZD6 was increased in colorectal cancer (CRC) patients when compared to that in nontumor tissues. Furthermore, they discovered that FZD6 expression in CRC cells was negatively regulated by miR199a-5p (Kim et al., 2015). In recent research, Corda et al. (2017) observed that the Wnt receptor-encoding gene *FZD6* is often duplicated in breast cancer and confers a higher risk of triple-negative breast cancer. For the assembly of the fibronectin matrix, FZD6 signaling is intrinsically required and interferes with actin cytoskeletal organization. The researchers concluded that in highly metastatic forms of breast cancer, such as TNBC, the FZD6-fibronectin actin axis could be targeted for drug treatment (Corda et al., 2017). In our study, we observed the overexpression of frizzled class receptor 6 in OC cell lines, and the overexpression of FZD6, which acts as an adverse prognostic factor, was associated with decreased survival in OC patients.

Frizzled class receptor 8 is also a "frizzled" gene family member that serves as a Wnt signaling protein receptor (Bhanot et al., 1996). Most frizzled receptors are also associated with the canonical signaling pathway of beta-catenin (Dann et al., 2001). Li et al. (2017) reported a higher level of FZD8 expression in

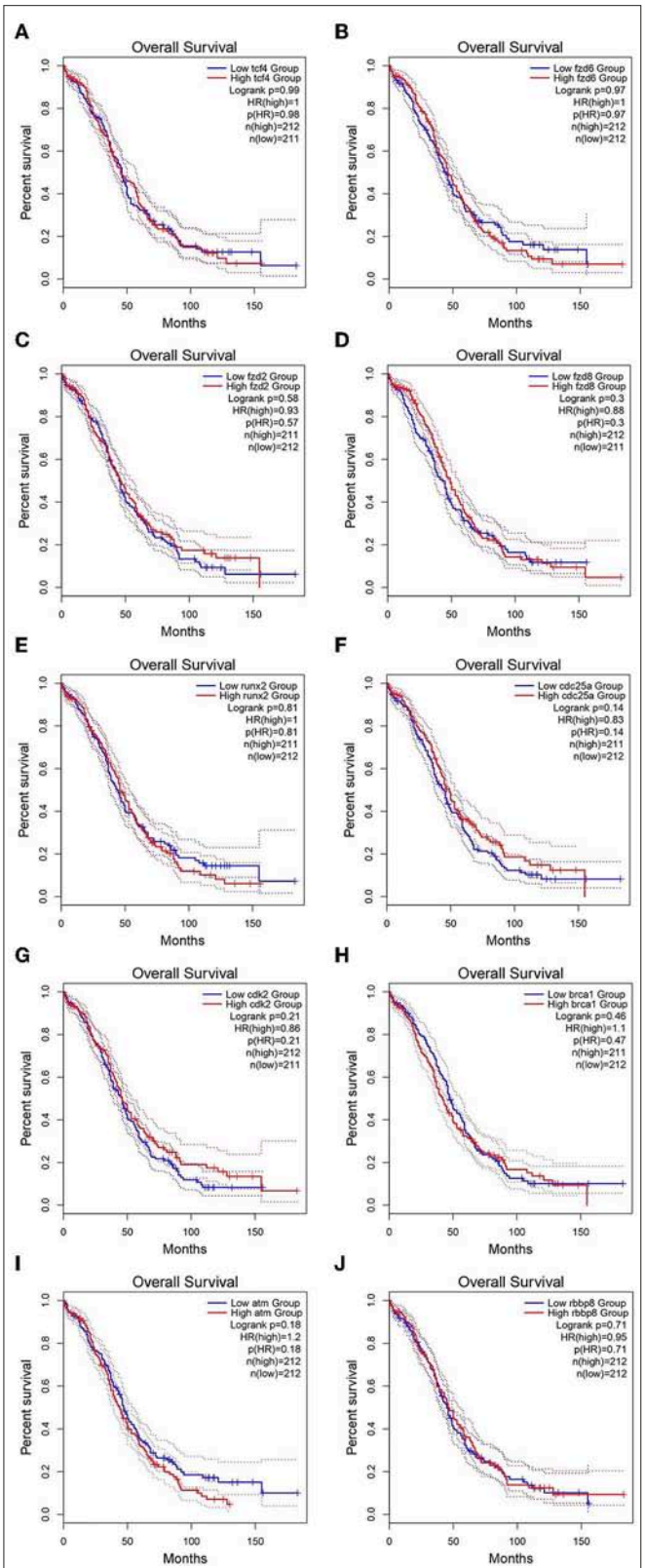


FIGURE 7 | Kaplan–Meier overall survival analysis of the hub genes expressed in OC and SINE-resistant OC cell lines. **(A)** TCF4 (Transcription factor 4), **(B)** *(Continued)*

FIGURE 7 | *FZD6* (Frizzled class receptor 6), **(C)** *FZD2* (Frizzled class receptor 2), **(D)** *FZD8* (Frizzled class receptor 8), **(E)** *RUNX2* (Runt related transcription factor 2), **(F)** *CDC25A* (Cell division cycle 25A), **(G)** *CDK2* (Cyclin-dependent kinase 2), **(H)** *BRCA1* (Breast cancer type 1 susceptibility protein), **(I)** *ATM* (ATM serine/threonine kinase), and **(J)** *RBBP8* (Retinoblastoma-binding protein 8). The survival curves were plotted using the GEPIA2 web server. The genes with high expression in the cohorts are shown in red, and the blue line indicates the low-expression cohort. The survival curves are represented as dotted lines, and the solid line represents the 95% confidence interval. HR stands for hazard ratio; patient number (n) = 212. The p-values were calculated using log-rank statistics.

bone metastases in prostate cancer (PCa), which is frequently diagnosed among men. This research group also found that the silencing of FZD8 suppressed the migration and invasion of cells and the occurrence of PCa bone metastasis *in vitro* and *in vivo* by activating the canonical β-catenin/Wnt signaling pathway, and the data suggest that FZD8 could be a potential therapeutic target for the treatment of bone metastasis in PCa (Li et al., 2017). Chakravarthi et al. (2018) reported that ETS-related gene (ERS) specifically targets and activates FZD8 directly by binding to its promoter region rather than ETV1 and suggested that the overexpression of ERG in PCa leads to FZD8 induction and the activation of the Wnt pathway (Chakravarthi et al., 2018). The research group led by He et al. recently found that miR-520b overexpression results in the inhibition of cell proliferation, migration, and invasion in human spinal osteosarcoma (OS) tissues and cell lines by inactivating the Wnt/β-catenin signaling pathway through the downregulation of FZD8 and thus provides a new spinal OS therapeutic target (Wang et al., 2017). Similarly, Liu et al. (2019) reported a reduced level of miR-99b-5p in non-small cell lung cancer (NSCLC) cell lines. They validated FZD8 as a specific target of miR-99b-5p and found that increased expression of miR-99b-5p inhibited NSCLC proliferation, migration, and invasion *in vitro* (Liu et al., 2019). The findings of our study suggest that the overexpression of FZDs in OC results in the anomalous activation of the canonical Wnt signaling pathway and may increase their function during the development of OC.

As seen in **Figure 6B**, CDK2 is mainly involved in the APC cell cycle regulation pathway, and the overexpression of CDK2 results in the upregulation of the G1/S phase transition, resulting in cancer cell proliferation. CDK2 and other relevant genes that are upregulated in OC (red circles in **Supplementary Figure 1B**) are shown in the closed network of the APC cell cycle regulation pathway; because of the increase in CDKs, APC failed to inactivate the CDK complexes by inducing their degradation. Liu et al. (2011) revealed that CDK2 expression was significantly higher in laryngeal squamous cell cancer tissues when compared to that in paired adjacent normal laryngeal tissues (Liu et al., 2011). Duong et al. (2012) reported that low-molecular-weight cyclin E (LMW-E) required kinase activity associated with CDK2 to induce the formation of mammary tumors by disrupting the growth of acinar cells. They used a combination of therapy with a CDK inhibitor (roscovitine) plus a b-Raf-targeting pan-kinase inhibitor (sorafenib) or an mTOR inhibitor (rapamycin) to arrest

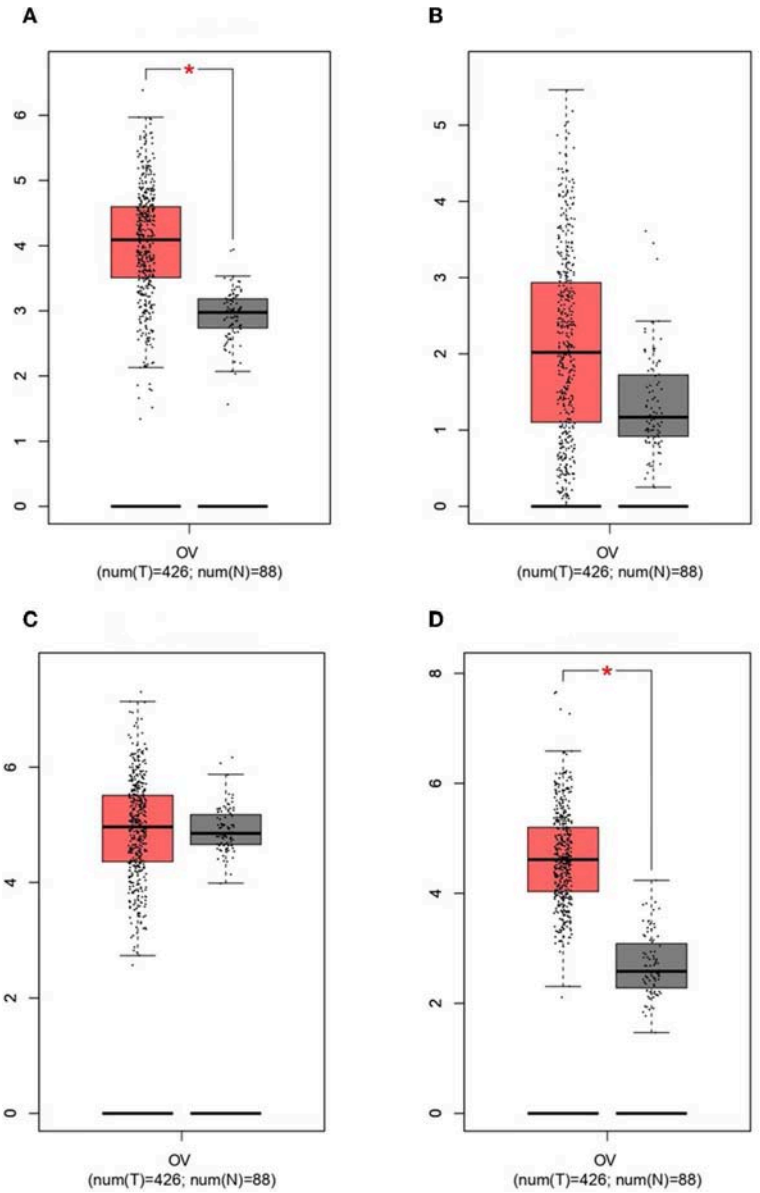


FIGURE 8 | Based on TCGA and GTEx data in GEPIA, we validated the expression levels of the four core genes in ovarian cancer ($n = 426$) and normal tissues ($n = 88$). **(A)** *FZD6*. **(B)** *FZD8*. **(C)** *CDK2*. **(D)** *RBBP8*.

the G1/S cell cycle in breast cancer cells; thus, the b-Raf-ERK1/2-mTOR signaling pathway could be suppressed (Duong et al., 2012). Kanwal et al. (2016) found that the expression of CDK2 is significantly increased in OC tissues when compared to that in normal ovarian tissues (Kanwal et al., 2016). The pathways involving cyclin-dependent kinase (CDK) are significant and well-established cancer treatment targets. The role of CDK2 remains controversial in several cancer types (McCurdy et al., 2017). Many studies have suggested that CDK2 could be a crucial factor in the progression of cancer by regulating several pathways and might be a prospective biomarker and indicator of prognosis. Therefore, CDK2 and its cyclin binding partners are possible therapeutic targets for future cancer treatments (Yin et al., 2018; Zhang et al., 2018; Wood et al., 2019).

RBBP8, also known as retinoblastoma-binding protein 8, aids in regulating cell proliferation and DNA repair by homologous recombination (Fusco et al., 1998). Soria-Bretones et al. (2013) observed decreased or no expression of *RBBP8* in paraffin-embedded breast cancer biopsy tissues from high-grade breast cancer and nodal metastases that were acquired during tumor removal surgery (Soria-Bretones et al., 2013). The research group led by Rose et al. suggested that *RBBP8* was significantly hypermethylated in bladder cancer (BLCA) and was associated with more prolonged overall survival, and they indicated that

it may be used as a complementary marker for the detection of BLCA in urine (Mijnes et al., 2018). Miao et al. (2019) reported that the downregulation of the long noncoding RNA (lncRNA) cancer susceptibility candidate 2 (CASC2, enhanced tumor development, increased miR-18a-5p levels, and reduced the expression of *RBBP8* in nasopharyngeal carcinoma (NPC). The upregulation of CASC2 resulted in decreased proliferation and increased apoptotic cell death *in vivo* (Miao et al., 2019). Our data clearly showed that *RBBP8* was differentially expressed and involved in cell cycle regulation (**Supplementary Figure 1B**). Additionally, it contributes to the development of OC in both groups. However, the role of *RBBP8* in OC is unclear and requires further research.

Furthermore, we conducted an interrelation analysis of the identified hub genes to elucidate the interactions between them, primarily among the genes that interacted with one another directly or indirectly. As shown in **Figure 9**, the mitotic cell cycle phase transition pathway interacts with the regulation of G1/S phase transition and APC in the cell cycle regulating pathway via the essential genes *RBBP8*, *BRCA1*, *CDC25A*, *ATM*, and *CDK2* (D'Andrilli et al., 2004; Soria-Bretones et al., 2013; Xiao et al., 2019). In contrast, frizzled family proteins (*FZD2, FZD6,* and *FZD8*) are directly involved in the Wnt signaling pathways because they are receptors of Wnt proteins (Janda et al., 2012). The *TCF4* and *RUNX2* genes are involved directly or indirectly in the Wnt signaling network, resulting in tumorigenesis (Gaur et al., 2005; Hrckulak et al., 2018; Komori, 2019). Taken together, these findings showed that node genes involved in the development of OC could be significant factors in cell cycle regulation and the Wnt signaling pathway.

Overall, our systematic bioinformatics assessment demonstrated that DEGs might play a pivotal role in the incidence, prognosis, growth, and development of OC. In this study, a total of 2708 DEGs and 10 hub genes were identified, and *FZD6*, *FZD8*, *CDK2*, and *RBBP8* could be the core genes involved in OC and SINE-resistant OC. Expression analysis and the correlation of the multiple genes will undoubtedly aid in the understanding of the roles of such genes in the growth and development of OC. Several research groups have demonstrated that preclinical models have showed some success in reducing tumor growth and decreasing the side effects of existing chemotherapy drugs (Cicenas et al., 2015; Whittaker

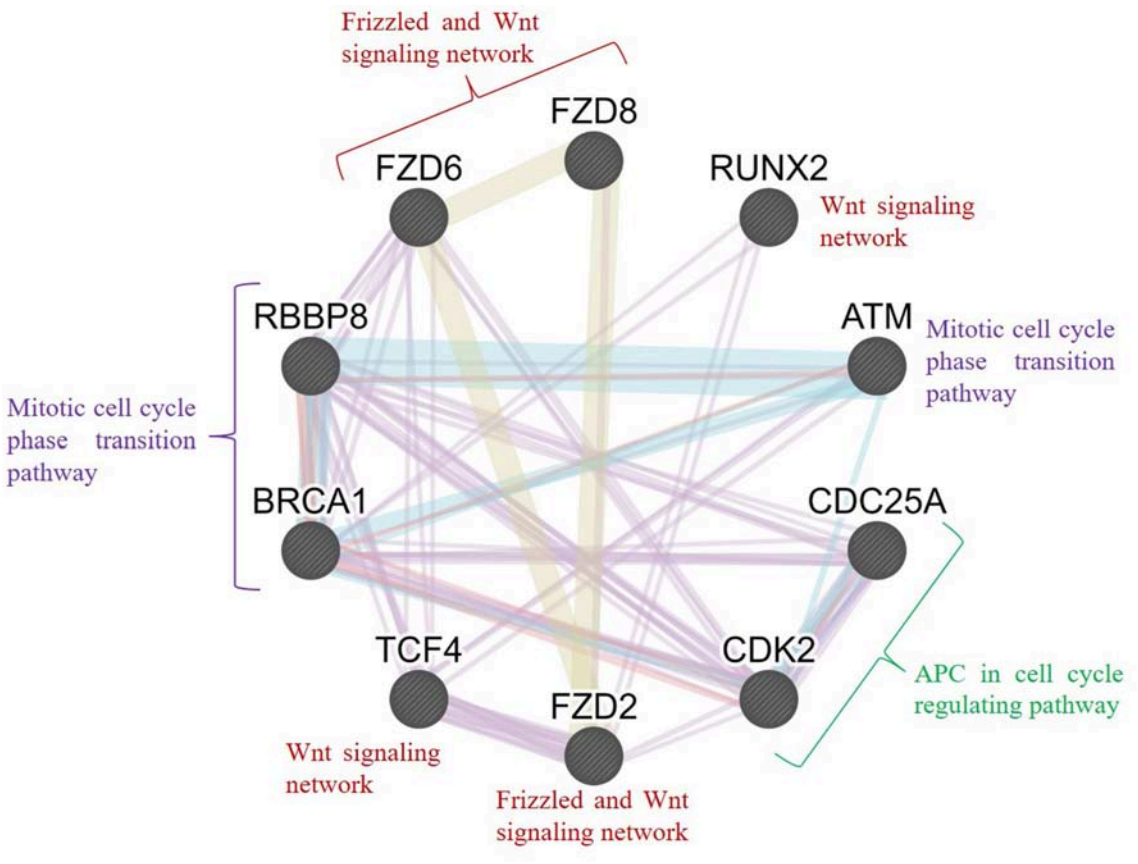


FIGURE 9 | Interrelation analysis of the hub genes identified from different pathways. GeneMANIA was used to plot the network, which was visualized in Cytoscape. Color code: physical interaction shown in red, coexpression shown in violet, predicted interaction shown in orange, common pathway shown in cyan, and colocalization shown in blue. The genes *FZD2*, *FZD6*, *FZD8*, *RUNX2*, and *TCF4* were involved in the frizzled and Wnt signaling network; *BRCA1*, *RBBP8*, and *ATM* were involved in the mitotic cell cycle phase transition pathway; and *CDC25A* and *CDK2* were involved in the APC cell cycle regulation pathway.

et al., 2017; Xia et al., 2018). We need to conduct a series of experimental studies to prove this hypothesis to obtain more precise correlation reports. However, the findings from this study could enhance the understanding of the molecular pathogenesis of OC. Furthermore, the core genes and pathways might be potential biomarkers that could be used for the detection and targeting of OC and SINE-resistant OC cells for therapy.

DATA AVAILABILITY STATEMENT

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

AUTHOR CONTRIBUTIONS

SK, DK, and CD were involved in the design of the study and the acquisition, analysis, and interpretation of the data. SK, DK, CD, and HZ were involved in the interpretation of the data and drafting the manuscript. CD, RS, and HZ supervised the entire study and were involved in study design, the acquisition, analysis, interpretation of the data, and drafting the manuscript. The manuscript was reviewed and approved by all the authors.

FUNDING

This work was supported by Qatar University Grant# QUST-2-CHS-2019-3.

ACKNOWLEDGMENTS

The authors would like to take this opportunity to thank the management of Vellore Institute of Technology for providing the necessary facilities and encouragement to carry out this work.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fbioe.2019.00391/full#supplementary-material

REFERENCES

Assenov, Y., Ramírez, F., Schelhorn, S.-E., Lengauer, T., and Albrecht, M. (2008). Computing topological parameters of biological networks. *Bioinforma. Oxf. Engl.* 24, 282–284. doi: 10.1093/bioinformatics/btm554

Aubert, J., Bar-Hen, A., Daudin, J. J., and Robin, S. (2004). Determination of the differentially expressed genes in microarray experiments using local FDR. *BMC Bioinformatics* 5:125. doi: 10.1186/1471-2105-5-125

Babu, M. M. (2004). An introduction to microarray data analysis. *Microarray Data Anal.* 225–249.

Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., et al. (2013). NCBI GEO: archive for functional genomics data sets–update. *Nucleic Acids Res.* 41, D991–D995. doi: 10.1093/nar/gks1193

Bhanot, P., Brink, M., Samos, C. H., Hsieh, J. C., Wang, Y., Macke, J. P., et al. (1996). A new member of the frizzled family from Drosophila functions as a Wingless receptor. *Nature* 382, 225–230. doi: 10.1038/382225a0

Brandes, U. (2001). A faster algorithm for betweenness centrality. *J. Math. Sociol.* 25, 163–177. doi: 10.1080/0022250X.2001.9990249

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

Chakravarthi, B. V. S. K., Chandrashekar, D. S., Hodigere Balasubramanya, S. A., Robinson, A. D., Carskadon, S., Rao, U., et al. (2018). Wnt receptor Frizzled 8 is a target of ERG in prostate cancer. *Prostate* 78, 1311–1320. doi: 10.1002/pros.23704

Cicenas, J., Kalyan, K., Sorokinas, A., Stankunas, E., Levy, J., Meskinyte, I., et al. (2015). Roscovitine in cancer and other diseases. *Ann. Transl. Med.* 3:135. doi: 10.3978/j.issn.2305-5839.2015.03.61

Cong, F., Schweizer, L., and Varmus, H. (2004). Wnt signals across the plasma membrane to activate the beta-catenin pathway by forming oligomers containing its receptors, Frizzled and LRP. *Dev. Camb. Engl.* 131, 5103–5115. doi: 10.1242/dev.01318

Corda, G., Sala, G., Lattanzio, R., Iezzi, M., Sallese, M., Fragassi, G., et al. (2017). Functional and prognostic significance of the genomic amplification of frizzled 6 (FZD6) in breast cancer. *J. Pathol.* 241, 350–361. doi: 10.1002/path.4841

D'Andrilli, G., Kumar, C., Scambia, G., and Giordano, A. (2004). Cell cycle genes in ovarian cancer: steps toward earlier diagnosis and novel therapies. *Clin. Cancer Res.* 10, 8132–8141. doi: 10.1158/1078-0432.CCR-04-0886

Dann, C. E., Hsieh, J. C., Rattner, A., Sharma, D., Nathans, J., and Leahy, D. J. (2001). Insights into Wnt binding and signalling from the structures of two Frizzled cysteine-rich domains. *Nature* 412, 86–90. doi: 10.1038/35083601

Duong, M. T., Akli, S., Wei, C., Wingate, H. F., Liu, W., Lu, Y., et al. (2012). LMW-E/CDK2 deregulates acinar morphogenesis, induces tumorigenesis, and associates with the activated b-Raf-ERK1/2-mTOR pathway in breast cancer patients. *PLoS Genet.* 8:e1002538. doi: 10.1371/journal.pgen.1002538

Franz, M., Rodriguez, H., Lopes, C., Zuberi, K., Montojo, J., Bader, G. D., et al. (2018). GeneMANIA update 2018. *Nucleic Acids Res.* 46, W60–W64. doi: 10.1093/nar/gky311

Fusco, C., Reymond, A., and Zervos, A. S. (1998). Molecular cloning and characterization of a novel retinoblastoma-binding protein. *Genomics* 51, 351–358. doi: 10.1006/geno.1998.5368

Gaur, T., Lengner, C. J., Hovhannisyan, H., Bhat, R. A., Bodine, P. V. N., Komm, B. S., et al. (2005). Canonical WNT signaling promotes osteogenesis by directly stimulating Runx2 gene expression. *J. Biol. Chem.* 280, 33132–33140. doi: 10.1074/jbc.M500608200

Gerecitano, J. (2014). SINE (selective inhibitor of nuclear export)–translational science in a new class of anti-cancer agents. *J. Hematol. Oncol.J Hematol Oncol* 7:67. doi: 10.1186/s13045-014-0067-3

Hrckulak, D., Janeckova, L., Lanikova, L., Kriz, V., Horazna, M., Babosova, O., et al. (2018). Wnt effector TCF4 is dispensable for Wnt signaling in human cancer cells. *Genes* 9:439. doi: 10.3390/genes9090439

Huang, D. W., Sherman, B. T., and Lempicki, R. A. (2009a). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.* 37, 1–13. doi: 10.1093/nar/gkn923

Huang, D. W., Sherman, B. T., and Lempicki, R. A. (2009b). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* 4, 44–57. doi: 10.1038/nprot.2008.211

Janda, C. Y., Waghray, D., Levin, A. M., Thomas, C., and Garcia, K. C. (2012). Structural basis of Wnt recognition by frizzled. *Science* 337, 59–64. doi: 10.1126/science.1222879

Kanwal, A., Kaur, M., Singh, A., Gupta, S., and Sachan, M. (2016). Hypo/unmethylated promoter status of Cdk2 gene correlates with its over-expression in ovarian cancer in north Indian population. *Cell. Mol. Biol. Noisy Gd. Fr.* 62, 67–72. doi: 10.14715/cmb/2016.62.1.13

Kastan, M. B., and Bartek, J. (2004). Cell-cycle checkpoints and cancer. *Nature* 432, 316–323. doi: 10.1038/nature03097

Kim, B.-K., Yoo, H.-I., Kim, I., Park, J., and Kim Yoon, S. (2015). FZD6 expression is negatively regulated by miR-199a-5p in human colorectal cancer. *BMB Rep.* 48, 360–366. doi: 10.5483/BMBRep.2015.48.6.031

Komori, T. (2019). Regulation of proliferation, differentiation and functions of osteoblasts by Runx2. *Int. J. Mol. Sci.* 20:1694. doi: 10.3390/ijms20071694

Lan, W., Wang, J., Li, M., Peng, W., and Wu, F. (2015). Computational approaches for prioritizing candidate disease genes based on PPI networks. *Tsinghua Sci. Technol.* 20, 500–512. doi: 10.1109/TST.2015.7297749

Li, Q., Ye, L., Zhang, X., Wang, M., Lin, C., Huang, S., et al. (2017). FZD8, a target of p53, promotes bone metastasis in prostate cancer by activating canonical Wnt/β-catenin signaling. *Cancer Lett.* 402, 166–176. doi: 10.1016/j.canlet.2017.05.029

Liu, J.-L., Ma, H.-P., Lu, X.-L., Sun, S.-H., Guo, X., and Li, F.-C. (2011). NF-κB induces abnormal centrosome amplification by upregulation of CDK2 in laryngeal squamous cell cancer. *Int. J. Oncol.* 39, 915–924. doi: 10.3892/ijo.2011.1125

Liu, R., Chen, Y., Shou, T., Hu, J., and Qing, C. (2019). miRNA-99b-5p targets FZD8 to inhibit non-small cell lung cancer proliferation, migration and invasion. *OncoTargets Ther.* 12, 2615–2621. doi: 10.2147/OTT.S199196

Malumbres, M., and Barbacid, M. (2009). Cell cycle, CDKs and cancer: a changing paradigm. *Nat. Rev. Cancer* 9, 153–166. doi: 10.1038/nrc2602

Matulonis, U. A., Sood, A. K., Fallowfield, L., Howitt, B. E., Sehouli, J., and Karlan, B. Y. (2016). Ovarian cancer. *Nat. Rev. Dis. Primer* 2:16061. doi: 10.1038/nrdp.2016.61

McCurdy, S. R., Pacal, M., Ahmad, M., and Bremner, R. (2017). A CDK2 activity signature predicts outcome in CDK2-low cancers. *Oncogene* 36, 2491–2502. doi: 10.1038/onc.2016.409

Mendonca, J., Sharma, A., Kim, H.-S., Hammers, H., Meeker, A., Marzo, A. D., et al. (2014). Selective inhibitors of nuclear export (SINE) as novel therapeutics for prostate cancer. *Oncotarget* 5, 6102–6112. doi: 10.18632/oncotarget.2174

Miao, W.-J., Yuan, D.-J., Zhang, G.-Z., Liu, Q., Ma, H.-M., and Jin, Q.-Q. (2019). lncRNA CASC2/miR-18a-5p axis regulates the malignant potential of nasopharyngeal carcinoma by targeting RBBP8. *Oncol. Rep.* 41, 1797–1806. doi: 10.3892/or.2018.6941

Mijnes, J., Veeck, J., Gaisa, N. T., Burghardt, E., de Ruijter, T. C., Gostek, S., et al. (2018). Promoter methylation of DNA damage repair (DDR) genes in human tumor entities: RBBP8/CtIP is almost exclusively methylated in bladder cancer. *Clin. Epigenetics* 10, 15. doi: 10.1186/s13148-018-0447-6

Miyake, T., and Sood, S. (2019). *Analysis of RNA Profiles in Parent and Selective Inhibitors of Nuclear Export (SINE) Resistant Ovarian Cancer Cells*. GEO Accession Viewer. Available online at: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE126519 (accessed April 29, 2019).

Newman, M. E. J. (2005). A measure of betweenness centrality based on random walks. *Soc. Netw.* 27, 39–54. doi: 10.1016/j.socnet.2004.11.009

Norquist, B. M., Harrell, M. I., Brady, M. F., Walsh, T., Lee, M. K., Gulsuner, S., et al. (2016). Inherited mutations in women with ovarian carcinoma. *JAMA Oncol.* 2, 482–490. doi: 10.1001/jamaoncol.2015.5495

Pennington, K. P., and Swisher, E. M. (2012). Hereditary ovarian cancer: beyond the usual suspects. *Gynecol. Oncol.* 124, 347–353. doi: 10.1016/j.ygyno.2011.12.415

Perez-Diez, A., Morgun, A., and Shulzhenko, N. (2013). "Microarrays for cancer diagnosis and classification," in *Landes Bioscience*. Available online at: https://www.ncbi.nlm.nih.gov/books/NBK6624/ (accessed July 19, 2019).

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

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

Russo, G., Zegar, C., and Giordano, A. (2003). Advantages and limitations of microarray technology in human cancer. *Oncogene* 22:6497. doi: 10.1038/sj.onc.1206865

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

Senapedis, W. T., Baloglu, E., and Landesman, Y. (2014). Clinical translation of nuclear export inhibitors in cancer. *Semin. Cancer Biol.* 27, 74–86. doi: 10.1016/j.semcancer.2014.04.005

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

Soria-Bretones, I., Sáez, C., Ruíz-Borrego, M., Japón, M. A., and Huertas, P. (2013). Prognostic value of CtIP/RBBP8 expression in breast cancer. *Cancer Med.* 2, 774–783. doi: 10.1002/cam4.141

Sriroopreddy, R., and Sudandiradoss, C. (2018). Integrative network-based approach identifies central genetic and transcriptomic elements in triple-negative breast cancer. *Funct. Integr. Genomics* 18, 113–124. doi: 10.1007/s10142-017-0579-3

Stelzl, U., Worm, U., Lalowski, M., Haenig, C., Brembeck, F. H., Goehler, H., et al. (2005). A human protein-protein interaction network: a resource for annotating the proteome. *Cell* 122, 957–968. doi: 10.1016/j.cell.2005.08.029

Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., et al. (2015). STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.* 43, D447–452. doi: 10.1093/nar/gku1003

Tang, Z., Li, C., Kang, B., Gao, G., Li, C., and Zhang, Z. (2017). GEPIA: a web server for cancer and normal gene expression profiling and interactive analyses. *Nucleic Acids Res.* 45, W98–W102. doi: 10.1093/nar/gkx247

Villaveces, J. M., Koti, P., and Habermann, B. H. (2015). Tools for visualization and analysis of molecular networks, pathways, and -omics data. *Adv. Appl. Bioinforma. Chem. AABC* 8, 11–22. doi: 10.2147/AABC.S63534

Wallingford, J. B., and Habas, R. (2005). The developmental biology of Dishevelled: an enigmatic protein governing cell fate and cell polarity. *Dev. Camb. Engl.* 132, 4421–4436. doi: 10.1242/dev.02068

Wang, J., Pang, W., Zuo, Z., Zhang, W., and He, W. (2017). MicroRNA-520b suppresses proliferation, migration, and invasion of spinal osteosarcoma cells via downregulation of frizzled-8. *Oncol. Res.* 25, 1297–1304. doi: 10.3727/096504017X14873430389189

Whittaker, S. R., Mallinger, A., Workman, P., and Clarke, P. A. (2017). Inhibitors of cyclin-dependent kinases as cancer therapeutics. *Pharmacol. Ther.* 173, 83–105. doi: 10.1016/j.pharmthera.2017.02.008

Wood, D. J., Korolchuk, S., Tatum, N. J., Wang, L.-Z., Endicott, J. A., Noble, M. E. M., et al. (2019). Differences in the conformational energy landscape of CDK1 and CDK2 suggest a mechanism for achieving selective CDK inhibition. *Cell Chem. Biol.* 26, 121–130.e5. doi: 10.1016/j.chembiol.2018.10.015

Xia, P., Liu, Y., Chen, J., Coates, S., Liu, D. X., and Cheng, Z. (2018). Inhibition of cyclin-dependent kinase 2 protects against doxorubicin-induced cardiomyocyte apoptosis and cardiomyopathy. *J. Biol. Chem.* 293, 19672–19685. doi: 10.1074/jbc.RA118.004673

Xiao, Y., Yu, Y., Gao, D., Jin, W., Jiang, P., Li, Y., et al. (2019). Inhibition of CDC25B With WG-391D impedes the tumorigenesis of ovarian cancer. *Front. Oncol.* 9:236. doi: 10.3389/fonc.2019.00236

Yin, X., Yu, J., Zhou, Y., Wang, C., Jiao, Z., Qian, Z., et al. (2018). Identification of CDK2 as a novel target in treatment of prostate cancer. *Future Oncol.* 14, 709–718. doi: 10.2217/fon-2017-0561

Younes, N., and Zayed, H. (2019). Genetic epidemiology of ovarian cancer in the 22 Arab countries: a systematic review. *Gene* 684, 154–164. doi: 10.1016/j.gene.2018.10.044

Zhang, R., Yang, X., Wang, J., Han, L., Yang, A., Zhang, J., et al. (2019). Identification of potential biomarkers for differential diagnosis between rheumatoid arthritis and osteoarthritis via integrative genome-wide gene expression profiling analysis. *Mol. Med. Rep.* 19, 30–40. doi: 10.3892/mmr.2018.9677

Zhang, X., Zhao, Y., Wang, C., Ju, H., Liu, W., Zhang, X., et al. (2018). Rhomboid domain-containing protein 1 promotes breast cancer progression by regulating the p-Akt and CDK2 levels. *Cell Commun. Signal. CCS* 16:65. doi: 10.1186/s12964-018-0267-5

Zyl, B., van, Tang, D., and Bowden, N. A. (2018). Biomarkers of platinum resistance in ovarian cancer: what can we use to improve treatment. *Endocr. Relat. Cancer* 25, R303–R318. doi: 10.1530/ERC-17-0336

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: 07 February 2020
doi: 10.3389/fgene.2019.01370

Identification of Biological Pathways Contributing to Marbling in Skeletal Muscle to Improve Beef Cattle Breeding

Zahra Roudbari[1,2*], *Susan L. Coort*[2], *Martina Kutmon*[2,3], *Lars Eijssen*[2], *Jonathan Melius*[2], *Tomasz Sadkowski*[4] *and Chris T. Evelo*[2,3]

[1] Department of Animal Science, Faculty of Agriculture, University of Jiroft, Jiroft, Iran, [2] Department of Bioinformatics-BiGCaT, NUTRIM School of Nutrition and Translational Research in Metabolism, Maastricht University, Maastricht, Netherlands, [3] Maastricht Centre for Systems Biology (MaCSBio), Maastricht University, Maastricht, Netherlands, [4] Department of Physiological Sciences, Faculty of Veterinary Medicine, Warsaw University of Life Sciences - SGGW, Warsaw, Poland

OPEN ACCESS

Edited by:
Jyoti Sharma, Institute of Bioinformatics (IOB), India

Reviewed by:
Duo Xu, Cornell University, United States
Abhishek Kumar, University of Kiel, Germany
Bin Tong, Inner Mongolia University, China

***Correspondence:**
Zahra Roudbari
roudbari.zahra@ujiroft.ac.ir

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

Received: *04 September 2019*
Accepted: *16 December 2019*
Published: *07 February 2020*

Citation:
Roudbari Z, Coort SL, Kutmon M, Eijssen L, Melius J, Sadkowski T and Evelo CT (2020) Identification of Biological Pathways Contributing to Marbling in Skeletal Muscle to Improve Beef Cattle Breeding. Front. Genet. 10:1370. doi: 10.3389/fgene.2019.01370

Red meat is an important dietary source that provides part of the nutritional requirements. Intramuscular fat, known as marbling, is located throughout skeletal muscle. Marbling is a trait of major economic relevance that positively influences sensory quality aspects. The aim of the present study was to identify and better understand biological pathways defining marbling in beef cattle. Pathway analysis was performed in PathVisio with publicly available transcriptomic data from semitendinosus muscle of well-marbled and lean-marbled beef. Moreover, for *Bos taurus* we created a gene identifier mapping database with bridgeDb and a pathway collection in WikiPathways. The regulation of marbling is possibly the result of the interplay between signaling pathways in muscle, fat, and intramuscular connective tissue. Pathway analysis revealed 17 pathways that were significantly different between well-marbled and lean-marbled beef. The MAPK signaling pathway was enriched, and the signaling pathways that play a role in tissue development were also affected. Interestingly, pathways related to immune response and insulin signaling were enriched.

Keywords: marbling, curation pathway for cow, signaling pathway, improve breeding selection, transcriptomics profiling

INTRODUCTION

Red meat is as an important dietary source that provides part of the nutritional requirements such as proteins, minerals, B-complex vitamins, and essential fatty acids (McAfee et al., 2010). Control of meat quality is very important for meat producers and meat sellers to satisfy customer's preferences (Bernard et al., 2007). Marbling, a trait that describes the presence intramuscular fat, is of major economic relevance for beef producing cattle that has a positive impact on sensory quality traits, such as flavor, juiciness, and tenderness of meat. Studies have shown that marbling depends on factors such as breed, genotype, age, diet, husbandry, and growth stages. Although in marbling the environmental factors play an important role, the genetic background of the animals is the major factor defining the marbling status (Yamada et al., 2006). O'Connor et al. studied the effect of breed-type on marbling, their results demonstrated that increase in meat marbling from *Bos taurus* cattle (Hereford, Red Angus, Angus, and Tarentaise breeds) can increase the tenderness, more than the Bos

indicus cattle (Braford, Red Brangus, and Simbrah breeds) (O'Connor et al., 1997). Although, Shackelford et al., found that, the lower tenderness of meat from Bos indicus cattle is mainly because of decreased postmortem proteolysis which result from elevated calpastatin activity that the possibility existed for an interaction between breed and the influences of marbling score on tenderness (Shackelford et al., 1991). Also, Wulf et al., reported that, in Charolais and Limousin breeds, marbling correlated with calpastatin activity and shear force. He suggested that selection for increased marbling based on these genetic effects, in those two breeds, might be effective for enhancing beef tenderness (Wulf et al., 1996). There is substantial evidence from transcriptomics studies that gene expression profiles affect phenotypic variation for marbling (Cesar et al., 2015). Understanding the signaling pathways that make up the regulatory network in the marbling process can help steer the breeding process (Thaller et al., 2003). Therefore, animal breeding specialists have attempted system-oriented approaches to investigate major economic traits (Lee et al., 2010). It has been shown that marbling differences may be a function of a number of complex interactions among biological pathways. Therefore, a pathway analysis with differential gene expression patterns can result in a better understanding of muscle physiological states and their influence on beef quality and animal welfare (Hocquette et al., 2012). The public availability of transcriptomics data from beef producing cattle, provides new opportunities to explore the global gene expression in muscle to investigate physiological processes and their influence on meat sensory quality traits (Lee et al., 2010; Hocquette et al., 2012).

Within the genomic region of marbling there are several genes considered as parts of QTLs such as EDGPR1, Titin, Akirin 2, and RPL27 (Takasuga et al., 2007) which were mapped in a half-sib family of Japanese Black cattle (Yamada et al., 2006). Thus, these genes were considered as positional functional candidates for the genes responsible for marbling. This study aims at identifying genes and biological pathways regulating marbling of muscle tissue in beef cattle based on publicly available transcriptomics data obtained from a study by Sadkowski and coworkers (2014). We updated and extended the pathway collection for *B. taurus* at WikiPathways (Slenter et al., 2017) an online pathway repository, and a *B. taurus* gene product identifier mapping BridgeDb database was created to allow mapping of expression data to the gene databases identifiers used in the pathways (van Iersel et al., 2010). Sadkowski et al., 2014 measured global gene expression in skeletal muscle of three cattle breeds, i.e., Limousin, Holstein-Friesian, and Hereford, using Agilent microarray chips. Pathway and network analysis were performed to select the important biological pathways involved in marbling and their interactions.

MATERIALS AND METHODS

Transcriptomics Data Set

The study by Sadkowski et al., 2014 compared gene expression in semitendinosus skeletal muscle of well-marbled beef (Holstein-Friesian and Hereford) versus lean-marbled beef (Limousin). Their publicly available microarray data set was used in the present study (NCBI GEO GSE46411). The Holstein-Friesian, Hereford, and Limousin groups consisted of four animals each. Samples for total RNA isolation were taken instantaneously after slaughter from semitendinosus muscle and were kept in liquid nitrogen for transportation and then at −80°C until analyzed. Quality of RNA samples was evaluated using Bioanalyzer 2100 (Agilent Technologies, USA). Only samples with RIN ≥ 8 were further analyzed (Sadkowski et al., 2014).

Agilent Microarray Data Analysis

Global gene expression was measured with Agilent Two-Color Mi Bovine (V2) 4 x 44K Gene Expression Microarray oligonucleotide slides (Agilent, USA). Sadkowski and coworkers checked the quality of the data and performed LOWESS normalization. The normalized transcriptomic data compared well-marbled beef Holstein-Friesian (n = 4) or Hereford (n = 4) to lean-marbled beef (n = 4). The four log10 fold change (log10FC) values for each group comparison were averaged to obtain an estimate of the 10logFC between the entire groups. Furthermore, a one-sample t-test was performed on both sets of four values, comparing those to 0 (giving a p-value indicating the significance of these values being different from 0 = no change). Bovine genes were considered to be significantly, differentially expressed with $p \leq 0.05$ and an absolute FC ≥ 1.3 (Sadkowski et al., 2014).

B. taurus Pathway Collection

The online biological pathway repository, WikiPathways (Slenter et al., 2017), contains pathways of different species, however a *B. taurus* collection was missing. We updated and extended the pathway collection for *B. taurus*. We also created a *B. taurus* gene identifier (ID) mapping database based on mappings present in the Ensembl-based BridgeDb framework (van Iersel et al., 2010). The newly created *B. taurus* ID mapping database was used to annotate genes and proteins in pathways from WikiPathways and to perform pathways analysis. A online and freely available version of the database for the Ensembl build 85 is accessible at (http://bridgedb.org/data/gene_database/archive/r85/Bt_Derby_Ensembl_85.bridge.zip). Second, the WikiPathways homology based the homology mapper which is available at GitHub (https://github.com/PathVisio/homology.mapper) was updated to improve homology coverage for gene products that were annotated with different data sources. The pathways were converted from human pathways, with a required minimum successful conversion of at least 50% of the original human genes. Third, we manually curated all converted pathways to check whether the genes were correctly annotated and pathways are relevant in *B. taurus*. Finally, new pathways directly derived from cow breeding literature and not present in the WikiPathways collection were designed in PathVisio (v3.2.0) (Kutmon et al., 2015), the pathway creation, visualization, and analysis tool. All pathways were uploaded in gpml format to WikiPathways using the WikiPathways plugin (https://www.pathvisio.org/plugin/wikipathways-plugin/) for PathVisio.

Pathway-Based Over-representation Analysis

To analyze and visualize the molecular changes in marbling at biological process level a pathway-based over-representation analysis was performed in PathVisio (v3.2.0) (Kutmon et al., 2015). The *B. taurus* WikiPathways pathway collection, containing 286 pathways (6/30/2015), and the *B. taurus* ID mapping database, was used in the analysis. The pathways are ranked based on a standardized difference score (Z-score) based on the expected value and standard deviation of the number of significantly ($p \leq 0.05$) and differentially (absolute $FC \geq 1.3$) expressed genes in a pathway. Biological pathways significantly changed when (i) Z-score > 1.96, (ii) permuted p-value < 0.05 and (iii) minimum number of changed genes is 3. Additionally, alterations in gene expression (log10FC and p value) when comparing Hereford to Limousin were visualized on the *B. taurus* pathways with PathVisio.

Gene Ontology Overrepresentation Analysis

To find the biological processes in which differentially expressed genes were over represented while no pathways for these processes were present in the *B. taurus* collection at WikiPathways we performed Gene Ontology (GO) analysis *via* the GO-Elite web-interface (Zambon et al., 2012). GO-Elite is a flexible tool for GO-based over-representation analysis. To identify GO processes the following settings in GO-Elite were used: (i) 2000 permutations, (ii) Z-score GO pruning algorithm, (iii) Z-score threshold >1.96, (iv) p-value threshold <0.05 and (v) minimum number of changed genes is 3 (apart from the method specific permutations those are the same criteria as used for the pathway analysis). This approach not only helps to unify the characteristics and functions of the genes but also to attain a broader perspective of the muscle physiological processes and their influence on meat quality.

Integrated Network Analysis

To visualize the pathway and GO analysis results and their interactions the network analysis and visualization tool, Cytoscape (version 3.2.0), was used (Shannon et al., 2003). First, all enriched pathways and the differentially expressed genes present in these pathways were selected. Second, all changed GO processes and the Differentially Expressed genes present in these GO classes were selected. Third, both results were combined into one network showing the interaction between pathways and GO classes based on corresponding differentially expressed genes. Finally, differences in gene expression between well-marbled and lean-marbled skeletal muscle were visualized in the network.

RESULTS

Identification of Differentially Expressed Genes Between Well-Marbled and Lean-Marbled Skeletal Muscle

In the selected transcriptomic data set of beef marbling 42,990 microarray reporters were measured in both lean marbling beef (Limousin) and well marbling beef (Hereford and Holstein-Friesian) animals. Statistical analysis was performed on 29,677 reported genes that remained from the 42,990 reporters after quality control and annotation with Ensembl gene IDs. In the Hereford breed compared to the Limousin breed, 1,513 were higher expressed and 1,556 lower expressed (absolute log10FC >0.11 and p-value <0.05). When comparing the Holstein-Friesian breed to the Limousin breed, 1,772 genes were higher expressed and 2,458 lower expressed in the Holstein Friesian breed. The genes that met these criteria were used for further analysis.

Creating of *B. taurus* Pathway Collection and Pathway Design

In total, 282 human pathways were converted from human pathways to cattle pathways. All these pathways were manually checked and are available at (https://www.pathvisio.org/downloads/download-pathways/). Moreover, 4 pathways were newly created based on the bovine breeding literature: Growth hormone signaling (WP2890) (Roudbari Z and Kutmon M: Growth Hormone (GH) Signaling (*B. taurus*); (https://www.wikipathways.org/instance/WP2890), Growth hormone receptor signaling (WP2891) (Roudbari Z, Hanspers K, Evelo C, Kutmon M: Growth Hormone Receptor (GHR) Signaling (*B. taurus*) (https://www.wikipathways.org/instance/WP2891), IGF1-signaling (WP2892) (Roudbari Z, Evelo C, Willighagen E, Mélius J, Hanspers K, Kutmon M: IGF1-signaling (*B. taurus*) (https://www.wikipathways.org/instance/WP2892), and Gonadotropin- releasing hormone signaling (WP2901) (Roudbari Z, Kutmon M, Pico A, Willighagen E, Mélius J: Gonadotropin-releasing hormone (GNRH) signaling (*B. taurus*) (https://www.wikipathways.org/instance/WP2901). As an example, the newly designed GNRH signaling pathway is shown in **Figure 1**. The elements of this process are the key factors stimulating gonadotropin release from the pituitary, which controls the release of luteinizing hormone and follicle-stimulating hormone, and reproductive development in mammals.

Pathway Analysis

When comparing Hereford with Limousin breed ten biological pathways that were formerly known to be involved in marbling (Cui et al., 2012; Lim et al., 2013; Silva-Vignato et al., 2017) were found to be significantly enriched in differentially expressed genes (z-score > 1.96) (**Table 1**). Four biological processes such as: The Hypertrophy Model, P38 MAPK signaling, IL-1 signaling, and insulin signaling pathways, which are known to be important in marbling development are described in more detail, and the pathways are shown in Results section. Interestingly, some pathways not yet known to play a role in marbling were also found to be enriched in differentially expressed genes when comparing Hereford with Limousin breed. These included histone modifications and vitamin D metabolism pathways, in addition to Hereford breed that is used for meat production, the Holstein-Friesian breed that is a dairy cattle, was also

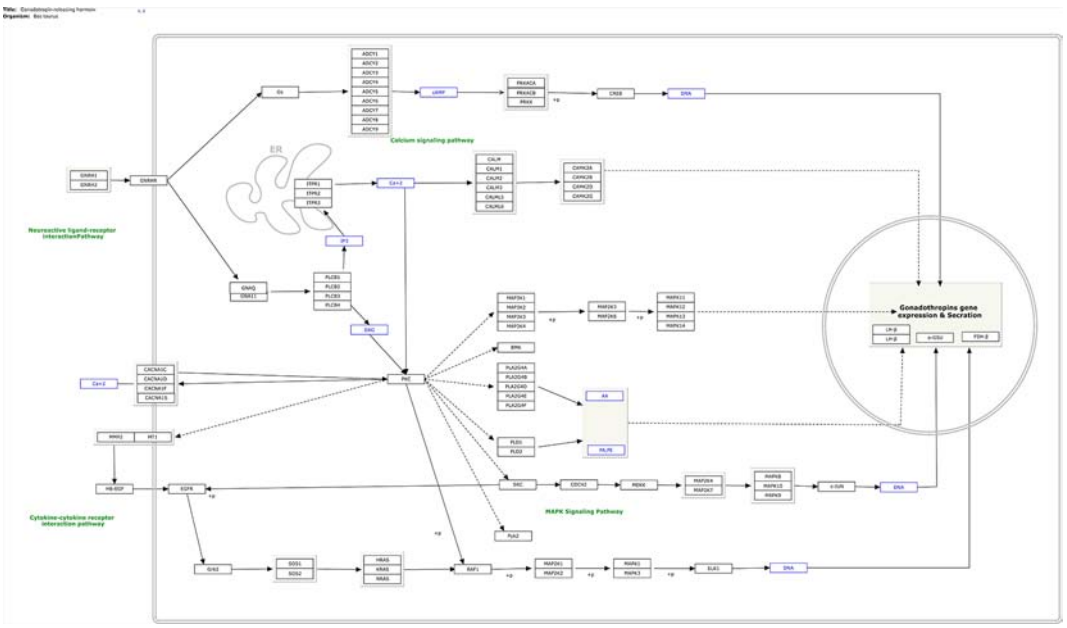


FIGURE 1 | Gonadotropin-releasing hormone (GnRH) signaling (*B. taurus*) (based on Widmann et al., 2013) and available at https://www.wikipathways.org/instance/WP2901.

compared with the Limousin breed. Pathway analysis revealed that some but not all of the marbling related pathways found for the comparison between Hereford breed and Limousin breed were also found when comparing Holstein-Friesian breed with Limousin. Examples of consistently affected pathways are the P38 MAPK signaling and the Hypertrophy Model pathways (**Table 2**).

Hypertrophy Model Pathway

Muscle hypertrophy is known to increase the muscle mass, and is determined by increased protein mass per fiber which results from an increase of protein synthesis (Glass, 2005). In the *B. taurus* Hypertrophy model (http://www.wikipathways.org/instance/WP982) the overall gene expression was higher in skeletal muscle of Hereford and Holstein-Friesian compared to Limousin (**Figure 2**). The expression of the Il1a, Ifrd1, Cyr61, ATF3, and Ankrd1 genes were significantly higher in Hereford in this pathway and the Il18, Eif4ebp1, and Il1r1 genes were significantly lower in the model (p-value < 0.05). Among them was IL-1 which plays a significant role in lipid metabolism by regulating insulin levels under physiological conditions (Matsuki et al., 2003); Atf3 which works together with p38c in a common pathway in the intestine to regulate lipid metabolism and immune homeostasis (Chakrabarti et al., 2014); and Frd1, Cyr61, and Ankrd1 genes. Two of the seven genes were significantly lower expressed including Ef4ebp1 contributing to the development of obesity through increased adipogenesis and fat metabolism alterations (Le Bacquer et al., 2007) and Il18 gene. TNF and interleukin (IL)-1 may cause negative inotropic effects indirectly through activation or release of IL-18 (Mann, 2015).

p38 MAPK Signaling Pathway

The p38 Mitogen-activated protein kinase (p38 MAPK) signaling pathway has found to be responsible for transduction of extracellular signals to their intracellular targets in different types of cells, including skeletal muscle cells and which leads to several biological effects for example proliferation, differentiation, migration, growth, apoptosis, and more specifically to muscle cells, hypertrophy (Yu et al., 2010; Silva-Vignato et al., 2017). The p38 MAPK is one intracellular signaling pathway activated during the differentiation of myogenic cell lines and this pathway is a chief regulator of skeletal muscle development (Keren et al., 2006). The p38 MAPK signaling pathway is a well-known pathway that affects lipid metabolism (Zhang and Liu, 2002). In the *B. taurus* p38 MAPK signal pathway five of the seven genes present were significantly higher expressed in Hereford and Holstein-Friesian compared to Limousin (**Figure 3**).

IL-1 Signaling Pathway

The IL-1 signal pathway is a major mediator of innate immune reactions.This pathway regulates extracellular and intracellular signaling of IL-1α or IL-1β including positive and negative-feedback mechanisms which strengthen or terminate the IL-1 response. In reply to ligand binding of the receptor, a complicated sequence of combinatorial phosphorylation and

TABLE 1 | The highest ranked pathways in skeletal muscle of Hereford compared to Limousin breed.

Pathway	Positive	Measured	Total	*Z* Score	P-value	Marbling
Hypertrophy Model	8	15	19	3.94	0.000	*
MAPK signaling pathway	34	124	167	3.51	0.002	*
Histone Modifications	13	35	43	3.41	0.000	-
IL-1 signaling pathway	15	41	54	3.17	0.003	*
P38 MAPK signaling pathway	10	27	36	2.98	0.006	*
Cardiac progenitor differentiation	11	33	54	2.72	0.006	-
T- Cell antigen Receptor signaling pathway	19	68	89	2.70	0.011	*
MicroRNAs in cardiomyocyte hypertrophy	20	73	102	2.67	0.006	*
Mitochondrial gene expression	5	11	23	2.66	0.006	*
Physiological and pathological hypertrophy of the heart	7	19	26	2.47	0.012	*
Insulin Signaling	27	111	157	2.41	0.012	*
Extracellular vesicle-mediated signaling in recipient cells	6	17	30	2.16	0.023	-
Toll-like receptor signaling pathway	18	71	92	2.15	0.032	*
Vitamin D metabolism	4	10	20	2.06	0.029	-
Alpha 6 Beta 4 signaling pathway	7	22	33	2.01	0.024	-

**Pathways previously known to be related to marbling.*
The pathways are ranked based on Z score. Per pathway the following is listed; Positive = amount of genes differentially expressed, Measured = the amount of genes measured in the study, Total = the amount of genes in the pathway and P-value = the significance level.

TABLE 2 | The highest ranked pathways in skeletal muscle of Holstein-Friesian compared to Limousin breed.

Pathway	Positive (*r*)	Measured (*n*)	Total	*Z* Score	P-value	Marbling
P38 MAPK signaling pathway	14	27	36	3.95	0.000	*
Quercetin and Nf-kB/AP-1 Induced Cell Apoptosis	7	10	26	3.77	0.000	-
Glycolysis and Gluconeogenesis	18	40	67	3.68	0.000	-
Hypertrophy Model	9	15	19	3.67	0.001	*
MAPK Signaling Pathway	42	124	167	3.48	0.000	*
Insulin Signaling	37	111	157	3.15	0.001	*
Eicosanoid Synthesis	8	15	38	3.04	0.000	-
Selenium Metabolism and Selenoproteins	11	26	48	2.63	0.009	-
IL1 and megakaryocytes in obesity	9	20	25	2.6	0.008	*
EGF/EGFR Signaling Pathway	34	110	156	2.5	0.019	-
Interferon type I signaling pathways	14	37	54	2.47	0.015	-
Cori Cycle	6	12	30	2.43	0.006	-
Integrated Cancer pathway	12	31	45	2.38	0.016	-
Myometrial Relaxation and Contraction Pathways	36	120	156	2.37	0.014	-
Pathogenic Escherichia coli infection	14	39	54	2.24	0.031	-

**Pathways previously known to be related to marbling.*
The pathways are ranked based on Z score. Per pathway the following is listed; Positive = amount of genes differentially expressed, measured = the amount of genes measured in the study, Total = the amount of genes in the pathway and P-value = the significance level.

ubiquitination events lead to activation of nuclear factor kB signaling and the JNK and p38 mitogen-activated protein kinase pathways (Weber et al., 2010). The members of the *B. taurus* IL-1 signaling pathway (http://www.wikipathways.org/instance/WP3271), such as IL-1α, IL-1β, MAP3K1, UBE2N, MAPK14, REL, ATF2, and JUN were significantly up-regulated in Hereford breed versus Limousin breed (**Figure 4**). Among them IL-1α which was found to play a role as an inhibitor of the expression of peroxisome proliferator-activated receptor gamma (PPARG), a key transcriptional factor for adipocytes differentiation, (Um et al., 2011). IL-1β has been reported to inhibit adipocyte differentiation from preadipocytes and to reduce the lipid content in mature adipocytes (Simons et al., 2005). Some of aforementioned genes: MAP3K1, UBE2N, MAPK14, REL, ATF2 can directly bind to the peroxisome proliferator-activated receptor promoter and activate transcription to regulate adipocyte differentiation (Maekawa et al., 2010). The significantly down-regulated genes are: RELA, MAPK1, IKBKG, MAP2K4, SQSTM1, and IL1R1 (**Figure 4**). Some of them are known to participate in lipid metabolism processes; activation of p62/SQSTM1 and peroxisome proliferator-activated receptor gamma is induced by palmitate internalization, which triggers lipid metabolism and limits inflammation (Krausgruber et al., 2011).

Insulin Signaling Pathway

Genes engaged in the insulin signaling pathway regulate several aspects of cellular function, including most notably the regulation of cellular growth and maintaining glucose homeostasis (DeBosch and Muslin, 2008). For Hereford vs. Limousine comparison, twenty six genes present in the *B. taurus* Insulin signaling pathway (http://www.wikipathways.org/instance/WP966) showed significant expression differences. Fourteen genes were identified as up-regulated in Hereford breed (SOS2, PIK3R3, PIK3CA, PIK3C2A, CBLB, CBLC, SNAP25, JUN, EGR1, MAP3K1, MAPK14. ENPP1, and XBP1), and twelve were down-regulated in Hereford breed (PFKM, PFKL, ARF1, STXBP2, EIF4EBP1, PIK3CD, GAB1, IGF1R, MAPK1, MAPK13, MAP2K4, and ELK1) when compared to Limousin breed (**Figure 5**). The involvement of some of upregulated and downregulated genes in lipid accumulation processes were earlier confirmed, including; PIK3CA (Foukas et al., 2013), JUN (Guo et al., 2016), EGR1 (Singh et al., 2015), p38 MAPK lipid accumulation (Sun et al., 2012), NPP1 (Pan et al., 2011), XBP-1 (Zhao et al., 2012), 4E-BP1/2 (Singh et al., 2015), and IGF-1R (Freude et al., 2012).

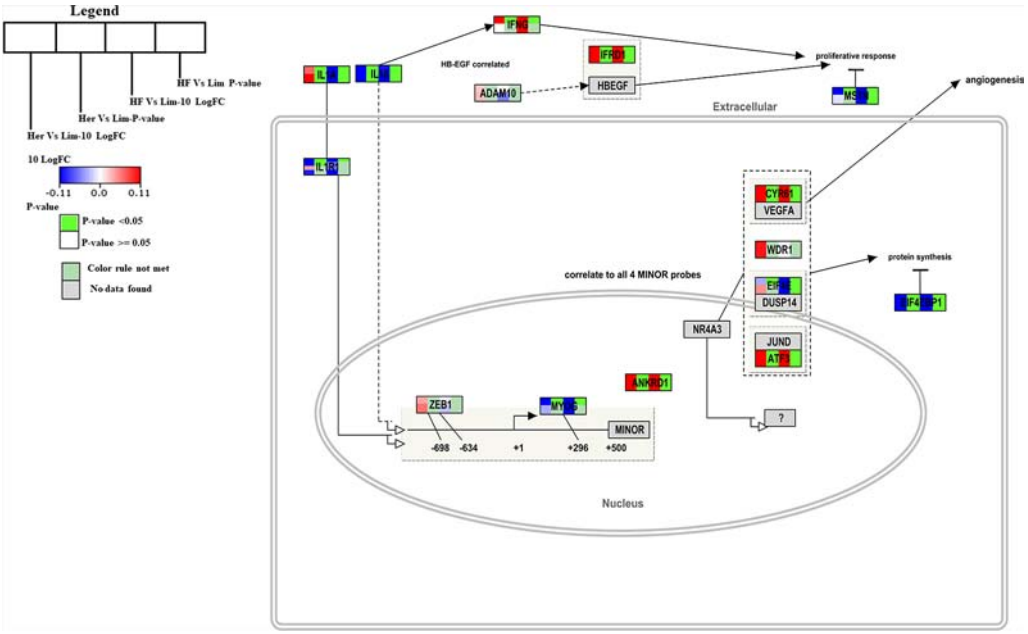


FIGURE 2 | Skeletal muscle gene expression in Hereford and Holstein-Friesian vs Limousin visualized on the Hypertrophy model. In the hypertrophy model from WikiPathways (WP982) the changes in gene expression between Hereford and Holstein-Friesian with Limousin in skeletal muscle are visualized. The logFC (Hereford and Holstein-Friesian vs Limousin) is indicated with a color gradient (blue to red over white), i.e., blue represents a negative value (= lower expressed in Hereford and Holstein-Friesian) and red a positive value (= higher expressed in Hereford and Holstein-Friesian). The p-value is colored based on a rule, i.e. p-value < = 0.05 (= significant) is shown in green and p-value > 0.05 in white.

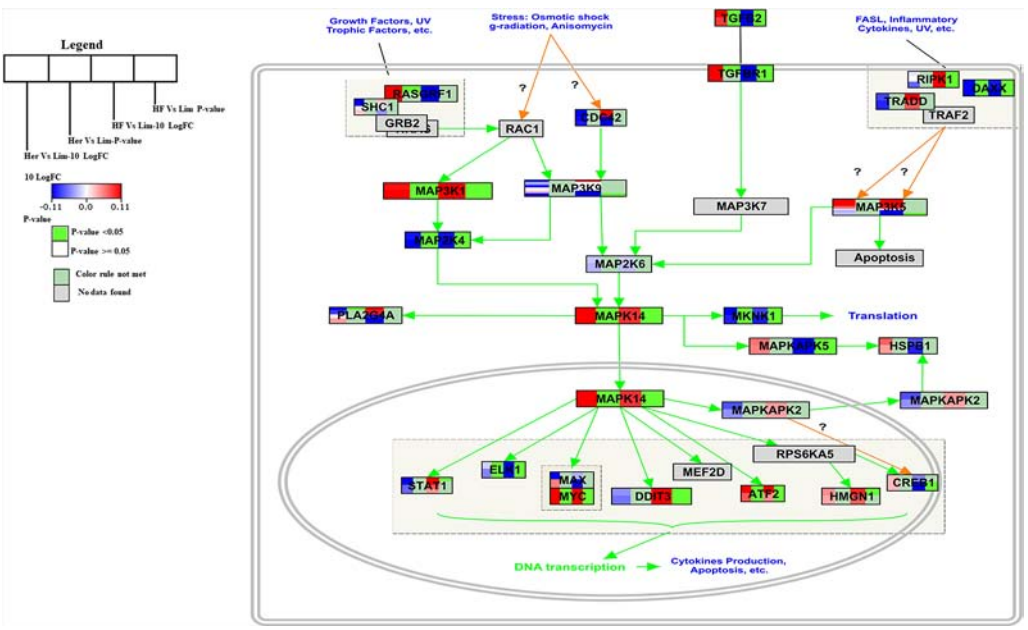


FIGURE 3 | Skeletal muscle gene expression in Hereford and Holstein-Friesian vs Limousin visualized on the p38 MAPK signaling. In the p38 MAPK signaling pathway from WikiPathways (WP1037) the changes in gene expression between Hereford and Holstein-Friesian with Limousin in skeletal muscle are visualized. The logFC (Hereford and Holstein-Friesian vs Limousin) is indicated with a color gradient (blue to red over white), i.e., blue represents a negative value (= lower expressed in Hereford and Holstein-Friesian) and red a positive value (= higher expressed in Hereford and Holstein-Friesian). The p-value is colored based on a rule, i.e., p-value < = 0.05 (= significant) is shown in green and p-value > 0.05 in white.

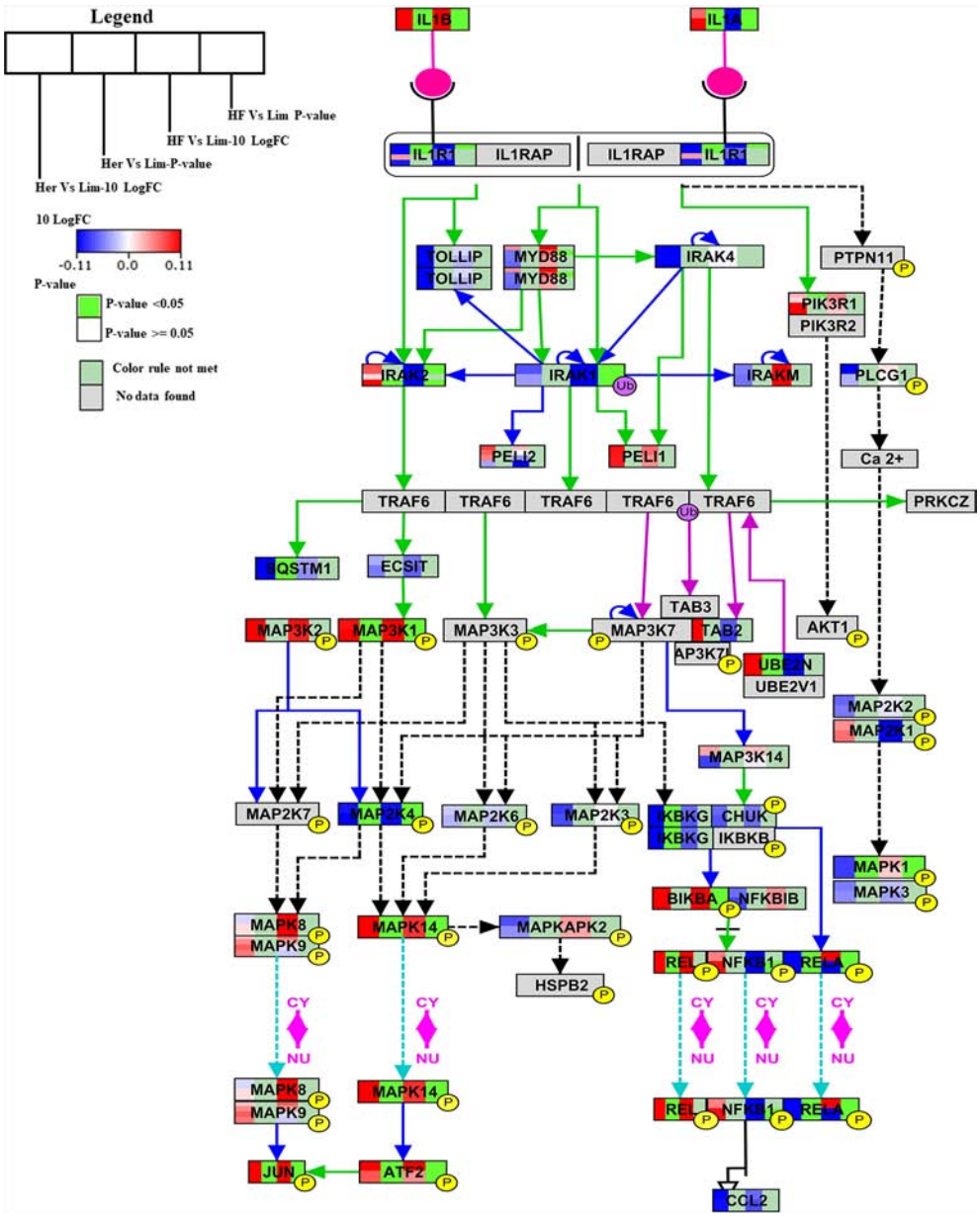


FIGURE 4 | Skeletal muscle gene expression in Hereford and Holstein-Friesian vs Limousin visualized on IL-1 signaling. In the IL-1 signaling pathway from WikiPathways (WP3271) the changes in gene expression between Hereford and Holstein-Friesian with Limousin in skeletal muscle are visualized. The logFC (Hereford and Holstein-Friesian vs Limousin) is indicated with a color gradient (blue to red over white), i.e., blue represents a negative value (= lower expressed in Hereford and Holstein-Friesian) and red a positive value (= higher expressed in Hereford and Holstein-Friesian). The p-value is colored based on a rule, i.e., p-value < = 0.05 (= significant) is shown in green and p-value > 0.05 in white.

Gene Ontology Analysis

Pathway analysis gave an insight in the biological processes involved in marbling. However, only 69% of all measured genes are present in the investigated pathways from the *B. taurus* WikiPathways collection. In order to obtain a better insight in the biological role of the differentially expressed genes not present in WikiPathways a GO analysis was performed. (**Table 3**). This approach not only helps to unify the characteristics and functions of the genes but also to attain a broader perspective of the muscle physiological processes and their influence on meat quality related to marbling.

Integrated Network of Altered Pathways With GO-Terms

The significant pathways and GO terms were merged together and shown in **Figure 6**. Some of the highly connected nodes are IL1α,

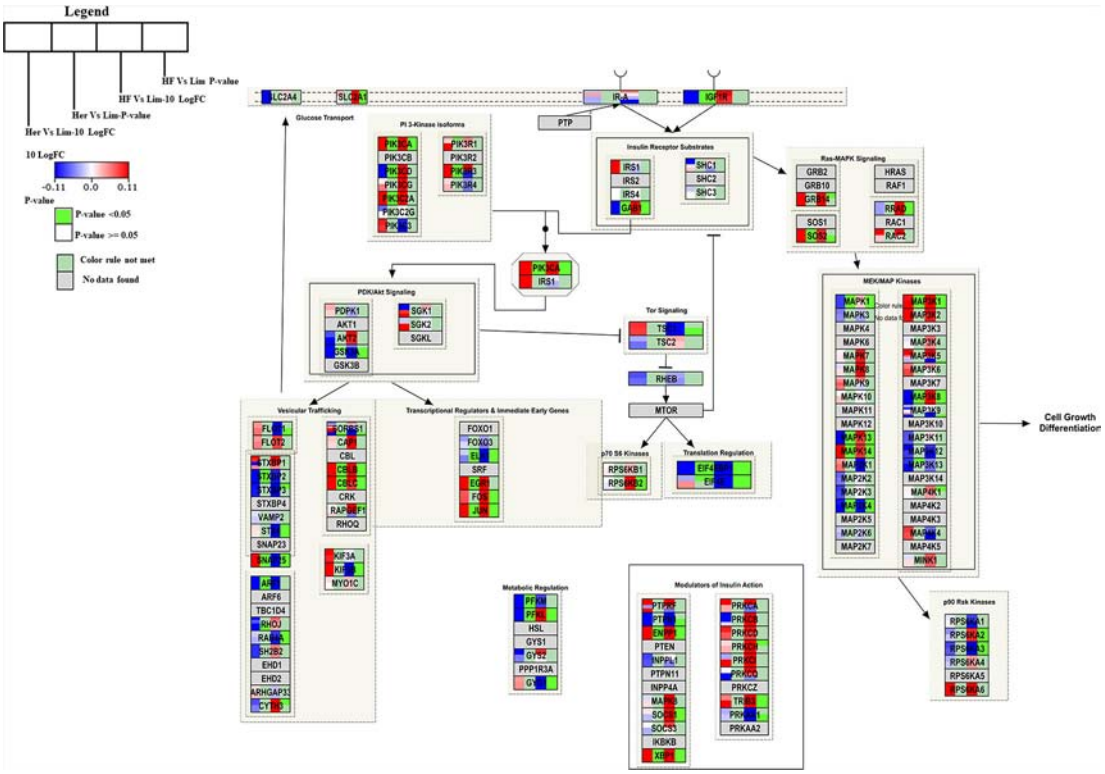


FIGURE 5 | Skeletal muscle gene expression in Hereford and Holstein-Friesian vs Limousin visualized on the Insulin signal pathway. In the Insulin signaling pathway from WikiPathways (WP966) the changes in gene expression between Hereford and Holstein-Friesian with Limousin in skeletal muscle are visualized. The logFC (Hereford and Holstein-Friesian vs Limousin) is indicated with a color gradient (blue to red over white), i.e., blue represents a negative value (= lower expressed in Hereford and Holstein-Friesian) and red a positive value (= higher expressed in Hereford and Holstein-Friesian). The p-value is colored based on a rule, i.e. p-value < = 0.05 (= significant) is shown in green and p-value > 0.05 in white.

TGFB2, PAK1, (as Pak1 deficiency led to upregulation of reverse cholesterol transporters in ApoE−/− mice in response to Western diet feeding, it might be suggested that Pak1 exerts a negative modulatory influence on these transporters and thereby might promote lipid retention in inflamed arteries which cause atherogenesis (Singh et al., 2015), TGFB2 (that TGF-β2 might control adipocyte differentiation in bone marrow stromal cells *in vivo* by inducing PPARγ phosphorylation. Whether Smad activation induced by TGFβ2 might play a role with MAPK in the inhibition of adipocyte differentiation induced by TGFβ2 *in vivo* requires more investigation (Ahdjoudj et al., 2005), IL1β, UBE2N, IRAK2, RCAN1, LRP6 (Treatment of LRP6 knockdown-Human mesenchymal stem cells with adipogenic supplements led to the accumulation of fat vacuoles, which was demonstrated by Oil Red O staining (Peröbner et al., 2012), IKBKG, TGFB3, RELA, BCL10, FGF2 (FGF-2 treatment of human preadipocytes also

TABLE 3 | The enriched processes found by GO-Analysis. A description of the process together with the positive gene number, Z Score, and P-value are given.

GOID	GO Name	GO Type	Gene Number	Z Score	P-value
GO:0051092	positive regulation of NF-kappa B transcription factor activity	Biological process	14	4.89	0.000
GO:0009826	unidimensional cell growth	Biological process	3	4.2	0.003
GO:0090257	regulation of muscle system process	Biological process	11	3.3	0.001
GO:0051059	NF-kappa B binding	Molecular function	5	3.3	0.002
GO:0051781	positive regulation of cell division	Biological process	8	3.28	0.001
GO:0003009	skeletal muscle contraction	Biological process	3	3.18	0.004
GO:0045444	fat cell differentiation	Biological process	13	3.01	0.000
GO:0048009	insulin-like growth factor receptor signaling pathway	Biological process	4	2.83	0.006
GO:0050431	transforming growth factor beta binding	Molecular function	3	2.82	0.008
GO:0048741	skeletal muscle fiber development	Biological process	3	2.26	0.020
GO:0045598	regulation of fat cell differentiation	Biological process	8	2.25	0.005
GO:0007528	neuromuscular junction development	Biological process	5	2.15	0.009

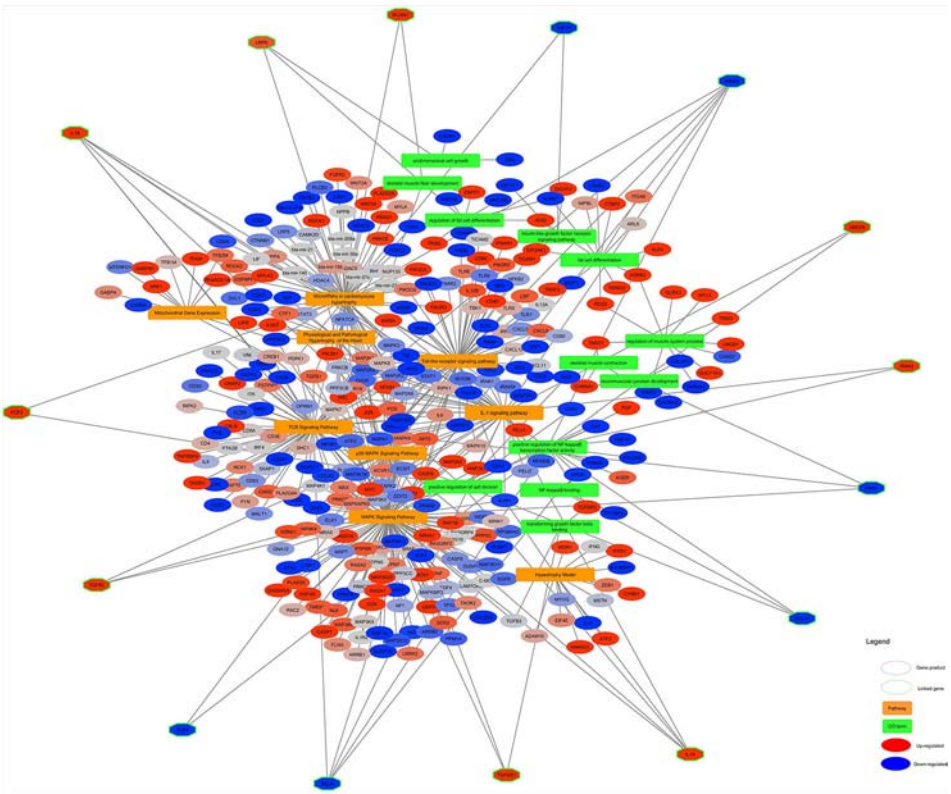


FIGURE 6 | Altered pathways and GO terms in skeletal muscle of Hereford vs Limousin. In the network, pathways are shown in orange rectangles and GO Terms are shown in green rectangles. Differentially expressed genes are shown in red (= higher expressed in Hereford) and blue (= lower expressed in Hereford). Linked genes common between two process pathway and GO terms (green diamonds) are depicted with a hexagon shape.

resulted in increased adipocyte differentiation, suggesting that this feature might be common to members of the fibroblast growth factor family, although FGF-1 was consistently the more potent adipogenic agent, particularly in cells from subcutaneous depots (Hutley et al., 2004), IGF1R, and TLR4 (TLR4 knockdown in H9C2 cardio myocytes decreases fatty acid-induced lipid accumulation (Dong et al., 2012) which are present in at least two different pathways and GO terms.

DISCUSSION

The aim of this study was to determine transcriptional profiles of high and low marbled beef with a focus on pathways of muscle cell origin that might play a role in the regulation of marbling development. The regulation of marbling is suggested to be the result of interaction of signaling pathways in muscle, fat, and intramuscular connective tissue (Hocquette, 2010). Identifying these processes with pathway analysis can help to decipher the key processes involved in marbling development. Pathway analysis revealed 17 pathways that were significantly different (z-score > 1.96) between well-marbled and lean marbled breeds. P38 MAPK signaling pathway well known to affects lipid metabolism and muscle development, was enriched when we compared gene expression in well and low marbling breeds. In addition, the signaling pathways "Hypertrophy Model", "MicroRNAs in cardiomyocyte hypertrophy" and "Physiological and pathological hypertrophy of the heart" that play a role in tissue development were affected. Interestingly, the analyses also demonstrated that pathways related to immune response (IL signaling, TCR signaling, and Toll-like receptor signaling pathways) and insulin signaling, mitochondrial gene expression and vitamin D metabolism were enriched and might act together with pathways related to lipid metabolism. We explored regulatory pathways that control gene expression in bovine muscle and the relationships between gene expression and the marbling trait to identify markers that effect on marbling. A similar study done by (Hong et al., 2014) investigated the biological characteristics of differentially expressed genes in high marbled muscle in pig compared to a low marbled muscle. They indicated that the differentially expressed genes were clustered to three group related to energy metabolism, protein synthesis, and immune response in high marbling pigs. These finding suggested that the genes related to energy metabolism, protein synthesis, and immune response contribute to growth performance and meat quality. Our results also showed differentially expressed genes take part in these processes. The hypertrophy model pathway was found

to be enriched with the highest Z score in the present study and during muscle hypertrophy there is an equilibrium between protein synthesis and degradation that might bring about protein deposition, and hence muscle growth. Together, these processes will lead to differences in muscle and fat deposition, and for this reason animals have different proportions of ribeye area and back fat thickness (Silva-Vignato et al.,2017). Mitogen-activated protein kinase (MAPK) signals have been shown to play a significant role in intracellular signaling associated with a variety of cellular activities including cell proliferation, differentiation, survival, and death (Yu et al., 2010). In mammalian cells, three MAPK families have been characterized: classical MAPK (also known as ERK), C-Jun N-terminal kinase/stress activated protein kinase and p38 kinase pathways (Zhang and Liu, 2002). Each mammalian MAPK signaling route comprises at least three components: a MAPK kinase kinase (MAP3K), a MAPK kinase (MAP2K), and a MAPK. Activated MAPKs phosphorylate various substrate proteins including transcription factors such as ATF2 and Jun (Kim & Choi, 2010). Philip and coworkers (Philip et al., 2005) discovered that the p38 MAPK played a key role in GDF-8-induced inhibition of proliferation and upregulation of the cyclin kinase inhibitor p21. In addition, their results showed a functional link between the p38 MAPK and GDF-8-activated Smad pathways, and identify an important role for the p38 MAPK in GDF-8's function as a negative regulator of muscle growth (Philip et al., 2005). In comparative muscle transcriptome associated with carcass traits of Nellore cattle, Silva-Vignato and colleagues indicated that MAPK signaling pathway involved in muscle and fat deposition, which are economically important carcass traits for beef production (Silva-Vignato et al., 2017). The third pathway found in the present study was IL-1, the IL-1 family of cytokines includes 11 proteins encoded by 11 different genes and gene regulation of IL-1 signal is activation of MKK4, MKK3, and MKK6 gene that activate NF-κB and p38 MAPK pathways (Weber et al., 2010). These two signaling pathway are needed to upregulate the expression of the key E3 ligases, *MuRF1*, which mediate the inhibition of protein synthesis (Clarke et al., 2007). Moreover, the insulin-signal transduction pathway, which was another pathway identified in the present study, is a highly conserved pathway that regulates cellular growth and when insulin binding to its cell-surface receptor, insulin receptor, activates a complex intracellular signaling network through insulin substrate proteins and the canonical PI3K and ERK cascades (Hocquette et al., 2010). Interestingly, insulin signaling is one of important factors involved in muscle development since stimulation of glucose utilization in fat and muscle cells in calves is occurring by enhancing insulin intracellular signaling (Jovanović et al., 2017).

Currently, systems biology approaches have become one of the most effective manners to accelerate the genetic improvement of beef and dairy cattle herds (Kadarmideen, 2014). It allows the selection of desired characteristics through the use of transcriptome profiles. The development of high throughput data and bioinformatics tools allow the selection of superior breeds without wasting time and money, contributing to the widespread use of transcriptome analysis in beef cattle operations. Our study shows in cattle that integration of pathway expression profiles in a systems biology approach will contribute to a better understanding of the genes and regulatory processes involved in marbling. These novel insights can be used in the future to take into account when improving the meat quality in beef cattle. The molecular mechanisms which underlie fat content in muscle can provide vital information for the production of healthier beef for human consumption.

CONCLUSIONS

The outcome of our research is the identification of biological pathways where we highlighted changed genes which are related with marbling in beef cattle. These results give a better understanding of mechanisms involving marbling in beef cattle, which is economically important carcasses trait for meat quality. Moreover, the genes involved in the *highlighted pathways can* potentially be utilized as an early biological marker for marbling fat content in breed-specific differences in growth performance and meat quality.

DATA AVAILABILITY STATEMENT

The datasets analyzed for this study can be found in the NCBI Gene Expression Omnibus (GEO) under the accession number GSE46411 and in the wikipathways.org knowledge base under the accession numbers: WikiPathways: WP2890, WikiPathways: WP2891, WikiPathways: WP2891, and WikiPathways: WP2901.

ETHICS STATEMENT

Ethical review and approval was not required for the animal study because this study is just analysis.

AUTHOR CONTRIBUTIONS

ZR, SC, and CE conceptualized and designed the study. ZR, SC, MK, LE, and JM analyzed the data. ZR collected and assembled the data and wrote the manuscript and SC, CE, and TS contributed to reviewing and editing the manuscript and also read and approved the final manuscript.

ACKNOWLEDGMENTS

The publication of the manuscript was financed by the KNOW (Leading National Research Centre) Scientific Consortium "Healthy Animal–Safe Food," under Ministry of Science and Higher Education Decision No. 05-1/KNOW2/2015 (Poland).

REFERENCES

Ahdjoudj, S., Kaabeche, K., Holy, X., Fromigue, O., Modrowski, D., Zerath, E., et al. (2005). Transforming growth factor-β inhibits CCAAT/enhancer-binding protein expression and PPARγ activity in unloaded bone marrow stromal cells. *Exp. Cell Res.* 303 (1), 138–147. doi: 10.1016/j.yexcr.2004.09.013

Bernard, C., Cassar-Malek, I., Le Cunff, M., Dubroeucq, H., Renand, G., and Hocquette, J. F. (2007). New indicators of beef sensory quality revealed by expression of specific genes. *J. Agric. Food Chem.* 55 (13), 5229–5237. doi: 10.1021/jf063372l

Cesar, A. S. M., Regitano, L. C. A., Koltes, J. E., Fritz-Waters, E. R., Lanna, D. P. D., Gasparin, G., et al. (2015). Putative regulatory factors associated with intramuscular fat content. *PloS One* 10 (6), e0128350. doi: 10.1371/journal.pone.0128350

Chakrabarti, S., Poidevin, M., and Lemaitre, B. (2014). The Drosophila MAPK p38c regulates oxidative stress and lipid homeostasis in the intestine. *PloS Genet.* 10 (9), e1004659. doi: 10.1371/journal.pgen.1004659

Clarke, B. A., Drujan, D., Willis, M. S., Murphy, L. O., Corpina, R. A., Burova, E., et al. (2007). The E3 Ligase MuRF1 degrades myosin heavy chain protein in dexamethasone treated skeletal muscle. *Cell Metab.* 6 (5), 376–385. doi: 10.1016/j.cmet.2007.09.009

Cui, H.-X., Liu, R.-R., Zhao, G.-P., Zheng, M.-Q., Chen, J. L., and Wen, J. (2012). Identification of differentially expressed genes and pathways for intramuscular fat deposition in pectoralis major tissues of fast-and slow-growing chickens. *BMC Genomics* 13 (1), 213. doi: 10.1186/1471-2164-13-213

DeBosch, B. J., and Muslin, A. J. (2008). Insulin signaling pathways and cardiac growth. *J. Mol. Cell. Cardiol.* 44 (5), 855–864. doi: 10.1016/j.yjmcc.2008.03.008

Dong, B., Qi, D., Yang, L., Huang, Y., Xiao, X., Tai, N., et al. (2012). TLR4 regulates cardiac lipid accumulation and diabetic heart disease in the nonobese diabetic mouse model of type 1 diabetes. *Am. J. Physiol.-Heart Circulatory Physiol.* 303 (6), H732–H742. doi: 10.1152/ajpheart.00948.2011

Foukas, L. C., Bilanges, B., Bettedi, L., Pearce, W., Ali, K., Sancho, S., et al. (2013). Long-term p110α PI3K inactivation exerts a beneficial effect on metabolism. *EMBO Mol. Med.* 5 (4), 563–571. doi: 10.1002/emmm.201201953

Freude, S., Schilbach, K., Hettich, M. M., Brönneke, H. S., Zemva, J., Krone, W., et al. (2012). Neuron-specific deletion of a single copy of the insulin-like growth factor-1 receptor gene reduces fat accumulation during aging. *Hormone Metab. Res.* 44 (2), 99–104. doi: 10.1055/s-0031-1298018

Glass, D. J. (2005). Skeletal muscle hypertrophy and atrophy signaling pathways. *Int. J. Biochem. Cell Biol.* 37 (10), 1974–1984. doi: 10.1016/j.biocel.2005.04.018

Guo, J., Fang, W., Sun, L., Lu, Y., Dou, L., Huang, X., et al. (2016). Reduced miR-200b and miR-200c expression contributes to abnormal hepatic lipid accumulation by stimulating JUN expression and activating the transcription of srebp1. *Oncotarget* 7 (24), 36207. doi: 10.18632/oncotarget.9183

Hocquette, J. F., Gondret, F., Baeza, E., Medale, F., Jurie, C., and Pethick, D. W. (2010). Intramuscular fat content in meat-producing animals:development, genetic and nutritional control, and identification of putative markers. *Animal* 4 (2), 303–319. doi: 10.1017/S1751731109991091

Hocquette, J. F., Cassar-Malek, I., Jurie, C., Bauchart, D., Picard, B., and Renand, G. (2012). Relationships between muscle growth potential, intramuscular fat content and different indicators of muscle fibre types in young Charolais bulls. *Anim. Sci. J.* 83 (11), 750–758. doi: 10.1111/j.1740-0929.2012.01021.x

Hocquette, J. F. (2010). Endocrine and metabolic regulation of muscle growth and body composition in cattle. Animal. *Int. J. Anim. Biosci.* 4 (11), 1797–1809. doi: 10.1017/S1751731110001448

Hong, X. U., Huang, Y., Li, W., Yang, M., Ge, C., Zhang, X., et al. (2014). Muscle biological characteristics of differentially expressed genes in wujin and landrace pigs. *J. Integr. Agric.* 13 (10), 2236–2242. doi: 10.1016/S2095-3119(13)60605-X

Hutley, L., Shurety, W., Newell, F., McGeary, R., Pelton, N., Grant, J., et al. (2004). Fibroblast growth factor 1: a key regulator of human adipogenesis. *Diabetes* 53 (12), 3097–3106. doi: 10.2337/diabetes.53.12.3097

Jovanović, L., Pantelić, M., Prodanović, R., Vujanac, I., Đurić, M., Tepavčević, S., et al. (2017). Effect of peroral administration of chromium on insulin signaling pathway in skeletal muscle tissue of Holstein calves. *Biol. Trace Element Res.* 180 (2), 223–232. doi: 10.1007/s12011-017-1007-1

Kadarmideen, H. N. (2014). Genomics to systems biology in animal and veterinary sciences: Progress, lessons and opportunities. *Livestock Sci.* 166, 232–248. doi: 10.1016/j.livsci.2014.04.028

Keren, A., Tamir, Y., and Bengal, E. (2006). The p38 MAPK signaling pathway: a major regulator of skeletal muscle development. *Mol. Cell. Endocrinol.* 252 (1–2), 224–230. doi: 10.1016/j.mce.2006.03.017

Kim, E. K., and Choi, E.-J. (2010). Pathological roles of MAPK signaling pathways in human diseases. *Biochim. Biophys. Acta (BBA)-Mol. Basis Dis.* 1802 (4), 396–405. doi: 10.1016/j.bbadis.2009.12.009

Krausgruber, T., Blazek, K., Smallie, T., Alzabin, S., Lockstone, H., Sahgal, N., et al. (2011). IRF5 promotes inflammatory macrophage polarization and T H 1-T H 17 responses. *Nat. Immunol.* 12 (3), 231. doi: 10.1038/ni.1990

Kutmon, M., van Iersel, M. P., Bohler, A., Kelder, T., Nunes, N., Pico, A. R., et al. (2015). PathVisio 3: an extendable pathway analysis toolbox. *PloS Comput. Biol.* 11 (2), e1004085. doi: 10.1371/journal.pcbi.1004085

Le Bacquer, O., Petroulakis, E., Paglialunga, S., Poulin, F., Richard, D., Cianflone, K., et al. (2007). Elevated sensitivity to diet-induced obesity and insulin resistance in mice lacking 4E-BP1 and 4E-BP2. *J. Clin. Invest.* 117 (2), 387–396. doi: 10.1172/JCI29528

Lee, S.-H., Gondro, C., van der Werf, J., Kim, N.-K., Lim, D., Park, E.-W., et al. (2010). Use of a bovine genome array to identify new biological pathways for beef marbling in Hanwoo (Korean Cattle). *BMC Genomics* 11 (1), 1–11. doi: 10.1186/1471-2164-11-623

Lim, D., Lee, S. H., Kim, N. K., Cho, Y. M., Chai, H. H., Seong, H. H., et al. (2013). Gene co-expression analysis to characterize genes related to marbling trait in Hanwoo (Korean) cattle. *Asian-Australasian J. Anim. Sci.* 26 (1), 19–29. doi: 10.5713/ajas.2012.12375

Maekawa, T., Jin, W., and Ishii, S.. (2010). The role of ATF-2 family transcription factors in adipocyte differentiation: antiobesity effects of p38 inhibitors. *Mol. Cel. Biol.* 30 (3), 613C625. doi: 10.1128/MCB.00685-09

Mann, D. L. (2015). Innate immunity and the failing heart: the cytokine hypothesis revisited. *Circ. Res.* 116 (7), 1254–1268. doi: 10.1161/CIRCRESAHA.116.302317

Matsuki, T., Horai, R., Sudo, K., and Iwakura, Y. (2003). IL-1 plays an important role in lipid metabolism by regulating insulin levels under physiological conditions. *J. Exp. Med.* 198 (6), 877–888. doi: 10.1084/jem.20030299

McAfee, A. J., McSorley, E. M., Cuskelly, G. J., Moss, B. W., Wallace, J. M. W., Bonham, M. P., et al. (2010). Red meat consumption: an overview of the risks and benefits. *Meat Sci.* 84 (1), 1–13. doi: 10.1016/j.meatsci.2009.08.029

O'Connor, S. F., Tatum, J. D., Wulf, D. M., Green, R. D., and Smith, G. C. (1997). Genetic effects on beef tenderness in Bos indicus composite and Bos taurus cattle. *J. Anim. Sci.* 75 (7), 1822–1830. doi: 10.2527/1997.7571822x

Pan, W., Ciociola, E., Saraf, M., Tumurbaatar, B., Tuvdendorj, D., Prasad, S., et al. (2011). Metabolic consequences of ENPP1 overexpression in adipose tissue. *Am. J. Physiol.-Endocrinol. Metab.* 301 (5), E901–E911. doi: 10.1152/ajpendo.00087.2011

Peröbner, I., Karow, M., Jochum, M., and Neth, P. (2012). LRP6 mediates Wnt/β-catenin signaling and regulates adipogenic differentiation in human mesenchymal stem cells. *Int. J. Biochem. Cell Biol.* 44 (11), 1970–1982. doi: 10.1016/j.biocel.2012.07.025

Philip, B., Lu, Z., and Gao, Y. (2005). Regulation of GDF-8 signaling by the p38 MAPK. *Cell. Signalling* 17 (3), 365–375. doi: 10.1016/j.cellsig.2004.08.003

Sadkowski, T., Ciecierska, A., Majewska, A., Oprządek, J., Dasiewicz, K., Ollik, M., et al. (2014). Transcriptional background of beef marbling—novel genes implicated in intramuscular fat deposition. *Meat Sci.* 97 (1), 32–41. doi: 10.1016/j.meatsci.2013.12.017

Shackelford, S. D., Koohmaraie, M., Miller, M. F., Crouse, J. D., and Reagan, J. O. (1991). An evaluation of tenderness of the longissimus muscle of angus by Hereford versus brahman crossbred heifers. *J. Anim. Sci.* 69 (1), 171–177. doi: 10.2527/1991.691171x

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., et al. (2003). Cytoscape : a software environment for integrated models of biomolecular interaction networks. *Genome Res.* 13 (11), 2498–2504. doi: 10.1101/gr.1239303.metabolite

Silva-Vignato, B., Coutinho, L. L., Cesar, A. S. M., Poleti, M. D., Regitano, L. C. A., and Balieiro, J. C. C. (2017). Comparative muscle transcriptome associated with carcass traits of nellore cattle. *BMC Genomics* 18 (1), 506. doi: 10.1186/s12864-017-3897-x

Simons, P. J., van den Pangaart, P. S., van Roomen, C. P. A. A., Aerts, J. M. F. G., and Boon, L. (2005). Cytokine-mediated modulation of leptin and adiponectin secretion during in vitro adipogenesis: evidence that tumor necrosis factor-α-

and interleukin-1β-treated human preadipocytes are potent leptin producers. *Cytokine* 32 (2), 94–103. doi: 10.1016/j.cyto.2005.08.003

Singh, M., Shin, Y.-K., Yang, X., Zehr, B., Chakrabarti, P., and Kandror, K. V. (2015). 4E-BPs control fat storage by regulating the expression of Egr1 and ATGL. *J. Biol. Chem.* 290 (28), 17331–17338. doi: 10.1074/jbc.M114.631895

Singh, N. K., Kotla, S., Dyukova, E., Traylor, J. G.Jr., Orr, A. W., Chernoff, J., et al. (2015). Disruption of p21-activated kinase 1 gene diminishes atherosclerosis in apolipoprotein E-deficient mice. *Nat. Commun.* 6, 7450. doi: 10.1038/ncomms8450

Slenter, D. N., Kutmon, M., Hanspers, K., Riutta, A., Windsor, J., Nunes, N., et al. (2017). WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research. *Nucleic Acids Res.* 46 (D1), D661–D667. doi: 10.1093/nar/gkx1064

Sun, C., Qi, R., Wang, L., Yan, J., and Wang, Y. (2012). p38 MAPK regulates calcium signal-mediated lipid accumulation through changing VDR expression in primary preadipocytes of mice. *Mol. Biol. Rep.* 39 (3), 3179–3184. doi: 10.1007/s11033-011-1084-8

Takasuga, A., Watanabe, T., Mizoguchi, Y., Hirano, T., Ihara, N., Takano, A., et al. (2007). Identification of bovine QTL for growth and carcass traits in japanese black cattle by replication and identical-by-descent mapping. *Mamm. Genome* 18 (2), 125–136. doi: 10.1007/s00335-006-0096-5

Thaller, G., Kühn, C., Winter, A., Ewald, G., Bellmann, O., Wegner, J., et al. (2003). DGAT1, a new positional and functional candidate gene for intramuscular fat deposition in cattle. *Anim. Genet.* 34 (5), 354–357. doi: 10.1046/j.1365-2052.2003.01011.x

Um, J.-Y., Rim, H.-K., Kim, S.-J., Kim, H.-L., and Hong, S.-H. (2011). Functional polymorphism of IL-1 alpha and its potential role in obesity in humans and mice. *PloS One* 6 (12), e29524. doi: 10.1371/journal.pone.0029524

van Iersel, M. P., Pico, A. R., Kelder, T., Gao, J., Ho, I., Hanspers, K., et al. (2010). The BridgeDb framework: standardized access to gene, protein and metabolite identifier mapping services. *BMC Bioinf.* 11 (1), 5. doi: 10.1186/1471-2105-11-5

Weber, A., Wasiliew, P., and Kracht, M. (2010). Interleukin-1 (IL-1) pathway. *Sci. Signal.* 3 (105), cm1–cm1. doi: 10.1126/scisignal.3105cm1

Widmann, P., Reverter, A., Fortes, M. R. S., Weikard, R., Suhre, K., Hammon, H., et al. (2013). A systems biology approach using metabolomic data reveals genes and pathways interacting to modulate divergent growth in cattle. *BMC Genomics*, 14 (1), 798. doi: 10.1186/1471-2164-14-798

Wulf, D. M., Tatum, J. D., Green, R. D., Morgan, J. B., Golden, B. L., and Smith, G. C. (1996). Genetic influences on beef longissimus palatability in charolais- and limousin-sired steers and heifers. *J. Anim. Sci.* 74 (10), 2394–2405. doi: 10.2527/1996.74102394x

Yamada, T., Taniguchi, Y., Nishimura, S., Yoshioka, S., Takasuga, A., Sugimoto, Y., et al. (2006). Radiation hybrid mapping of genes showing intramuscular fat deposition-associated expression changes in bovine musculus longissimus muscle. *Anim. Genet.* 37 (2), 184–185. doi: 10.1111/j.1365-2052.2006.01426.x

Yu, W., Chen, C., Fu, Y., Wang, X., and Wang, W. (2010). Insulin signaling: a possible pathogenesis of cardiac hypertrophy. *Cardiovasc. Ther.* 28 (2), 101–105. doi: 10.1111/j.1755-5922.2009.00120.x

Zambon, A. C., Gaj, S., Ho, I., Hanspers, K., Vranizan, K., Evelo, C. T., et al. (2012). GO-Elite: a flexible solution for pathway and ontology over-representation. *Bioinformatics* 28 (16), 2209–2210. doi: 10.1093/bioinformatics/bts366

Zhang, W., and Liu, H. T. (2002). MAPK signal pathways in the regulation of cell proliferation in mammalian cells. *Cell Res.* 12 (1), 9. doi: 10.1038/sj.cr.7290105

Zhao, S., Zhu, L., Duan, H., Liu, S., Liu, Q., Liu, W., et al. (2012). PI3K/Akt pathway mediates high glucose-induced lipid accumulation in human renal proximal tubular cells via spliced XBP-1. *J. Cell. Biochem.* 113 (10), 3288–3298. doi: 10.1002/jcb.24207

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: 27 February 2020
doi: 10.3389/fgene.2020.00091

Genome-Wide Identification and Characterization of DNA Methylation and Long Non-Coding RNA Expression in Gastric Cancer

Peng Song[1†], *Lei Wu*[2†] *and Wenxian Guan*[1*]

[1] Department of General Surgery, Nanjing Drum Tower Hospital, The Affiliated Hospital of Nanjing University Medical School, Nanjing, China, [2] Department of Laboratory Medicine, The First Affiliated Hospital of Nanjing Medical University, Nanjing, China

OPEN ACCESS

Edited by:
Manoj Kumar Kashyap, Amity University Gurgaon, India

Reviewed by:
Arivusudar Everad John, Mazumdar Shaw Medical Centre, India
Yashwanth Subbannayya, Norwegian University of Science and Technology, Norway

***Correspondence:**
Wenxian Guan
guan_wenxian@sina.com

[†]*These authors have contributed equally to this work*

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

Received: *26 November 2019*
Accepted: *27 January 2020*
Published: *27 February 2020*

Citation:
Song P, Wu L and Guan W (2020) Genome-Wide Identification and Characterization of DNA Methylation and Long Non-Coding RNA Expression in Gastric Cancer. Front. Genet. 11:91. doi: 10.3389/fgene.2020.00091

Abnormal DNA methylation, an epigenetic modification, has increasingly been linked to the pathogenesis of many human cancers. However, there has been little focus on the DNA methylation patterns of genes encoding long noncoding RNAs (lncRNAs) in gastric cancer (GC). This study comprehensively determined DNA methylation and lncRNA expression profiles in GC through genome-wide analysis. Differentially methylated loci and lncRNAs were identified by integrating multi-omics data. In total, 548 differentially methylated CpG sites in lncRNA promoters and 2,399 differentially expressed lncRNAs were screened that were capable of distinguishing GC from normal tissues. Among them, 22 differentially methylation sites in 17 lncRNAs were inversely related to expression levels. Further analysis of DNA methylation status and gene expression level in GC revealed that three CpG sites (cg01550148, cg22497867, and cg20001829) and two lncRNAs (RP11-366F6.2 and RP5-881L22.5) were significantly associated with GC patient overall survival. Molecular function analysis showed that these abnormally methylated lncRNAs were mainly involved in transcriptional activator activity. Our study identified several lncRNAs regulated by aberrant DNA methylation that have clinical utility as novel prognostic biomarkers in GC. These findings help improve the understanding of methylated patterns of lncRNAs and further our knowledge of the role of epigenetics in cancer development.

Keywords: DNA methylation, long non-coding RNA, epigenetics, prognosis, gastric cancer

INTRODUCTION

Gastric carcinoma (GC) is the fourth most prevalent malignancy and third leading cause of cancer death worldwide (Torre et al., 2015). Histologically, GC demonstrates marked heterogeneity at the cytologic level, resulting in the classification of tumor subtypes. Distinct molecular genetic profiles, morphology, and expression of specific markers have been used to investigate the diversity and characteristics of GC (Zouridis et al., 2012; Cancer Genome Atlas Research Network, 2014). Therefore, identifying potential biomarkers to further understand the pathogenesis of GC is critical.

Long noncoding RNAs (lncRNAs) are loosely defined as RNAs more than 200 bases in length with no apparent coding capacity (Mattick and Rinn, 2015). LncRNAs regulate gene expression at transcriptional and post-transcriptional levels and thus are involved in diverse biological functions. Furthermore, recent studies have demonstrated a role for lncRNAs in carcinogenesis (Spizzo et al., 2012; Zhuo et al., 2019). DNA methylation, a key epigenetic mechanism, plays a crucial role in the regulation of gene expression, genomic imprinting, genome stabilization, and chromatin modification. Aberrant DNA methylation has been reported to be involved in the formation and progressions of diseases, especially cancers (Guo et al., 2018; Xu et al., 2019). Recent studies have showed that expression alterations of lncRNA-encoding genes mediated by changes in methylation can subsequently affect their downstream targets. For instance, the lncRNA C5orf66-AS1 functions as a tumor suppressor gene in GC, and aberrant hypermethylation of the regions around its transcription start site (TSS) is associated with its expression and is cancer-specific (Guo et al., 2018). This study indicated that hypermethylation of the C5orf66-AS1 promoter may serve as a potential prognostic marker in predicting GC patient survival. Shahabi et al. identified an epigenetically deregulated lncRNA linc00261, whose expression was lost in lung adenocarcinoma through DNA methylation silencing. The authors found that linc00261 acted upstream of ATM activation to facilitate DNA damage response activation and its loss resulted in malignant phenotypes and predisposed lung cells to cancer development (Shahabi et al., 2019). Additionally, lncRNAs showing aberrant DNA methylation may serve as potential epigenetically-based diagnostic factors. Silencing from CpG-island methylation of promoter-induced transcribed ultraconserved regions (T-UCRs) is common in many tumors and linked to colorectal cancer diagnosis (Kottorou et al., 2016; Honma et al., 2017). Therefore, elucidating the relationship between DNA methylation and lncRNA expression is essential for understanding GC development and potentially identifying new prognostic or diagnostic markers.

Here, we employed multigenomic data from The Cancer Genome Atlas (TCGA) and Gene Expression Omnibus (GEO) datasets to systematically characterize global DNA methylation levels, lncRNA expression profiles, and clinical features in GC. Our results decode the landscape of DNA methylation-mediated regulation for lncRNAs and provide promising biomarkers in the diagnosis and treatment of GC.

METHODS

DNA Methylation and Gene Expression Data

The DNA methylation array data (Illumina Infinium Human Methylation27, 450 BeadChip) were downloaded from the UCSC Xena browser (https://xenabrowser.net/). A Human Methylation27 BeadChip array of GC (GSE30601) was obtained from the GEO database (https://www.ncbi.nlm.nih.gov/geo/). Level-3 RNA-sequencing data (HTSeq-Counts and HTSeq-FPKM-UQ) and the clinicopathological and survival data of GC patients were also downloaded from the Xena website.

Analysis of DNA Methylation Data

Differentially methylated CpG sites (DMCs) and differentially methylated regions (DMRs) between GC samples and adjacent tissues were identified using the minfi package (Version: 1.32.0; http://www.bioconductor.org/packages/release/bioc/html/minfi.html) (Fortin et al., 2017). Bump hunting method was applied to identify DMRs. False discovery rate (FDR) was calculated from multiple testing corrections of raw P-value by the Benjamini and Hochberg method (Benjamini et al., 2001). The genomic annotation of each CpG site was conducted using the hm27.hg38.manifest file (http://zwdzwd.io/InfiniumAnnotation). The coordinates of the individual lncRNA were extracted from GENCODE v22 (https://www.gencodegenes.org/human/release_22.html). After the preprocessing the coordinates of CpG sites and lncRNAs, we further integrated both information based on the genomic location, considering differentially methylated loci within promoter regions (DNA sequences between –2,500 and 1,000 bp relative to the putative TSS). Manhattan plot was constructed to depict the distribution of CpG sites according to FDR *via* qqman package (Version: 0.1.4; https://cran.r-project.org/web/packages/qqman/index.html) (Turner, 2018).

Differential Long Non-Coding RNA Expression Analysis

Read count tables were imported into the edgeR package for identifying differentially expressed transcripts (Version: 3.7, https://bioconductor.org/packages/release/bioc/html/edgeR.html) (McCarthy et al., 2012). LncRNA catalogue was retrieved from GENCODE v22. Genes with FDR < 0.05 and absolute fold change (FC) > 2 were considered differentially expressed lncRNAs (DElncs).

Integrated Analysis of DNA Methylation and Long Non-Coding RNA Expression

The correlation analysis between DMCs and DElncs was calculated and those with |coefficient of correlation| > 0.3 and P-value < 0.05 were considered significant. The visualization of the potential regulation of CpG sites to genes was constructed in Cytoscape 3.7.1 (Shannon et al., 2003).

Functional Annotation and Enrichment Analysis for Long Non-Coding RNAs

ClusterProfiler tool (Version: 3.8.1, https://bioconductor.org/packages/release/bioc/html/clusterProfiler.html) was used to perform Gene Ontology (GO) function and Gene Set Enrichment Analysis (GSEA) for DElncs with the DMCs (Yu et al., 2012). Spearman's correlation coefficients of expression levels between DElncs and protein-coding RNAs were calculated. The deregulated protein-coding genes were considered for GO analysis, setting parameters as "pAdjustMethod" = "BH," "pvalueCutoff" = 0.05, and "qvalueCutoff" = 0.05 for multiple

comparisons. Terms from GO (molecular function) database slice were tested for enrichment. Mappings between GO terms and Entrez Gene IDs relied on the regularly updated R package org.Hs.eg.db (Version 3.10.0; https://bioconductor.org/packages/release/data/annotation/html/org. Hs.eg.db.html). The value of $\log_2$ (FC) calculated by edgeR package was used as ranking metric for GSEA. We used the canonical pathways sub-collection of the C2 collection in the Molecular Signatures Database as the gene sets in the analysis. The leading-edge subset of genes in an enriched gene set are defined as those that appear in the ranked list before the point at which the running sum reaches its maximum deviation from zero.

Statistical Analysis

Cox proportional hazard regression analyses were carried out to compare clinical features, DNA methylation, DElncs expression, and GC patients' prognosis. Survival curves were compared using Kaplan-Meier Plotter with log-rank test. All statistical analysis was two-sided and $P < 0.05$ was defined as statistically significant. Statistical analysis was performed using R programming language v.3.5.3.

RESULTS

Characteristics of the DNA Methylation Pattern in Gastric Cancer

Because the TCGA-450k set contained only two normal samples and the number was too small to reach statistical significance with respect to determining the DNA methylation profile of GC, we used the TCGA-27k set to identify DMCs and DMRs. We obtained 6,404 CpG sites with FDR < 0.05 between 48 GC and 25 non-tumor samples and identified 1,078 DMCs with a delta-beta value > 0.2. A total of 103 DMRs were identified based on the following parameters: resamples = 100, cut off = 0.2, and probe number ≥ 2. The 103 DMRs included 65 hypermethylated regions and 38 hypomethylated regions.

To identify DNA methylation alterations in lncRNA promoter regions, 3,010 CpG sites were examined. The methylation distribution in lncRNAs showed a V-shaped curve around the TSSs, indicating a relative reduction of the methylation density at the TSS (**Figure 1A**). As shown by Manhattan plot (**Figure 1B**), the CpG sites were distributed in all chromosomes, and 698 probes were found using the threshold of FDR < 0.05. We subsequently validated the DNA methylation patterns of the CpG sites of interest in an independent cohort (GSE30601) and found that 548 probes overlapped with the 698 sites reported in TCGA-27k set (**Supplementary Table S1**).

Characteristics of Long Non-Coding RNA Expression in Gastric Cancer

To determine the lncRNA expression profile in GC, RNA-seq data of 375 GC tumors and 32 normal tissues were retrieved from TCGA. Among the 6,820 lncRNAs, we identified 2,399 DElncs, including upregulated 1,830 lncRNAs and 569 downregulated lncRNAs, using the criteria of FDR < 0.05 and absolute FC > 2 (**Figure 2A**, **Supplementary Table S2**). We then analyzed the categories of the 2,399 DElncs, as shown in **Figure 2B**. Long intergenic non-coding RNAs (lincRNAs) accounted for 54.3% of all DElncs, followed by antisense transcripts (30.3%). The remaining non-coding transcript types were sense_intronic transcripts (4.5%), processed_transcripts (3.7%), and sense_overlapping transcripts (1.3%).

Integrated Analysis of Differential Methylation and Long Non-Coding RNA Expression Data

After the profiles of DNA methylation and lncRNA expression were preprocessed, we combined the two omics data for further analysis. By associating the 548 DMCs to 2,399 DElncs, 31

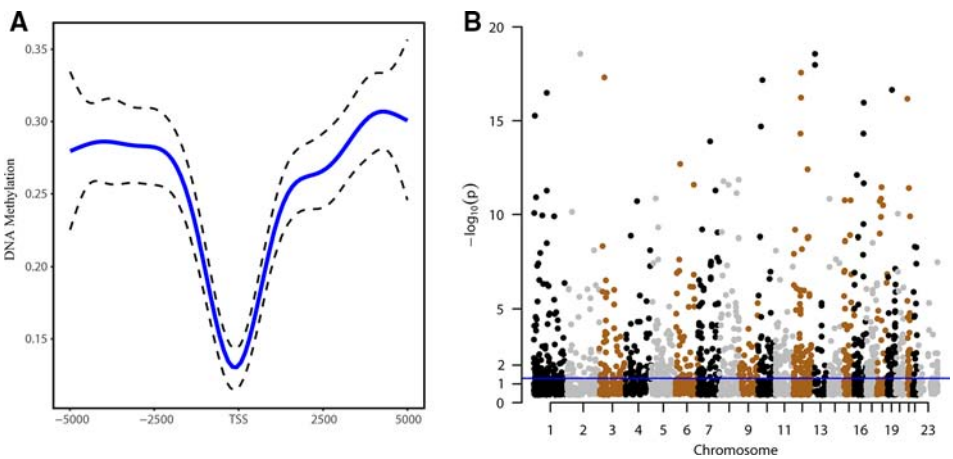


FIGURE 1 | DNA methylation patterns of genes encoding long noncoding RNAs (lncRNAs). **(A)** Distribution of the methylation levels around lncRNA genes in sperm ranging from 5 kb upstream to 5 kb downstream of the transcription start site (TSS). **(B)** Manhattan plot of CpG sites in the promoter regions of lncRNA genes; dots above the blue line indicate CpG sites with P value < 0.05.

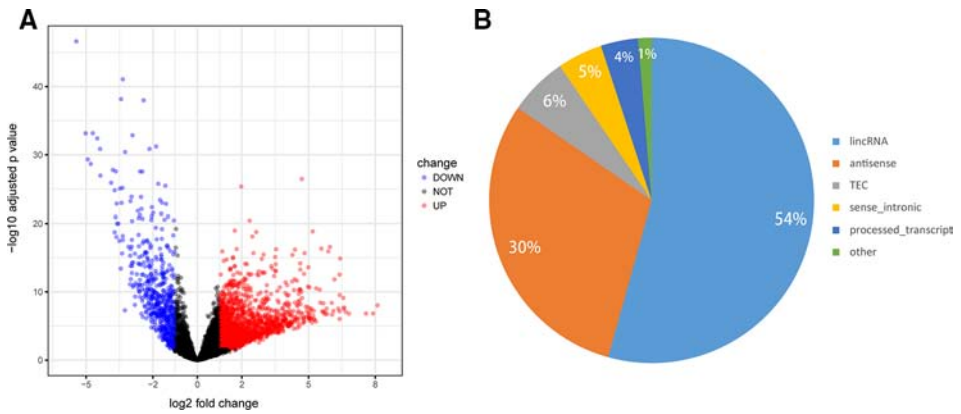


FIGURE 2 | Differential expression profiles of long noncoding RNAs (lncRNAs) in gastric cancer (GC). **(A)** Volcano plot of the differentially expressed lncRNAs between GC tumors and normal tissues. The red points represent lncRNAs that are significantly upregulated in GC while blue points represent downregulated lncRNAs (absolute FC >2 and FDR < 0.05). **(B)** Pie chart shows the number of differentially expressed lncRNAs in each category.

negative correlated pairs and 1 positive correlated pair were obtained in TCGA-27k set. DNA methylation in promoters is well known to negatively correlate with corresponding gene expression (Mosquera Orgueira, 2015). The 31 negative correlated pairs were validated in TCGA-450k set, and 22 probes showed a significantly inverse correlation with the promoter methylation of 17 aberrantly expressed lncRNAs (**Supplementary Table S3**). We used the negatively correlated pairs to construct a DNA methylation-regulated network that was composed of 39 nodes, including 6 hypermethylated DMCs, 16 hypomethylated DMCs, 13 upregulated lncRNAs, and 4 downregulated lncRNAs (**Figure 3A**). As shown in **Figure 3B**,

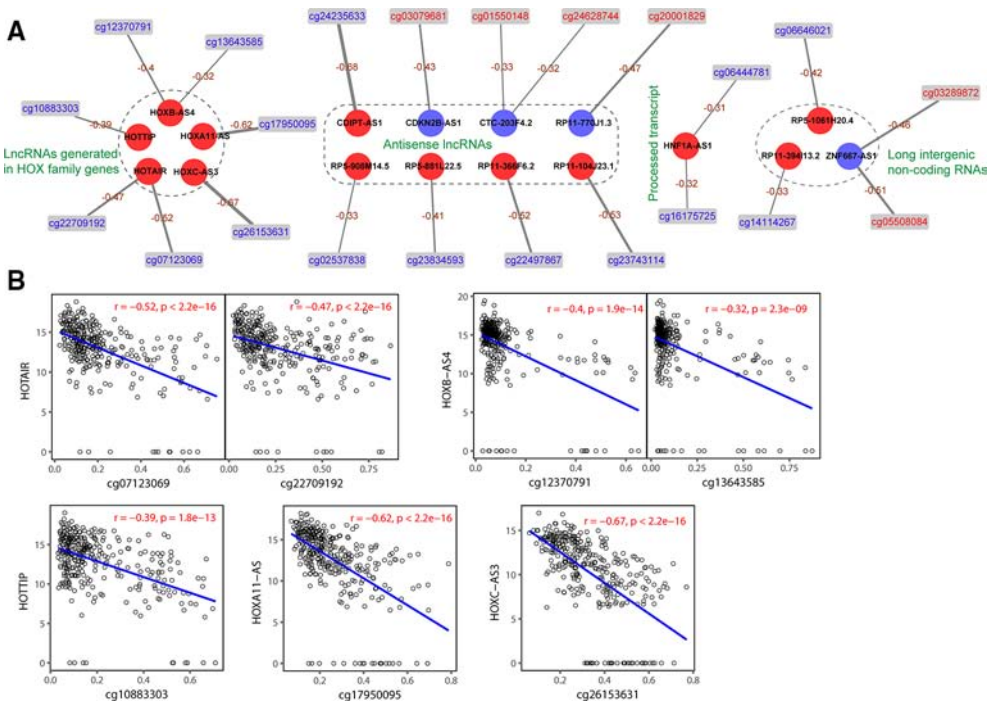


FIGURE 3 | Relation between DNA methylation and long noncoding RNA (lncRNA) expression. **(A)** Correlation between differentially methylated CpGs (DMCs) and lncRNAs. Circles and rectangles represent lncRNAs and DMCs, respectively. Red color indicates upregulated or hypermethylated, and blue indicates downregulated or hypomethylated. **(B)** Correlation (P values derive from Spearman's correlation) between DNA methylation and the expression of HOX family genes associated with five lncRNAs in matched samples.

the expression of 5 lncRNAs (HOTAIR, HOTTIP, HOXA11-AS, HOXB-AS4, and HOXC-AS3) generated in HOX family genes were negatively correlated with their methylation levels. The functions of these 17 lncRNAs are listed in **Supplementary Table S4**; only 6 of the lncRNAs have been reported to function GC (Zhang et al., 2014; Liu et al., 2015; Sun et al., 2016; Wu et al., 2017; Liu et al., 2018; Zhang et al., 2018; Song et al., 2019).

Impact of DNA Methylation and Long Non-Coding RNA Expression on Gastric Cancer Survival

Univariate Cox regression was used to evaluate the association of the 22 probes and 17 lncRNAs with overall survival in GC and the results identified three CpG sites (cg01550148, cg22497867, and cg20001829) and two lncRNAs (RP11-366F6.2 and RP5-881L22.5) with $P < 0.05$. Forest plot demonstrated that the methylation of the three probes and two lncRNAs were associated with the overall survival time of GC patients (**Figure 4A**). **Figure 4B** shows the Kaplan-Meier curves for survival in GC patients according to RP11-366F6.2 and RP5-881L22.5 expression. High expression of RP11-366F6.2 was significantly correlated with poor survival compared with low expression. Additionally, poor survival was observed for patients with low expression of RP5-881L22.5 compared with patients with high levels ($P = 0.039$). In the multivariate Cox analysis, even after adjustment by tumor stage and other covariates, the expression of the two lncRNAs was still significantly associated with patient survival ($P < 0.05$, **Table 1**).

Association of Deregulated Long Non-Coding RNAs With Biological Pathways and Processes

To better understand the biological function of the 17 DElncs, we constructed a co-expression network of deregulated protein-coding genes and lncRNAs. Using the Spearman's correlation coefficient above 0.8, a total of 32 deregulated mRNAs co-expressed with five lncRNAs (ZNF667-AS1, RP5-881L22.5, HOTAIR, HOTTIP, and HOXC-AS3) were acquired for GO enrichment analysis. Only nine protein-coding genes associated with HOTAIR, HOXC-AC3, and

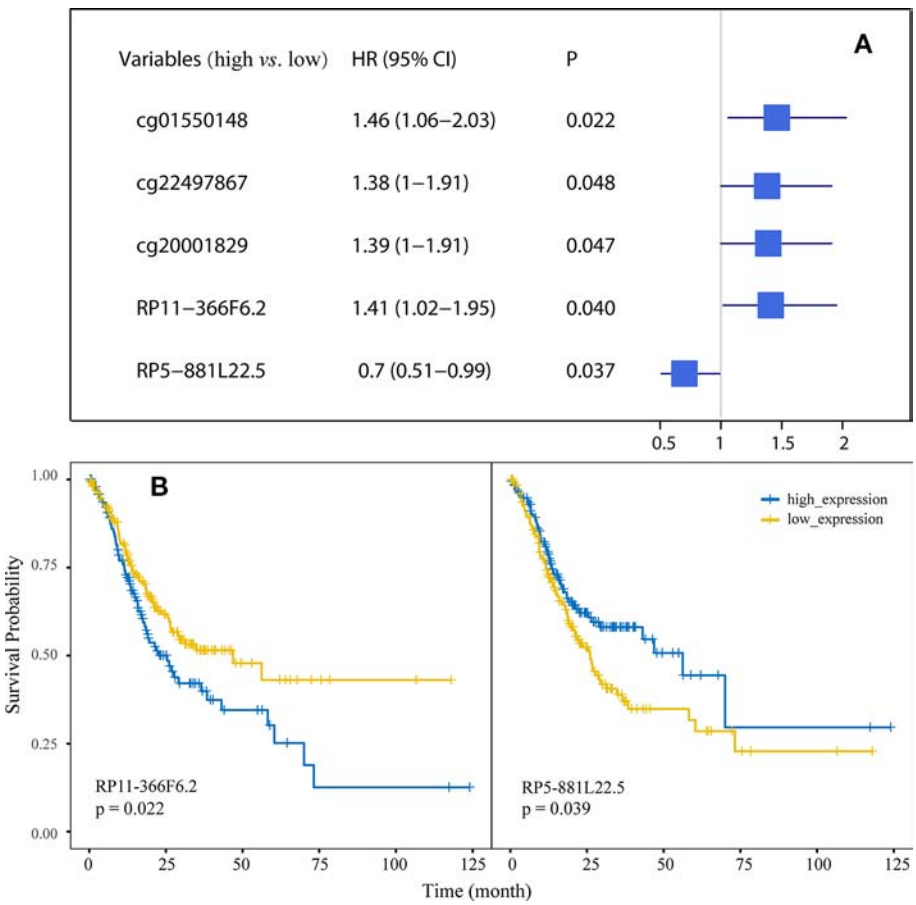


FIGURE 4 | Association of the methylation of differentially methylated CpGs (DMCs) and expression of differentially expressed long noncoding RNAs (DElncs) with survival of GC patients. **(A)** Forest plot depicting correlations between the methylation of DMCs with the survival of GC patients, using the median expression of probes as the cut-off value. **(B)** Kaplan-Meier analysis of overall survival for GC patients according to RP11-366F6.2 and RP5-881L22.5 expression.

TABLE 1 | Univariate and multivariate Cox regression analysis of variables associated with gastric cancer (GC) patient survival.

Variables	Univariate analysis			Multivariate analysis		
N=351	HR	95% CI	P	HR	95% CI	P
Age (≥67/ < 67)	1.44	1.04–2.00	0.029	1.49	1.06–2.11	0.023
Sex (male/female)	1.33	0.93–1.89	0.115	1.31	0.91–1.89	0.151
Tumor_stage (III+IV/I+II)	1.85	1.29–2.63	<0.001	1.89	1.32–2.70	<0.001
RP11-366F6.2 (high/low) *	1.41	1.02–1.95	0.040	1.45	1.03–2.05	0.033
RP5-881L22.5 (high/low) †	0.70	0.51–0.99	0.037	0.68	0.49–0.96	0.030

HR, hazard ratio; CI, confidence interval.

**Using the mean expression of genes as the cut-off value.*

† Using the median expression of genes as the cut-off value.

ZNF667-AS1 were assigned GO molecular function, involving oxidoreductase activity, nucleotide diphosphatase activity, and transcriptional activator activity (**Figure 5A**). In addition, GSEA analysis was performed to identify the associated biological processes and signaling pathways for these deregulated lncRNAs. As an example, we explored a functionally unknown lncRNA, RP5-881L22.5, with significantly hypomethylated promoter regions that was upregulated in GC tissue and implicated with prognosis (**Figure 5B**). The expression of RP5-881L22.5 was positively correlated with "VECCHI_GASTRIC_CANCER_EARLY_UP" set, in which upregulated genes could differ early GC and normal tissue samples. The "NABA_ECM_GLYCOPROTEINS" set was enriched in the RP5-881L22.5 low expression group, which implied that this lncRNA could suppress tumor metastasis (**Figure 5C**).

DISCUSSION

Cancer involves a complex regulatory network, and therefore integrating multiple omics data is required in the era of precision medicine (Olivier et al., 2019). The increasing applications of multi-omic profiling of GC have delivered new insight into the dynamics of this cancer type. In this study, we characterized DNA methylation in the promoters of lncRNA-encoding genes and inferred the potential lncRNAs regulated by aberrant DNA methylation in GC. Differential analyses were performed to compare DNA methylation and gene expression patterns between GC and normal tissues, and 548 DMCs and 2,399 DElncs were obtained. We thus identified lncRNAs (such as HOTAIR, HOTTIP, HOXA11-AS, HOXB-AS4, and HOXC-AS3) that could be regulated by aberrant DNA methylation *via* combination analysis. We further divided the potentially epigenetic regulated lncRNAs into different groups to explore their biological and clinical relationships with GC and found that the expressions of RP11-366F6.2 and RP5-881L22.5 were related to the prognosis of GC.

Methylation that interferes with transcription machinery binding to DNA has been reported to be highly associated with repression of gene transcription (Suzuki and Bird, 2008). Our research showed that a large number of lncRNAs were epigenetically deregulated by promoter methylation, and lncRNAs were globally hypomethylated, which was consistent with the previous observation that increased global DNA hypomethylation was a major event for the development and progression of cancer (Zeng et al., 2017). HOX genes are a subset of evolutionarily conserved homeobox genes that encode a class of important transcription factors that function in numerous developmental processes (Quinonez and Innis, 2014). By characterizing the transcriptional landscape of the four human HOX loci (A–D) at five base pair resolution in 11 anatomic sites, Rinn et al. identified 231 HOX non-coding RNAs (Rinn et al., 2007). HOTAIR is located in the HOXC cluster and serves as a scaffold protein by binding Polycomb repressive complex 2 (PRC2, including SUZ12, EED, and EZH2) *via* its 5′-domain and the LSD1/CoREST/REST complex *via* its 3′-domain, mediating gene silencing and reprogramming the overall chromatin dynamics in GC (Qu et al., 2019). Like HOTAIR, HOXA11-AS recruits EZH2 along with the histone demethylase LSD1 or DNMT1, which promotes proliferation and invasion of GC (Sun et al., 2016). HOTTIP enhances the expression of neighboring HOXA genes, particularly HOXA13 (Chang et al., 2016). HOXC-AS3, an antisense transcript of HOXC10, mediates gene transcriptional regulation in the tumorigenesis of GC by binding to YBX1 (Zhang et al., 2018). Genome-wide screening isolated HOXB-AS4 as specifically methylated in pancreatic cancer cells, which was useful to assess a cancer cell fraction in DNA samples (Ishihara et al., 2018).

Therapeutic targets and prognosis prediction from a comprehensive analysis of multi-omics data and clinical profiles is a critical for better understanding the biological complexity of GC. We identified two hypomethylated DElncs (RP11-366F6.2 and RP5-881L22.5) in GC that were significantly associated with overall survival. RP11-366F6.2, also called MAGEA4-AS1, is located in chrX and was reported to have significantly high expression in several tumor tissues, such as breast cancer and laryngeal squamous cell carcinoma (Yuan et al., 2017; Liu and Ye, 2019). Although pre-ranked GSEA analysis for RP11-366F6.2 returned no significantly gene sets, examining the functional roles of deregulated genes (such as MAGEA4, MAGEA10, HOXD10 and IGF2BP1) in the leading edge set indicated RP11-366F6.2 might be associated with tumor invasion and metastasis (Suzuki et al., 2008; Schultz-Thater et al., 2011; Xu et al., 2019). Regarding RP5-881L22.5, Zhu et al. developed an eleven-lncRNA signature, including this lncRNA, which could provide an effective individual mortality risk prediction and risk stratification in GC patients (Zhu et al., 2018). However, the biological functions of RP5-881L22.5 have

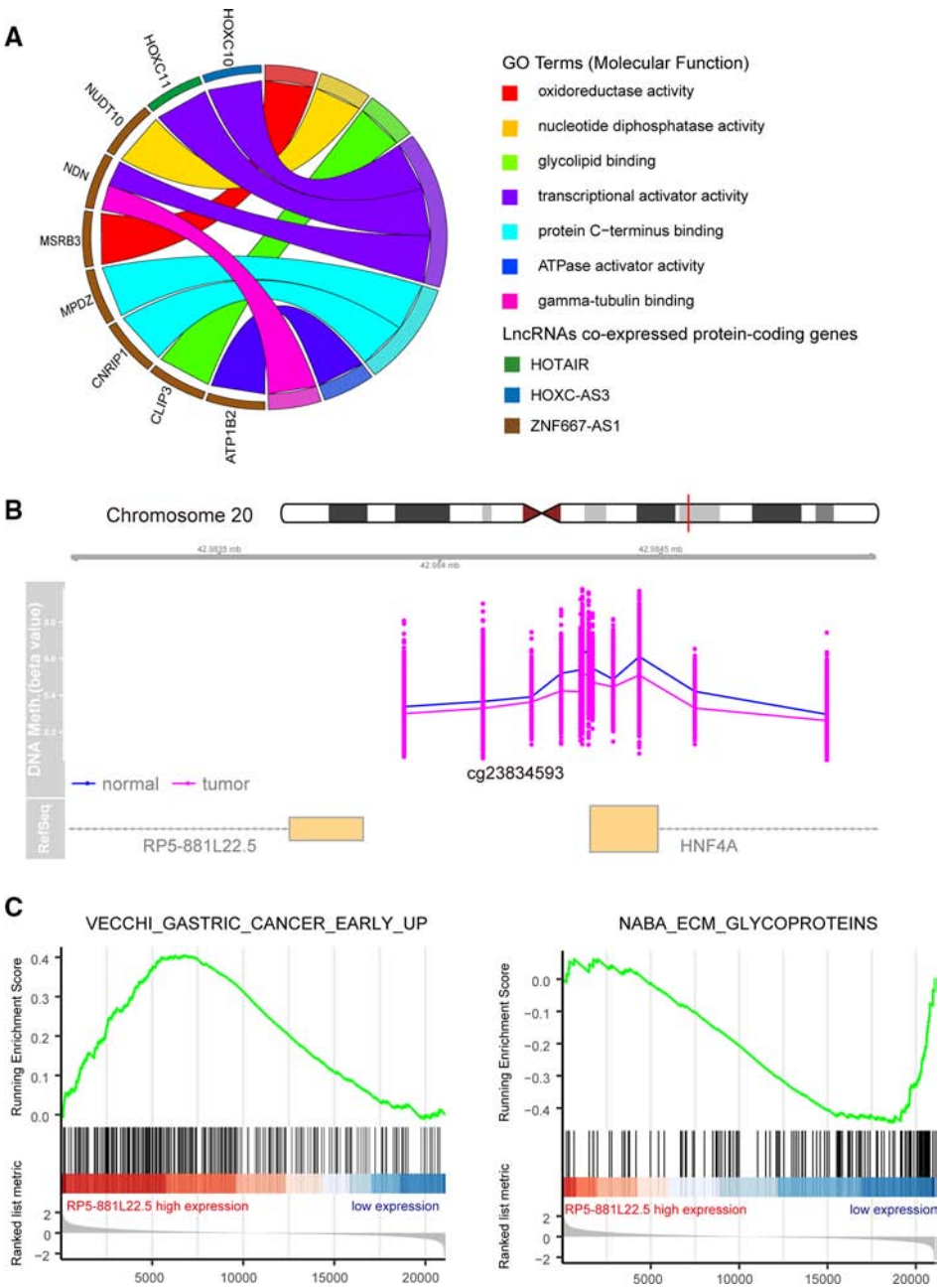


FIGURE 5 | Functional annotation for differentially methylated differentially expressed long noncoding RNAs (DElncs). **(A)** Circular plot of molecular function regulator Gene Ontology Term. **(B)** RP5-881L22.5 promoter hypermethylation in gastric cancer (GC) tumors compared with normal tissues. **(C)** Gene set enrichment analysis of RP5-881L22.5 in the The Cancer Genome Atlas (TCGA) dataset.

not been determined. GSEA results revealed that RP5-881L22.5 was likely to be involved in an extracellular matrix (ECM) interaction pathway. Glycoproteins make the ECM a cohesive network of molecules by linking cells together with structural components (Nallanthighal et al., 2019). Adhesive glycoproteins can bind to ECM components to activate downstream signaling pathways to regulate epithelial-mesenchymal transition, self-renewal, migration, differentiation, and proliferation (Naba et al., 2016; Song et al., 2017). For example, the adhesion of cancer cells to fibronectin, a major adhesive ECM glycoprotein, remodels the tumor vasculature, enhances tumorigenicity, and facilitates metastasis. This mechanism could partly explain why decreased RP5-881L22.5 expression indicated a poor prognosis for GC patients.

To investigate the effect of epigenetically deregulated lncRNA in biological processes and pathways, an integrated analysis of

DElncs and predicted mRNAs expression was performed. GO analysis revealed that these lncRNAs were involved in dysregulated transcriptional programs that invariably lead to cancer. We also found that predicted mRNAs HOXC10 and MSRB3 in GO analysis were significantly associated with overall survival in GC patients (Kim et al., 2019; Ma et al., 2019).

Several limitations in the present study should be pointed out. First, integrated analysis of genome-wide DNA methylation and lncRNA expression was based on the 27K Illumina array platform, which only contained 27,578 individual registered probes, and thus some possibly important methylation differences may be lacking from the current results. Second, the results of the present study are preliminary and primarily derived from bioinformatics analysis, and lack functional validation of the epigenetically deregulated lncRNAs. Third, due to limited availability of clinical data, it was not possible to obtain deeper insights into characterizing phenotype-genotype relationships.

In conclusion, the present results provide evidence for the changes of widespread DNA methylation of lncRNA-encoding genes in GC patients. The candidate factors identified in this study might function as potential molecular phenotypic biomarkers, especially RP11-366F6.2 and RP5-881L22.5, which were associated with prognosis. Our results help elucidate a more detailed explanation of epigenetic mechanisms for GC and deepen our understanding of the aberrantly methylated patterns in lncRNA-encoding genes.

DATA AVAILABILITY STATEMENT

The following information was supplied regarding data availability:.

The level 3 TCGA data for DNA methylation arrays and lncRNA expression are available in Xena website.

HumanMethylation27: https://gdc.xenahubs.net/download/TCGA-STAD.methylation27.tsv.gz

HumanMethylation450: https://gdc.xenahubs.net/download/TCGA-STAD.methylation450.tsv.gz

HTSeq-Counts: https://gdc.xenahubs.net/download/TCGA-STAD.htseq_counts.tsv.gz

HTSeq-FPKM-UQ: https://gdc.xenahubs.net/download/TCGA-STAD.htseq_fpkm-uq.tsv.gz

Phenotype: https://gdc.xenahubs.net/download/TCGA-STAD.GDC_phenotype.tsv.gz

Survival data: https://gdc.xenahubs.net/download/TCGA-STAD.survival.tsv.gz

The microarray data that support this study are available through the NCBI database under accession GSE30601.

The data used to support the findings of this study are included within the article.

AUTHOR CONTRIBUTIONS

PS and WG designed the study. PS and LW collected, analysed and interpreted the data. PS and LW wrote the draft. PS and WG edited the manuscript.

ACKNOWLEDGMENTS

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.00091/full#supplementary-material

REFERENCES

Benjamini, Y., Drai, D., Elmer, G., Kafkafi, N., and Golani, I. (2001). Controlling the false discovery rate in behavior genetics research. *Behav. Brain Res.* 125 (1-2), 279–284. doi: 10.1016/s0166-4328(01)00297-2

Cancer Genome Atlas Research Network (2014). Comprehensive molecular characterization of gastric adenocarcinoma. *Nature* 513 (7517), 202–209. doi: 10.1038/nature13480

Chang, S., Liu, J., Guo, S., He, S., Qiu, G., Lu, J., et al. (2016). HOTTIP and HOXA13 are oncogenes associated with gastric cancer progression. *Oncol. Rep.* 35 (6), 3577–3585. doi: 10.3892/or.2016.4743

Fortin, J.-P., Triche, T. J., and Hansen, K. D. (2017). Preprocessing, normalization and integration of the Illumina HumanMethylationEPIC array with minfi. *Bioinf. (Oxford England)* 33 (4), 558–560. doi: 10.1093/bioinformatics/btw691

Guo, W., Lv, P., Liu, S., Xu, F., Guo, Y., Shen, S., et al. (2018). Aberrant methylation-mediated downregulation of long noncoding RNA C5orf66-AS1 promotes the development of gastric cardia adenocarcinoma. *Mol. Carcinog.* 57 (7), 854–865. doi: 10.1002/mc.22806

Honma, R., Goto, K., Sakamoto, N., Sekino, Y., Sentani, K., Oue, N., et al. (2017). Expression and function of Uc.160+, a transcribed ultraconserved region, in gastric cancer. *Gastric Cancer* 20 (6), 960–969. doi: 10.1007/s10120-017-0714-9

Ishihara, H., Yamashita, S., Amano, R., Kimura, K., Hirakawa, K., Ueda, T., et al. (2018). Pancreatic cancer cell fraction estimation in a DNA sample. *Oncology* 95 (6), 370–379. doi: 10.1159/000491637

Kim, J., Bae, D. H., Kim, J. H., Song, K. S., Kim, Y. S., and Kim, S. Y. (2019). HOXC10 overexpression promotes cell proliferation and migration in gastric cancer. *Oncol. Rep.* 42 (1), 202–212. doi: 10.3892/or.2019.7164

Kottorou, A., Antonacopoulou, A., Dimitrakopoulos, F.-I., Diamantopoulou, G., Theodorakopoulos, T., Oikonomou, C., et al. (2016). Diagnostic value of methylation status of T-UCRs for colorectal cancer. *Ann. Oncol.* 27 (suppl_6), vi188. doi: 10.1093/annonc/mdw370.118

Liu, Y., and Ye, F. (2019). Construction and integrated analysis of crosstalking ceRNAs networks in laryngeal squamous cell carcinoma. *PeerJ* 7, e7380. doi: 10.7717/peerj.7380

Liu, Y. W., Sun, M., Xia, R., Zhang, E. B., Liu, X. H., Zhang, Z. H., et al. (2015). LincHOTAIR epigenetically silences miR34a by binding to PRC2 to promote the epithelial-to-mesenchymal transition in human gastric cancer. *Cell Death Dis.* 6, e1802. doi: 10.1038/cddis.2015.150

Liu, H.-T., Liu, S., Liu, L., Ma, R.-R., and Gao, P. (2018). EGR1-mediated transcription of lncRNA-HNF1A-AS1 promotes cell-cycle progression in gastric cancer. *Cancer Res.* 78 (20), 5877–5890. doi: 10.1158/0008-5472.CAN-18-1011

Ma, X., Wang, J., Zhao, M., Huang, H., and Wu, J. (2019). Increased expression of methionine sulfoxide reductases B3 is associated with poor prognosis in gastric cancer. *Oncol. Lett.* 18 (1), 465–471. doi: 10.3892/ol.2019.10318

Mattick, J. S., and Rinn, J. L. (2015). Discovery and annotation of long noncoding RNAs. *Nat. Struct. Mol. Biol.* 22 (1), 5–7. doi: 10.1038/nsmb.2942

McCarthy, D. J., Chen, Y., and Smyth, G. K. (2012). Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Res.* 40 (10), 4288–4297. doi: 10.1093/nar/gks042

Mosquera Orgueira, A. (2015). Hidden among the crowd: differential DNA methylation-expression correlations in cancer occur at important oncogenic pathways. *Front. In Genet.* 6, 163. doi: 10.3389/fgene.2015.00163

Naba, A., Clauser, K. R., Ding, H., Whittaker, C. A., Carr, S. A., and Hynes, R. O. (2016). The extracellular matrix: tools and insights for the "omics" era. *Matrix Biol.: J. Int. Soc. Matrix Biol.* 49, 10–24. doi: 10.1016/j.matbio.2015.06.003

Nallanthighal, S., Heiserman, J. P., and Cheon, D.-J. (2019). The role of the extracellular matrix in cancer stemness. *Front. In Cell Dev. Biol.* 7, 86. doi: 10.3389/fcell.2019.00086

Olivier, M., Asmis, R., Hawkins, G. A., Howard, T. D., and Cox, L. A. (2019). The need for multi-omics biomarker signatures in precision medicine. *Int. J. Mol. Sci.* 20 (19), 4781. doi: 10.3390/ijms20194781

Qu, X., Alsager, S., Zhuo, Y., and Shan, B. (2019). HOX transcript antisense RNA (HOTAIR) in cancer. *Cancer Lett.* 454, 90–97. doi: 10.1016/j.canlet.2019.04.016

Quinonez, S. C., and Innis, J. W. (2014). Human HOX gene disorders. *Mol. Genet. Metab.* 111 (1), 4–15. doi: 10.1016/j.ymgme.2013.10.012

Rinn, J. L., Kertesz, M., Wang, J. K., Squazzo, S. L., Xu, X., Brugmann, S. A., et al. (2007). Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. *Cell* 129 (7), 1311–1323. doi: 10.1016/j.cell.2007.05.022

Schultz-Thater, E., Piscuoglio, S., Iezzi, G., Le Magnen, C., Zajac, P., Carafa, V., et al. (2011). MAGE-A10 is a nuclear protein frequently expressed in high percentages of tumor cells in lung, skin and urothelial malignancies. *Int. J. Cancer* 129 (5), 1137–1148. doi: 10.1002/ijc.25777

Shahabi, S., Kumaran, V., Castillo, J., Cong, Z., Nandagopal, G., Mullen, D. J., et al. (2019). LINC00261 is an epigenetically regulated tumor suppressor essential for activation of the DNA damage response. *Cancer Res.* 79 (12), 3050–3062. doi: 10.1158/0008-5472.can-18-2034

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., et al. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* 13 (11), 2498–2504. doi: 10.1101/gr.1239303

Song, P., Jiang, B., Liu, Z., Ding, J., Liu, S., and Guan, W. (2017). A three-lncRNA expression signature associated with the prognosis of gastric cancer patients. *Cancer Med.* 6 (6), 1154–1164. doi: 10.1002/cam4.1047

Song, Y., Wang, R., Li, L.-W., Liu, X., Wang, Y.-F., Wang, Q.-X., et al. (2019). Long non-coding RNA HOTAIR mediates the switching of histone H3 lysine 27 acetylation to methylation to promote epithelial-to-mesenchymal transition in gastric cancer. *Int. J. Oncol.* 54 (1), 77–86. doi: 10.3892/ijo.2018.4625

Spizzo, R., Almeida, M. I., Colombatti, A., and Calin, G. A. (2012). Long non-coding RNAs and cancer: a new frontier of translational research? *Oncogene* 31 (43), 4577–4587. doi: 10.1038/onc.2011.621

Sun, M., Nie, F., Wang, Y., Zhang, Z., Hou, J., He, D., et al. (2016). LncRNA HOXA11-AS promotes proliferation and invasion of gastric cancer by scaffolding the chromatin modification factors PRC2, LSD1, and DNMT1. *Cancer Res.* 76 (21), 6299–6310. doi: 10.1158/0008-5472.can-16-0356

Suzuki, M. M., and Bird, A. (2008). DNA methylation landscapes: provocative insights from epigenomics. *Nat. Rev. Genet.* 9 (6), 465–476. doi: 10.1038/nrg2341

Suzuki, S., Sasajima, K., Sato, Y., Watanabe, H., Matsutani, T., Iida, S., et al. (2008). MAGE-A protein and MAGE-A10 gene expressions in liver metastasis in patients with stomach cancer. *Br. J. Cancer* 99 (2), 350–356. doi: 10.1038/sj.bjc.6604476

Torre, L. A., Bray, F., Siegel, R. L., Ferlay, J., Lortet-Tieulent, J., and Jemal, A. (2015). Global cancer statistics 2012. *CA Cancer J. Clin.* 65 (2), 87–108. doi: 10.3322/caac.21262

Turner, S. D. (2018). qqman: an R package for visualizing GWAS results using Q-Q and manhattan plots. *J. Open Source Softw.* 3 (25), 731. doi: 10.21105/joss.00731

Wu, D.-C., Wang, S. S. W., Liu, C.-J., Wuputra, K., Kato, K., Lee, Y.-L., et al. (2017). Reprogramming antagonizes the oncogenicity of HOXA13-long noncoding RNA HOTTIP axis in gastric cancer cells. *Stem Cells (Dayton Ohio)* 35 (10), 2115–2128. doi: 10.1002/stem.2674

Xu, F., Li, C. H., Wong, C. H., Chen, G. G., Lai, P. B. S., Shao, S., et al. (2019). Genome-wide screening and functional analysis identifies tumor suppressor long noncoding RNAs epigenetically silenced in hepatocellular carcinoma. *Cancer Res.* 79 (7), 1305–1317. doi: 10.1158/0008-5472.CAN-18-1659

Yu, G., Wang, L. G., Han, Y., and He, Q. Y. (2012). clusterProfiler: an R package for comparing biological themes among gene clusters. *Omics* 16 (5), 284–287. doi: 10.1089/omi.2011.0118

Yuan, N., Zhang, G., Bie, F., Ma, M., Ma, Y., Jiang, X., et al. (2017). Integrative analysis of lncRNAs and miRNAs with coding RNAs associated with ceRNA crosstalk network in triple negative breast cancer. *Onco Targets Ther.* 10, 5883–5897. doi: 10.2147/ott.s149308

Zeng, X.-Q., Wang, J., and Chen, S.-Y. (2017). Methylation modification in gastric cancer and approaches to targeted epigenetic therapy (Review). *Int. J. Oncol.* 50 (6), 1921–1933. doi: 10.3892/ijo.2017.3981

Zhang, E.-b., Kong, R., Yin, ,.D.-d., You, ,.L.-h., Sun, M., Han, L., et al. (2014). Long noncoding RNA ANRIL indicates a poor prognosis of gastric cancer and promotes tumor growth by epigenetically silencing of miR-99a/miR-449a. *Oncotarget* 5 (8), 2276–2292. doi: 10.18632/oncotarget.1902

Zhang, E., He, X., Zhang, C., Su, J., Lu, X., Si, X., et al. (2018). A novel long noncoding RNA HOXC-AS3 mediates tumorigenesis of gastric cancer by binding to YBX1. *Genome Biol.* 19 (1), 154. doi: 10.1186/s13059-018-1523-0

Zhu, M., Wang, Q., Luo, Z., Liu, K., and Zhang, Z. (2018). Development and validation of a prognostic signature for preoperative prediction of overall survival in gastric cancer patients. *Onco Targets Ther.* 11, 8711–8722. doi: 10.2147/ott.s181741

Zhuo, W., Liu, Y., Li, S., Guo, D., Sun, Q., Jin, J., et al. (2019). Long noncoding RNA GMAN, up-regulated in gastric cancer tissues, is associated with metastasis in patients and promotes translation of Ephrin A1 by competitively binding GMAN-AS. *Gastroenterology* 156 (3), 676–691.e611. doi: 10.1053/j.gastro.2018.10.054

Zouridis, H., Deng, N., Ivanova, T., Zhu, Y., Wong, B., Huang, D., et al. (2012). Methylation subtypes and large-scale epigenetic alterations in gastric cancer. *Sci. Transl. Med.* 4 (156), 156ra140. doi: 10.1126/scitranslmed.3004504

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 Bioengineering and Biotechnology

ORIGINAL RESEARCH
published: 06 March 2020
doi: 10.3389/fbioe.2020.00179

Identification of Familial Hodgkin Lymphoma Predisposing Genes Using Whole Genome Sequencing

Aayushi Srivastava[1,2,3,4†], *Sara Giangiobbe*[1,4†], *Abhishek Kumar*[1], *Nagarajan Paramasivam*[5], *Dagmara Dymerska*[6], *Wolfgang Behnisch*[7], *Mathias Witzens-Harig*[4], *Jan Lubinski*[6], *Kari Hemminki*[1,8], *Asta Försti*[1,2,3] and *Obul Reddy Bandapalli*[1,2,3,4*]

[1] Division of Molecular Genetic Epidemiology, German Cancer Research Center (DKFZ), Heidelberg, Germany, [2] Hopp Children's Cancer Center (KiTZ), Heidelberg, Germany, [3] Division of Pediatric Neurooncology, German Cancer Research Center (DKFZ), German Cancer Consortium (DKTK), Heidelberg, Germany, [4] Medical Faculty, Heidelberg University, Heidelberg, Germany, [5] Computational Oncology, Molecular Diagnostics Program, National Center for Tumor Diseases (NCT), Heidelberg, Germany, [6] Department of Genetics and Pathology, International Hereditary Cancer Centre, Pomeranian Medical University, Szczecin, Poland, [7] Department of Pediatric Oncology, Hematology and Immunology, University of Heidelberg, Heidelberg, Germany, [8] Faculty of Medicine and Biomedical Center in Pilsen, Charles University in Prague, Pilsen, Czechia

OPEN ACCESS

Edited by:
Lavanya Balakrishnan, Mazumdar Shaw Medical Centre, India

Reviewed by:
Prashanth N. Suravajhala, Birla Institute of Scientific Research, India
Raghu Metpally, Geisinger Health System, United States

***Correspondence:**
Obul Reddy Bandapalli
o.bandapalli@kitz-heidelberg.de

[†] *These authors have contributed equally to this work*

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Bioengineering and Biotechnology

Received: *03 December 2019*
Accepted: *21 February 2020*
Published: *06 March 2020*

Citation:
Srivastava A, Giangiobbe S, Kumar A, Paramasivam N, Dymerska D, Behnisch W, Witzens-Harig M, Lubinski J, Hemminki K, Försti A and Bandapalli OR (2020) Identification of Familial Hodgkin Lymphoma Predisposing Genes Using Whole Genome Sequencing. Front. Bioeng. Biotechnol. 8:179. doi: 10.3389/fbioe.2020.00179

Hodgkin lymphoma (HL) is a lymphoproliferative malignancy of B-cell origin that accounts for 10% of all lymphomas. Despite evidence suggesting strong familial clustering of HL, there is no clear understanding of the contribution of genes predisposing to HL. In this study, whole genome sequencing (WGS) was performed on 7 affected and 9 unaffected family members from three HL-prone families and variants were prioritized using our Familial Cancer Variant Prioritization Pipeline (FCVPPv2). WGS identified a total of 98,564, 170,550, and 113,654 variants which were reduced by pedigree-based filtering to 18,158, 465, and 26,465 in families I, II, and III, respectively. In addition to variants affecting amino acid sequences, variants in promoters, enhancers, transcription factors binding sites, and microRNA seed sequences were identified from upstream, downstream, 5′ and 3′ untranslated regions. A panel of 565 cancer predisposing and other cancer-related genes and of 2,383 potential candidate HL genes were also screened in these families to aid further prioritization. Pathway analysis of segregating genes with Combined Annotation Dependent Depletion Tool (CADD) scores >20 was performed using Ingenuity Pathway Analysis software which implicated several candidate genes in pathways involved in B-cell activation and proliferation and in the network of "Cancer, Hematological disease and Immunological Disease." We used the FCVPPv2 for further *in silico* analyses and prioritized 45 coding and 79 non-coding variants from the three families. Further literature-based analysis allowed us to constrict this list to one rare germline variant each in families I and II and two in family III. Functional studies were conducted on the candidate from family I in a previous study, resulting in the identification and functional validation of a novel heterozygous missense variant in the tumor suppressor gene *DICER1* as potential HL predisposition factor. We aim to identify the individual genes responsible for predisposition in the remaining two families and will functionally validate these in further studies.

Keywords: familial Hodgkin lymphoma, whole genome sequencing, predisposing genes, germline variants, variant prioritization, next generation sequencing, genetic predisposition to disease

INTRODUCTION

Hodgkin lymphoma (HL) is a lymphoproliferative malignancy originated in germinal center B-cells and is reported to account for about 10% of newly diagnosed lymphomas and 1% of all *de novo* neoplasms worldwide with an incidence of about 3 cases per 100,000 people in Western countries (Diehl et al., 2004). It is one of the most common tumors in young adults in economically developed countries, with one peak of incidence in the third decade of life and a second peak after 50 years of age.

Based on differences in the morphology and phenotype of the lymphoma cells and the composition of the cellular infiltrate, HL is subdivided into classical Hodgkin lymphoma (cHL) that accounts for about 95% of cases and nodular lymphocyte-predominant Hodgkin lymphoma (NLPHL) that accounts for the remaining 5% of cases (Kuppers, 2009).

Although familial risk for HL is reported to be among the highest of all cancers (Kharazmi et al., 2015), not many genetic risk factors have been identified. An association between various HLA class I and class II alleles and increased risk of HL has been reported (Diepstra et al., 2005), while other non-HLA susceptibility loci have been detected through genome-wide association studies (Frampton et al., 2013; Cozen et al., 2014; Kushekhar et al., 2014). The identification of major predisposing genes is a more daunting task, however, rare germline variants in *KLDHC8B*, *NPAT*, *ACAN*, *KDR*, *DICER1,* and *POT1* gene have been reported by different groups in high-risk HL families (Salipante et al., 2009; Saarinen et al., 2011; Ristolainen et al., 2015; Rotunno et al., 2016; Bandapalli et al., 2018; Mcmaster et al., 2018).

Here we report the results of whole genome sequencing (WGS) performed in three families with documented recurrence of HL. We used our Familial Cancer Variant Prioritization Pipeline (FCVPPv2) (Kumar et al., 2018) as well as two gene/variant panels based on cancer predisposing genes and variants prioritized in the largest familial HL cohort study to date in order to identify possible disease-causing high-penetrance germline variants in each family (Zhang et al., 2015; Rotunno et al., 2016). Pathway and network analyses using Ingenuity Pathway Analysis software also allowed us to gain insight into the molecular mechanisms of the pathogenesis of HL. We hope that these results can be used in the development of targeted therapy and in the screening of other individuals at risk of developing HL.

MATERIALS AND METHODS

Patient Samples

Three families with documented recurrence of HL were analyzed in this study, with a total number of 16 individuals (7 affected and 9 unaffected). HL family I and family III were recruited at the University Hospital of Heidelberg, Germany, while family II was recruited at the Pomeranian Medical University, Szczecin, Poland.

The study was approved by the Ethics Committee of the University of Heidelberg and Pomeranian Medical University, Poland. Collection of blood samples and clinical information from subjects was undertaken with a written informed consent in accordance with the tenets of the Declaration of Helsinki.

Germline DNA samples used for genome sequencing were isolated from peripheral blood using QIAamp® DNA Mini kit (Qiagen, Cat No. 51104) according to the manufacturer's instructions.

Whole Genome Sequencing, Variant Calling, Annotation and Filtering

Whole genome sequencing (WGS) of available affected and unaffected members of the three HL families was performed using Illumina-based small read sequencing. Mapping to reference human genome (assembly version Hs37d5) was performed using BWA mem (version 0.7.8) (Li and Durbin, 2009) and duplicates were removed using biobambam (version 0.0.148). The SAMtools suite (Li, 2011) was used to detect single nucleotide variants (SNVs) and Platypus (Rimmer et al., 2014) to detect indels. Variants were annotated using ANNOVAR, 1000 Genomes, dbSNP, and ExAC (Smigielski et al., 2000; Wang et al., 2010; The Genomes Project Consortium et al., 2015; Lek et al., 2016). Variants with a quality score >20 and a coverage >5×, SNVs that passed the strand bias filter (a minimum one read support from both forward and reverse strand) and indels that passed all the Platypus internal filters were evaluated further for minor allele frequencies (MAFs) with respect to the 1,000 Genomes Phase 3 and non-TCGA ExAC data. Variants with a MAF <0.1% were deduced from these two datasets. A pairwise comparison of shared rare variants was performed to check for sample swaps and family relatedness.

Data Analysis and Variant Prioritization

Prioritization of Coding Variants

Variant evaluation was performed using the criteria of our in-house developed variant prioritization pipeline (FCVPPv2) (Kumar et al., 2018). Shortly, variants with MAF < 0.1% were first filtered based on the pedigree data considering cancer patients as cases and unaffected persons as controls. The probability of an individual being a Mendelian case or true control was considered.

Variants were then ranked using the CADD tool v1.3 (Kircher et al., 2014). Only variants with a scaled PHRED-like CADD score >10, i.e., variants belonging to the top 1% of probable deleterious variants in the human genome, were considered further. Genomic Evolutionary Rate Profiling (GERP) (Cooper et al., 2005), PhastCons (Siepel et al., 2005), and PhyloP (Pollard et al., 2010) were used to evaluate the evolutionary conservation of a particular variant. GERP scores > 2.0, PhastCons scores > 0.3, and PhyloP scores ≥ 3.0 were indicative of a good level of conservation and were therefore used as thresholds in the selection of potentially causative variants.

Next, all missense variants were assessed for deleteriousness using 10 tools accessed using dbNSFP (Liu et al., 2016), namely SIFT, PolyPhen V2-HDV, PolyPhen V2-HVAR, LRT, MutationTaster, Mutation Assessor, FATHMM, MetaSVM, MetLR, and PROVEAN. Variants predicted to be deleterious by at least 60% of these tools were analyzed further. Prediction scores for nonsense variants were attained via VarSome (Kopanos et al., 2018), the final verdict on pathogenicity

offered by VarSome was based on the following tools: DANN, MutationTaster, FATHMM-MKL, FATHMM-XF, ALoFT, EIGEN, EIGEN PC, and PrimateAI.

Lastly, three different intolerance scores derived from NHLBI-ESP6500 (Petrovski et al., 2013), ExAC (Lek et al., 2016) and a local dataset, all of which were developed with allele frequency data, were included to evaluate the intolerance of genes to functional mutations. However, these scores were used merely to rank the variants and not as cut-offs for selection. The ExAC consortium has developed two additional scoring systems using large-scale exome sequencing data including intolerance scores (pLI) for loss-of-function variants and Z-scores for missense and synonymous variants. These were used for nonsense and missense variants, respectively.

Structural variants were analyzed using Canvas (version 1.40.0.1613) (https://academic.oup.com/bioinformatics/article/32/15/2375/1743834) program's SmallPedigree-WGS separately to detect the larger copy number variants. The joint genotyped VCF for all the samples in a family generated via Platypus was used as the b-allele input file along with the BAM files, and the rest of the parameters were kept default. Variants with "PASS" filters and present in all the cases in a family were processed further and variants overlapping common structural variants (AF > 1%) from gnomAD (version 2.1) were marked as common and removed. The remaining rare structural variants that affects the known cancer predisposition genes were selected for the manual inspection in IGV.

Analysis of Non-coding Variants

Variants located in the 3′ and 5′ untranslated regions (UTRs) were prioritized based on their location in regulatory regions. Putative miRNA targets at variant positions within 3′ UTRs and 1 kb downstream of transcription end sites were detected by scanning the entire dataset of the human miRNA target atlas from TargetScan 7.0 (Agarwal et al., 2015) using the intersect function of bedtools. Similarly, 5′ UTRs and regions 1 kb upstream of transcription start sites were scanned for putative enhancers and promoters using merged enhancer and promoter data from the FANTOM5 consortium as well as super-enhancer data from the super-enhancer archive (SEA) and dbSUPER. These regions were also scanned for transcription factor binding sites using SNPnexus (Dayem Ullah et al., 2018).

The regulatory nature and the possible functional effects of non-coding variants were evaluated using CADD v1.3, HaploReg V4 (Ward and Kellis, 2012), and RegulomeDB (Boyle et al., 2012), primarily based on ENCODE data (Birney et al., 2007). Epigenomic data and marks from 127 cell lines from the NIH Roadmap Epigenomics Mapping Consortium were accessed via CADD v1.3, which gave us information on chromatin states from ChromHmm and Segway. CADD also provided mirSVR scores to rank predicted microRNA target sites by a down-regulation score. These scores are based on a new machine learning method based on sequence and contextual features extracted from miRanda-predicted target sites (Betel et al., 2010). Furthermore, SNPnexus was used to access non-coding scores for each variant and to identify regulatory variants located in CpG islands.

The final selection of 3′ UTR and downstream variants was based on their CADD scores > 10 and whether or not they had predicted miRNA target site matches. Similarly, upstream and 5′ UTR variants in enhancers, promoters, super-enhancers or transcription factor binding sites with CADD scores >10 were short-listed.

Presence of Candidate Variants in 565 Cancer Predisposing and Other Cancer-Related Genes

In a study on cancer predisposing genes (CPGs) in pediatric cancers, Zhang et al. compiled 565 CPGs based on review of the American College of Medical Genetics and Genomics (ACMG) and medical literature (Zhang et al., 2015). The categories included genes associated with autosomal dominant cancer-predisposition syndromes (60), genes associated with autosomal recessive cancer-predisposition syndromes (29), tumor-suppressor genes (58), tyrosine kinase genes (23), and other cancer genes (395). We checked a list of genes corresponding to our shortlisted coding and non-coding variants for their presence in the list of genes in the aforementioned study.

Presence of Candidate Variants in Prioritized HL Genes From a Large WES-Based Familial HL Study

In a study by Rotunno et al. (2016) 2,699 variants corresponding to 2,383 genes were identified in 17 HL discovery families after filtering and prioritization. We intersected our list of candidate genes with this list of 2,383 HL genes to identify coding and non-coding variants from our shortlist in potentially causative HL genes.

Variant Validation

Specific variants of interest mentioned throughout the text (*DICER1, HLTF, LPP, PLK3, RAD51D, RELB, SH3GL2,* and *SPTAN1*) and highlighted as bold in the tables were validated using specific primers for polymerase chain reaction amplification designed with Primer3 (http://bioinfo.ut.ee/primer3-0.4.0/) and Sanger sequencing on a 3,500 Dx Genetic Analyzer (Life Technologies, CA, USA), using ABI PRISM 3.1 Big Dye terminator chemistry, according to the manufacturer's instructions. The electrophoretic profiles were analyzed manually. Segregation analysis of the prioritized variants was performed in additional family members when DNA was available. Primer details are available on request.

Ingenuity Pathway Analysis (IPA)

IPA (Qiagen; http://www.qiagen.com/ingenuity; analysis date 15/10/2019) was used to perform a core analysis to identify enriched canonical pathways, diseases, biological functions, and molecular networks among genes that passed the allele frequency cut-off, fulfilled family-based segregation criteria, met the CADD score cut-off and were not intergenic or intronic variants. Data were analyzed for all three families together. Top canonical pathways were identified from the IPA pathway library and ranked according to their significance to our input data. This significance was determined by *p*-values calculated using the right tailed Fisher's exact test.

IPA was also used to generate gene networks in which upstream regulators were connected to the input dataset genes while taking advantage of paths that involved more than one link (i.e., through intermediate regulators). These connections represent experimentally observed cause-effect relationships that relate to expression, transcription, activation, molecular modification and transport as well as binding events. The networks were ranked according to scores that were generated by considering the number of focus genes (input data) and the size of the network to approximate the relevance of the network to the original list of focus genes.

RESULTS

Whole Genome Sequencing Results

In our study, we analyzed three families with reported recurrence of Hodgkin lymphoma. Their respective pedigrees are shown in **Figure 1**.

In family I (**Figure 1A**), the proband (III-1) and her mother (II-2) were diagnosed with two different histological subtypes of classical Hodgkin lymphoma (cHL) at the ages of 7 and 34, respectively. The daughter was diagnosed with nodular sclerosis cHL and the mother with lymphocyte-rich cHL. The sample of the unaffected father (II-1) was also sequenced. Family II (**Figure 1B**) is characterized by a strong recurrence of HL. Five family members were diagnosed with HL (II-3, II-4, III-3, III-4, and III-5), of which three (III-3, III-4, and III-5) underwent WGS. In addition, the family member (II-6), who was considered as an obligatory carrier of the mutation, was sequenced as were samples and four healthy family members (III-1, III-2, III-6, and III-8) and one family member diagnosed with uterine cancer (II-1) as controls. In family 3 (**Figure 1C**), II-1 and II-2 were diagnosed with cHL, at the age of 27 and 24, respectively. Their parents (I-1, I-2) were not affected, however one of them is expected to be a carrier and analyzed accordingly.

WGS of 7 affected and 9 unaffected members from the three studied families identified a total number of 98,564, 170,550, and 113,654 variants which were reduced by pedigree-based filtering to 18,158, 465, and 26,465 in families I, II, and III, respectively.

Prioritization of Candidates According to the FCVPPv2

After pedigree-based filtering, 130, 7, and 196 exonic variants were left in families I, II, and III, respectively, with a prevalence of non-synonymous and synonymous SNVs. The predominant type of substitution was the C>T transition. Among exonic variants fulfilling pedigree-based criteria, only variants with CADD scores >10 were taken into further consideration and prioritized according to deleteriousness, intolerance, and conservational scores, as detailed in the methods section. At the end of this process, 37 potential missense variants and 9 potential nonsense mutations were prioritized for families I—III and are shown in **Tables 1**, **2**.

Pedigree-based filtering also reduced the number of potentially interesting variants located in the untranslated regions to 523 for 5'UTR variants (130 in family I, 5 in family II, and 314 in family III) and 854 for 3'UTR variants (347 in family I, 10 in family II, and 497 in family III). These variants were further prioritized based on their CADD score >10 and their localization in known regulatory regions (**Supplementary Table 1**). 5′UTR variants were analyzed by the SNPNexus tool, which allowed us to identify 4 variants located in transcription factors binding sites. In addition, the intersect function of bedtools was used to identify further 15 variants located in promoter regions and 4 located in super-enhancer regions. Among variants located in the 3′UTR region, 56 variants located in miRNA seed sequences were selected.

Analysis of structural variants resulted in identification of a large deletion in exons 9 and 10 (del5395) of Chek2 kinase gene (*CHEK2)* in family 1 that segregates with the disease.

Candidate Variants in 565 CPGs and 2383 Potentially Causative HL Genes

Intersecting our prioritized list of candidate genes with the list of 565 CPGs, we identified 11 variants in nine genes in coding and selected non-coding regions (upstream and downstream variants, 3′ and 5′ UTRs) of the known CPGs. These include *FUBP1, SEPT6, DICER1, EZR,* and *NCOA1* from family 1 and *BCL6*, *RAD51D, LPP,* and *PTCH1* from family 3 (**Table 3**). *DICER1* and *PTCH1* are known in autosomal dominant cancer-predisposition syndromes, whereas the rest are categorized as being "other cancer genes."

In addition to the identification of 11 variants in CPGs, we intersected our prioritized list of genes with a list of 2,383 genes with potentially causative variants from a large WES-based familial HL study. We found 25 variants in the coding and non-coding regions in 23 of the HL genes, with 7 coming from family I and 18 from family III (**Table 4**).

Network and Pathway Analysis With IPA

Pathway analysis of the selected variants performed with IPA showed an enrichment of mutations in genes involved in pathways essential for B-cell proliferation and activation, specifically B-cell receptor signaling, and PI3K signaling in B lymphocytes and B cell activating factor signaling (**Supplementary Table 2A**, **Supplementary Figures 1A,B**).

Similarly, the IPA network analysis generated a comprehensive picture of possible gene interactions between our candidate genes (**Supplementary Table 2B**). The top network is related to cancer, hematological disease and immunological disease, which is in complete coherence with the pathogenesis of HL. Many genes from the prioritized list of top candidates are shown to play a role in the top networks (**Figure 2**).

Literature Mining, Consolidation of Results, and Selection of Candidates

With the aim of identifying one highly penetrant dominant variant per family, we used our pipeline results and literature-based mining to determine the genes' link to Hodgkin lymphoma or immune-related processes. For family 1, we have short-listed 5 potential candidates (*DICER1, HLTF, NOTCH3, PLK3, and RELB*). Based on segregation, confirmation and functional validation, we identified *DICER1* as a candidate HL predisposing gene by showing significant down-regulation of

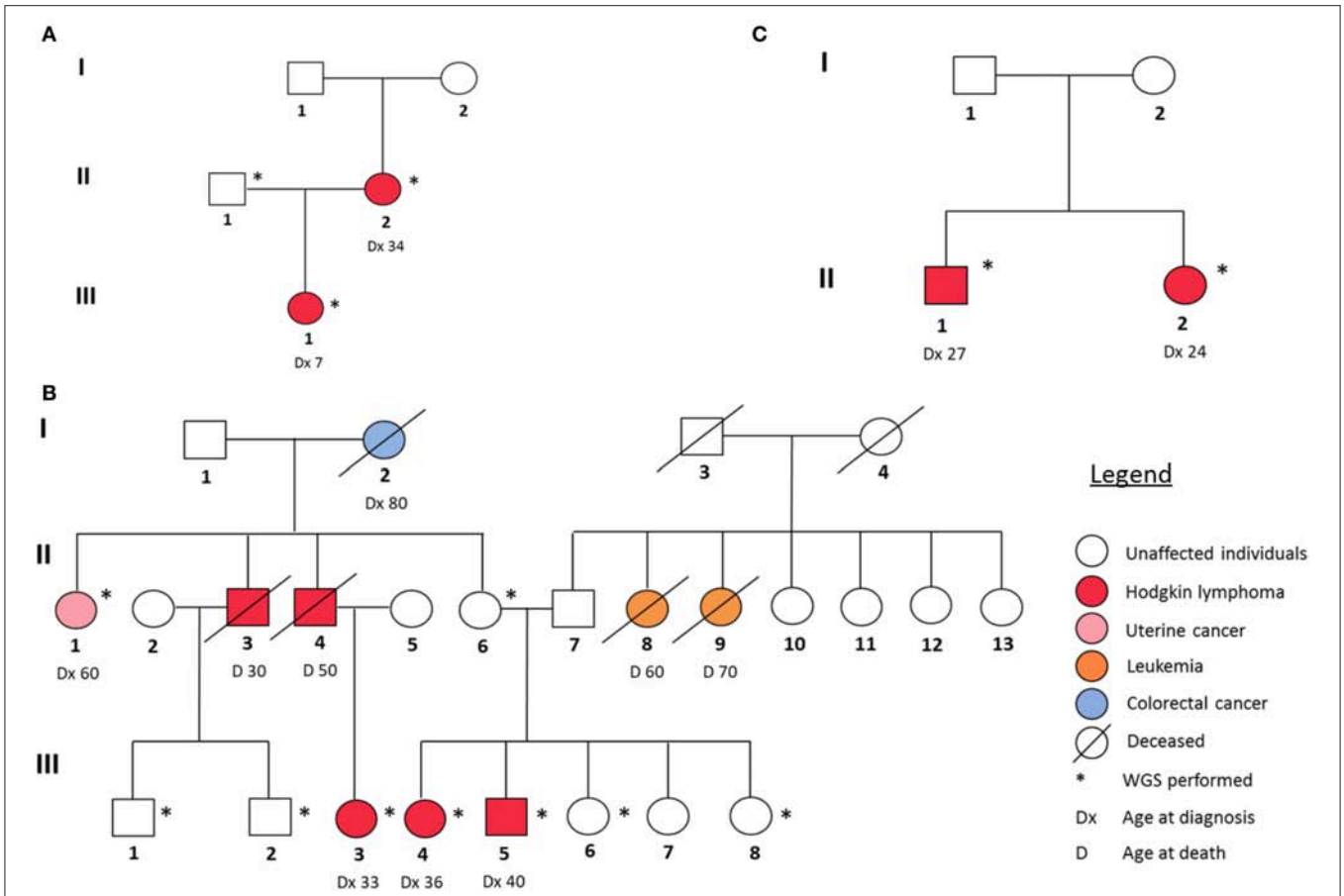


FIGURE 1 | Pedigrees of the three HL families analyzed in this study. **(A)** Family 1, **(B)** Family 2, and **(C)** Family 3.

tumor suppressor miRNAs in *DICER1*-mutated family members (Bandapalli et al., 2018). The presence of *DICER1* in the list of 565 CPGs also reinforces its status as the disease-causing variant in this family.

In family 2, three exonic variants made it to the final list (*ALAD, CERCAM,* and *SPTAN1)* of which *SPTAN1* was shown to be among the genes in one of the top IPA networks (Network 3; **Figure 2C**). No coding or non-coding variants intersected with the panel of CPGs or HL candidate genes.

Two genes stand out in family 3, namely *LPP* and *RAD51D*. Both genes were found in the list of 565 CPGs and *LPP* was additionally found in the gene list from the large cohort of HL families. Three variants in *LPP* were prioritized by the FCVPPv2 and made it to the shortlist including one stopgain variant (3_188123978_G_T), one 3′ UTR (3_188608373_A_T) and one non-synonymous missense variant (3_188123979_A_T). LPP (LIM domain containing preferred translocation partner in lipoma) is a member of the zyxin family of LIM proteins that is characterized as a promoter of mesenchymal/fibroblast cell migration. LPP has been shown to be a critical inducer of tumor cell migration, invasion and metastasis by virtue of its ability to localize to adhesions and to promote invadopodia formation (Ngan et al., 2018). A genome-wide association study of 253 Chinese individuals with B-cell NHL also identified a new susceptibility locus between *BCL6* and *LPP* that was significantly associated with the increased risk of B-cell NHL (Tan et al., 2013). On the other hand, there are no reports of an association between *RAD51D* and lymphomas; however, it is a well-established susceptibility gene in Breast-Ovarian Cancer, Familial 4 and Hereditary Breast Ovarian Cancer Syndrome (Loveday et al., 2011; Chen et al., 2018). The final selection of a candidate in this family will be based on further functional studies.

DISCUSSION

In summary, WGS data analysis of three families with reported recurrence of HL allowed us to prioritize 45 coding and 79 non-coding variants from which we subsequently selected and validated one for family I (*DICER1*), short-listed three in family II (*ALAD*, *CERCAM*, and *SPTAN1*) and two in family III (*RAD51* and *LPP*), to investigate further with validation and functional studies. For family I we have already functionally validated *DICER1* as the candidate predisposing gene in a previous study (Bandapalli et al., 2018). However, it was important to include the family in this paper, especially with regard to the integrity of the pathway and network analyses. We identified pathways related to

TABLE 1 | Top missense variants prioritized using the FCVPPv2.

Family ID	Position (Hg 19)	Gene	Effect	CADD PHRED	Int (n/3)	Del (*n*/10)	VarSome Score
Family_1	17_48746518_C_T	ABCC3	p.P652L	22.6	2	7	Uncertain significance
Family_1	1_49052793_G_A	AGBL4	p.R384C	35	2	7	Uncertain significance
Family_1	5_139909090_A_G	ANKHD1, ANKHD1-EIF4EBP3	p.N2187D	25.2	.	6	Uncertain significance
Family_1	1_160164884_T_C	CASQ1	p.I183T	26.5	2	10	Likely Benign
Family_1	**14_95560456_A_C**	**DICER1**	**p.I1711M**	**24**	**3**	**7**	**Uncertain significance**
Family_1	6_159206584_G_A	EZR	p.P75L	32	3	9	Uncertain significance
Family_1	12_8192537_G_A	FOXJ2	p.G37R	29.9	3	9	Uncertain significance
Family_1	14_88729713_C_T	KCNK10	p.A79T	27.1	3	6	Uncertain significance
Family_1	10_88705360_G_A	MMRN2	p.P58L	29.8	2	7	Uncertain significance
Family_1	5_36962227_G_A	NIPBL	p.R154Q	27.5	3	8	Uncertain significance
Family_1	2_206614449_A_G	NRP2	p.D596G	23.1	2	9	Uncertain significance
Family_1	**1_45268632_C_T**	**PLK3**	**p.T252M**	**25.3**	**3**	**6**	**Uncertain significance**
Family_1	**19_45515485_T_C**	**RELB**	**p.I152T**	**26**	**3**	**6**	**Uncertain significance**
Family_1	6_52372363_G_C	TRAM2	p.A205G	29.8	3	10	Uncertain significance
Family_1	22_18613830_C_T	TUBA8	p.A450V	24.7	3	10	Uncertain significance
Family_1	X_47272364_G_A	ZNF157	p.G298R	27.6	2	6	Uncertain significance
Family_2	9_116151739_G_C	ALAD	p.I243M	22.9	2	6	Uncertain significance
Family_2	9_131196759_G_T	CERCAM	p.A468S	24.5	2	7	Uncertain significance
Family_2	**9_131367689_C_T**	**SPTAN1**	**p.R1327C**	**34**	**3**	**6**	**Uncertain significance**
Family_3	9_139917418_C_T	ABCA2	p.G83S	26.3	2	7	Uncertain significance
Family_3	17_40971572_G_C	BECN1	p.P85R	23.4	3	6	Uncertain significance
Family_3	8_67968830_G_T	COPS5	p.P131T	23.9	3	9	Uncertain significance
Family_3	3_5246773_C_T	EDEM1	p.T160M	34	3	9	Uncertain significance
Family_3	6_131191103_G_A	EPB41L2	p.S736F	22.1	3	7	Uncertain significance
Family_3	8_28575243_G_A	EXTL3	p.R172H	23	3	6	Likely Benign
Family_3	**3_188123979_A_T**	**LPP**	**p.E24V**	**32**	**2**	**6**	**Uncertain significance**
Family_3	14_74970734_C_T	LTBP2	p.G1493R	27.7	3	10	Uncertain significance
Family_3	3_196730925_C_A	MFI2	p.D662Y	34	3	6	Uncertain significance
Family_3	17_27441099_G_A	MYO18A	p.A843V	24.3	3	6	Uncertain significance
Family_3	19_14584756_A_G	PTGER1	p.L126P	25.9	2	6	Uncertain significance
Family_3	3_49138083_G_A	QARS	p.R301C	34	2	9	Uncertain significance
Family_3	**17_33428327_G_A**	**RAD51D,RAD51L3-RFFL**	**p.R266C**	**27**	**3**	**8**	**Benign**
Family_3	11_9838541_C_T	SBF2	p.R1275H	33	2	9	Likely Benign
Family_3	**9_17761502_A_G**	**SH3GL2**	**p.N14S**	**26.1**	**3**	**9**	**Uncertain significance**
Family_3	20_35467682_G_A	SOGA1	p.R46C	32	3	7	Uncertain significance
Family_3	1_43891311_G_A	SZT2	p.A96T	31	3	6	Uncertain significance

Chromosomal positions, classifications, PHRED-like CADD scores, protein changes and the number of positive intolerance (Int) and deleteriousness (Del) scores are shown for each variant. Variants of interest that were validated by Sanger sequencing in the provided family samples have been shown in bold.

B-cell proliferation and networks related to cancer, hematological disease, immunological disease, hereditary disorders, cell death and cell survival using IPA software, helping us to prioritize genes with functions in the pathogenesis of HL. Interestingly, several genes in our gene list were related to DNA repair (e.g., *NOTCH3, RAD51,* and *SPTAN1*).

In the current study, we also identified a deletion of exon 9 and 10 in *CHEK2* in family 1. The same deletion has been reported in several unrelated patients with breast cancer of Polish origin. In that study the deletion of exon 9 and 10 in *CHEK2* was shown to lead to a premature protein truncation at codon 381 and to evoke a 2-fold increase in the risk of prostate cancer and a 4-fold increase in the risk of familial prostate cancer (Cybulski et al., 2006). The detection of mRNA of abnormal length suggests that the deletion does not lead to complete transcript loss and therefore, the effect of this truncating mutation on cancer risk may differ or work in tandem with another genetic effect, may be with *DICER1* in this family but warrants further experiments. Personalized medicine is an upcoming and promising field of medicine in which medical decisions, practices, interventions,

TABLE 2 | Top non-sense variants prioritized using the FCVPPv2.

Family ID	Position (Hg 19)	Gene	Exonic classification	Effect	CADD	Int (*n*/3)	VarSome score [I]
Family_1	10_88911115_AGT_A	FAM35A	Frameshift deletion	p.2_2del	25.8	2	PM2
Family_1	**3_148802664_C_T**	**HLTF**	**Stopgain SNV**	**p.W11X**	**37**	**2**	**PP3 (4)**
Family_1	1_177923437_CTG_C	SEC16B	Frameshift deletion	p.481_481del	36	0	Uncertain significance
Family_1	15_91546350_TG_T	VPS33B	Frameshift deletion	p.P321fs	36	3	PVS1
Family_3	7_31683260_AT_A	CCDC129	Frameshift deletion	p.D611fs	34	0	Uncertain significance
Family_3	1_21267855_C_T	EIF4G3	Stopgain SNV	p.W7X	14.54	2	PVS1, PP3 (1)
Family_3	**3_188123978_G_T**	**LPP**	**Stopgain SNV**	**p.E24X**	**40**	**2**	**PM2, PP3 (4)**
Family_3	15_24921469_G_A	NPAP1	Stopgain SNV	p.W152X	24.8	0	PM2, PP3 (3)
Family_3	1_241958547_CAG_C	WDR64	Frameshift deletion	p.836_836del	37	0	Uncertain significance

[I]VarSome Scores.
PM2, Pathogenic Moderat; PP3, Pathogenic Supporting (no. of scores predicting pathogenicity); Uncertain Significance: No scores could be found for the variant in question, PVS1, Pathogenic Very Strong.
Chromosomal positions, classifications, PHRED-like CADD scores, protein changes, the number of positive intolerance (Int) and VarSome prediction scores are included for each variant. Variants of interest that were validated by Sanger sequencing in the provided family samples have been shown in bold.

TABLE 3 | Variants corresponding to genes present in the panel of 565 known cancer predisposition genes from a study by Zhang et al. (2015).

Gene ID	HL family	HL gene	HL variant	Variant type	Variant classification	HGNC approved name	CADD_PHRED	Familial syndrome	Category
10499	1	NCOA2	8_71316112_T_TCCTCCTCCC	Indel	Upstream	Nuclear receptor coactivator 2	15.56		Other CancerGene
8880	1	FUBP1	1_78414225_A_G	SNVs	UTR3	Far upstream element (FUSE) binding protein 1	13.59		Other CancerGene
23157	1	SEPT6	X_118751062_CGTGT_C	Indel	UTR3	Septin 6	10.56		Other CancerGene
23405	1	DICER1	14_95560456_A_C	SNVs	Non-synonymous SNV	Dicer 1, ribonuclease type III	24	DICER1 syndrome, Familial Multinodular Goiter	Autosomal Dominant
7430	1	EZR	6_159206584_G_A	SNVs	Non-synonymous SNV	Ezrin	32		Other CancerGene
604	3	BCL6	3_187463568_C_A	SNVs	Upstream; downstream	B-cell CLL/lymphoma 6	13		Other CancerGene
4026	3	LPP	3_188123978_G_T	SNVs	Stopgain SNV	LIM domain containing preferred translocation partner in lipoma	40		Other CancerGene
	3	LPP	3_188123979_A_T	SNVs	Non-synonymous SNV	LIM domain containing preferred translocation partner in lipoma	32		Other CancerGene
	3	LPP	3_188608373_A_T	SNVs	UTR3	LIM domain containing preferred translocation partner in lipoma	10.5		Other CancerGene
5892	3	RAD51D	17_33428327_G_A	SNVs	Non-synonymous SNV	RAD51 paralog D	27		Other CancerGene
5727	3	PTCH1	9_98270531_C_A	SNVs	Non-synonymous SNV	Patched 1	20.4	Gorlin syndrome	Autosomal Dominant

and products are tailored to the individual patient based on their predicted response or risk of disease. The scope of this field has advanced rapidly with the advent of genomics and other omics and the possibility of implicating one gene or a set of genes in the pathogenesis of a particular disease. Thus, the identification of germline predisposing genes could be of great value in the screening of individuals at risk of developing HL, as well as in the development of personalized adjuvant therapies based on the affected pathways. In this aspect, delta-aminolevulinate dehydratase (*ALAD)* from family 2 is interesting, as it is

TABLE 4 | Variants corresponding to genes intersecting with the list of 2,383 high-risk HL genes from a study by Rotunno et al. (2016).

HL variants from Rotunno et al. (2016)			Variant in matched gene from present study					
Gene symbol	**Variant position**	**IDS**	**HL family**	**Variant position**	**CADD_PHRED score**	**Variant type**	**Variant consequence**	**Protein effect**
ABHD16A	6_31670740_A_T		3	6_31671105_G_A	13.23	SNVs	UTR5	–
C6orf62	6_24719009_T_C	rs147402940	3	6_24705773_T_C	12.31	SNVs	UTR3	–
CEP120	5_122758609_G_T	rs141808885	1	5_122681069_C_T	12.31	SNVs	UTR3	–
EDEM1	3_5257909_A_G	rs139745426	3	3_5246773_C_T	34	SNVs	Non-synonymous SNV	p.T160M
EIF4G3	1_21494519_T_C		3	1_21267855_C_T	14.54	SNVs	stopgain SNV	p.W7X
EPB41L2	6_131202023_A_G		3	6_131191103_G_A	22.1	SNVs	Non-synonymous SNV	p.S736F
EXTL3	8_28609630_G_A	rs191528081	3	8_28575243_G_A	23	SNVs	Non-synonymous SNV	p.R172H
FAM35A	10_88917757_A_G	rs371636091	1	10_88911115_AGT_A	25.8	Indel	Frameshift deletion	p.2_2del
FUK	16_70507153_G_A		1	16_70501193_C_T	10.08	SNVs	UTR3	-
HLTF	3_148757909_A_G	rs61750365	1	3_148802664_C_T	37	SNVs	stopgain SNV	p.W11X
LPP	3_188464149_C_T		3	3_188608373_A_T	10.5	SNVs	UTR3	–
LPP	3_188464149_C_T		3	3_188123978_G_T	40	SNVs	stopgain SNV	p.E24X
LPP			3	3_188123979_A_T	32	SNVs	Non-synonymous SNV	p.E24V
LRP6	12_12419973_G_T		3	12_12272924_AATA TATATAT ATATATATATATATA TATATATA TATAT_A	12.55	Indel	UTR3	–
LTBP2	14_74983553_G_A	rs145851939	3	14_74970734_C_T	27.7	SNVs	Non-synonymous SNV	p.G1493R
MAPKAP1	9_128199718_AT_		3	9_128199770_TAA_T	14.3	Indel	UTR3	–
MARCH10	17_60827878_G_A	rs112201730	3	17_60885673_G_A	12.03	SNVs	UTR5	–
MYO18A	17_27448659_C_T	rs371862120	3	17_27441099_G_A	24.3	SNVs	Non-synonymous SNV	p.A843V
NCAM1	11_113113556_A_G		1	11_113134920_C_A	11.07	SNVs	UTR3	–
NIPBL	5_36876673_G_A		1	5_36962227_G_A	27.5	SNVs	Non-synonymous SNV	p.R154Q
PHC2	1_33820711_G_A		3	1_33896663_C_A	15.12	SNVs	upstream	–
RCN1	11_32126524_C_T		3	11_32112681_C_T	10.66	SNVs	UTR5	–
SBF2	11_9985135_TAAT_		3	11_9838541_C_T	33	SNVs	Non-synonymous SNV	p.R1275H
SLMAP	3_57914019_A_C	rs191613999	1	3_57742023_C_G	13.51	SNVs	UTR5	–
SZT2	1_43885320_C_T		3	1_43891311_G_A	31	SNVs	Non-synonymous SNV	p.A96T

Variant details from both databases the present study and the study by Rotunno et al. (2016) are shown.

involved in the catalysis of the second step in the biosynthesis of heme and also acts as an endogenous inhibitor of the 26 S proteasome, a multi-catalytic ATP-dependent protease complex that functions as the degrading arm of the ubiquitin system, which is the major pathway for regulated degradation of proteins in all eukaryotes. Down regulation of *ALAD* is shown to be associated with poor prognosis in patients with breast cancer (Ge et al., 2017) whereas the existing data on non-erythroid spectrin αII (*SPTAN1*) suggest that overexpression of *SPTAN1* in tumor cells reflects neoplastic and tumor promoting activity or tumor suppressing effects by enabling DNA repair through interaction with DNA repair proteins (Ackermann and Brieger, 2019). *CERCAM* is known as an unfavorable prognostic marker in urothelial, renal, and ovarian cancers implying the importance of the variants in these genes (Ma et al., 2016). RAD51D from family III is particularly interesting since it is involved in DNA repair through homologous recombination. Therefore, it is possible that carcinomas arising in patients carrying mutations in this gene will be sensitive to chemotherapeutic agents that target this pathway, such as cisplatin and the PARP (poly (ADP-ribose) polymerase) inhibitor olaparib. This has already been demonstrated in *BRCA1/2* mutation-carrier cancer patients

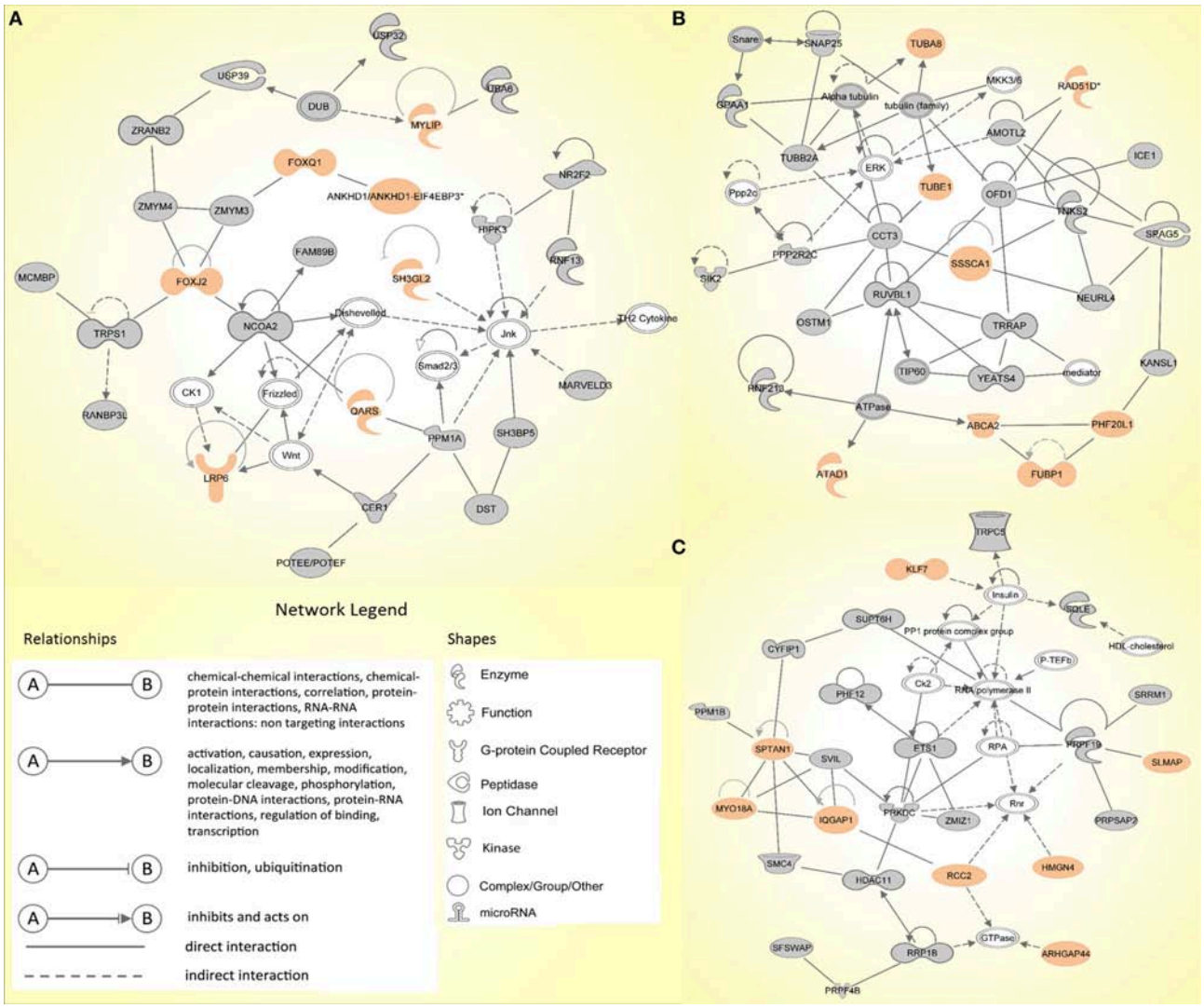


FIGURE 2 | The top three molecular networks identified by Ingenuity Pathway Analysis: **(A)** Network 1. Cancer, hematological disease, immunological disease; **(B)** Network 2. developmental disorder, endocrine system disorders, hereditary disorder; **(C)** Network 3. RNA post-transcriptional modification, cell death and survival, cellular movement. Genes from our input-data are shown in gray, genes from our prioritized candidate list are highlighted in peach.

(Banerjee et al., 2010; Loveday et al., 2011). This approach can also be applied to target pathways affected by the mutated genes. Several candidate genes were identified by IPA pathway analysis in B cell receptor pathways, offering a valuable target for other pharmaceutical drugs. The B cell receptor (BCR) signaling pathway, when dysregulated, is a potent contributor to lympomagenesis and tumor survival (Valla et al., 2018). This pathway has been targeted in B-cell lymphomas and leukemias with several BCR-directed agents, such as inhibitors of Bruton's tyrosine kinase (BTK9), spleen tyrosine kinase (SYK) and phosphatidylinositol-3-kinase (PI3K) (Buggy and Elias, 2012; Dreyling et al., 2017; Liu and Mamorska-Dyga, 2017). In one study, excellent response rates could be demonstrated in certain non-Hodgkin lymphoma subtypes, however, issues related to the development of resistance to BTK inhibitors need to be addressed (Valla et al., 2018).

Advancements in the field of genomics have allowed WGS to become the state-of-the-art tool for the identification of novel cancer predisposing genes in Mendelian diseases. It is still a challenge to appropriately interpret the immense amount of data generated by WGS, especially with respect to non-coding variants. In our study, we have attempted to interpret a selection of non-coding variants using *in silico* and bioinformatic tools, however, the adequate analysis of intronic and intergenic variants remains a challenge. There are several reports of WGS being successfully implemented to implicate rare, high-penetrance germline variants in cancer, for example *POT1* mutations in familial melanoma and Hodgkin lymphoma

(Mcmaster et al., 2018; Wong et al., 2019) and *POLE* and *POLD1* mutations in colorectal adenomas or carcinomas (Palles et al., 2013). In a previous study, we have used our pipeline (FCVPPv2) to prioritize novel variants in non-medullary thyroid cancer prone families (Srivastava et al., 2019). We have also successfully combined our pipeline with literature review and functional studies to identify *DICER1* as a candidate predisposing gene in one Hodgkin lymphoma family (Bandapalli et al., 2018). We aim to apply these methods in the remaining Hodgkin lymphoma families and hope that these results will facilitate personalized therapy in the studied families and contribute to the screening of other individuals at risk of developing HL.

DATA AVAILABILITY STATEMENT

Unfortunately, for reasons of ethics and patient confidentiality we are not able to provide the sequencing data into a public data base. The data underlying the results presented in the study are available from the corresponding author or from Dr. Asta Försti (Email: a.foersti@kitz-heidelberg.de).

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Ethics Committee of the University of Heidelberg, Germany & Ethics Committee of the Pomeranian Medical University, Poland. Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin.

AUTHOR CONTRIBUTIONS

OB, AF, and KH conceived and designed the study. WB, MW-H, DD, and JL provided the HL family samples. NP ran WGS pipeline and CNVs analysis. AS, OB, SG, and AK analyzed the data. OB and SG performed the experiments. AS and OB wrote the first draft of the manuscript. All authors read, commented on, and approved the manuscript.

FUNDING

This study was supported by the Harald Huppert Foundation and Transcan ERA-NET funding from the German Federal Ministry of Education and Research (BMBF).

ACKNOWLEDGMENTS

The authors thank the Genomics and Proteomics Core Facility (GPCF) of the German Cancer Research Center (DKFZ) for providing excellent library preparation and sequencing services and the Omics IT and Data Management Core Facility (ODCF) of the DKFZ for the whole genome sequencing data management.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fbioe.2020.00179/full#supplementary-material

REFERENCES

Ackermann, A., and Brieger, A. (2019). The role of nonerythroid spectrin αII in cancer. *J. Oncol.* 2019:7079604. doi: 10.1155/2019/7079604

Agarwal, V., Bell, G. W., Nam, J. W., and Bartel, D. P. (2015). Predicting effective microRNA target sites in mammalian mRNAs. *Elife* 4, 1–14. doi: 10.7554/eLife.05005.028

Bandapalli, O. R., Paramasivam, N., Giangiobbe, S., Kumar, A., Benisch, W., Engert, A., et al. (2018). Whole genome sequencing reveals DICER1 as a candidate predisposing gene in familial Hodgkin lymphoma. *Int J Cancer* 143, 2076–2078. doi: 10.1002/ijc.31576

Banerjee, S., Kaye, S. B., and Ashworth, A. (2010). Making the best of PARP inhibitors in ovarian cancer. *Nat. Rev. Clin. Oncol.* 7, 508–519. doi: 10.1038/nrclinonc.2010.116

Betel, D., Koppal, A., Agius, P., Sander, C., and Leslie, C. (2010). Comprehensive modeling of microRNA targets predicts functional non-conserved and non-canonical sites. *Genome Biol.* 11:R90. doi: 10.1186/gb-2010-11-8-r90

Birney, E., Stamatoyannopoulos, J. A., Dutta, A., Guigó, R., Gingeras, T. R., Margulies, E. H., et al. (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* 447, 799–816. doi: 10.1038/nature05874

Boyle, A. P., Hong, E. L., Hariharan, M., Cheng, Y., Schaub, M. A., Kasowski, M., et al. (2012). Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res.* 22, 1790–1797. doi: 10.1101/gr.137323.112

Buggy, J. J., and Elias, L. (2012). Bruton tyrosine kinase (BTK) and its role in B-cell malignancy. *Int. Rev. Immunol.* 31, 119–132. doi: 10.3109/08830185.2012.664797

Chen, X., Li, Y., Ouyang, T., Li, J., Wang, T., Fan, Z., et al. (2018). Associations between RAD51D germline mutations and breast cancer risk and survival in BRCA1/2-negative breast cancers. *Ann. Oncol.* 29, 2046–2051. doi: 10.1093/annonc/mdy338

Cooper, G. M., Stone, E. A., Asimenos, G., Program, N. C. S., Green, E. D., Batzoglou, S., et al. (2005). Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res.* 15, 901–913. doi: 10.1101/gr.3577405

Cozen, W., Timofeeva, M. N., Li, D., Diepstra, A., Hazelett, D., Delahaye-Sourdeix, M., et al. (2014). A meta-analysis of Hodgkin lymphoma reveals 19p13.3 TCF3 as a novel susceptibility locus. *Nat. Commun.* 5:3856. doi: 10.1038/ncomms4856

Cybulski, C., Wokolorczyk, D., Huzarski, T., Byrski, T., Gronwald, J., Gorski, B., et al. (2006). A large germline deletion in the Chek2 kinase gene is associated with an increased risk of prostate cancer. *J. Med. Genet.* 43, 863–866. doi: 10.1136/jmg.2006.044974

Dayem Ullah, A. Z., Oscanoa, J., Wang, J., Nagano, A., Lemoine, N. R., and Chelala, C. (2018). SNPnexus: assessing the functional relevance of genetic variation to facilitate the promise of precision medicine. *Nucleic Acids Res.* 46, W109–W113. doi: 10.1093/nar/gky399

Diehl, V., Thomas, R. K., and Re, D. (2004). Part II: Hodgkin's lymphoma–diagnosis and treatment. *Lancet Oncol.* 5, 19–26. doi: 10.1016/S1470-2045(03)01320-2

Diepstra, A., Niens, M., Vellenga, E., Van Imhoff, G. W., Nolte, I. M., Schaapveld, M., et al. (2005). Association with HLA class I in Epstein-Barr-virus-positive and with HLA class III in Epstein-Barr-virus-negative Hodgkin's lymphoma. *Lancet* 365, 2216–2224. doi: 10.1016/S0140-6736(05)66780-3

Dreyling, M., Santoro, A., Mollica, L., Leppa, S., Follows, G. A., Lenz, G., et al. (2017). Phosphatidylinositol 3-kinase inhibition by copanlisib in relapsed or refractory indolent lymphoma. *J. Clin. Oncol.* 35, 3898–3905. doi: 10.1200/JCO.2017.75.4648

Frampton, M., Da Silva Filho, M. I., Broderick, P., Thomsen, H., Forsti, A., Vijayakrishnan, J., et al. (2013). Variation at 3p24.1 and 6q23.3 influences the risk of Hodgkin's lymphoma. *Nat. Commun.* 4:2549. doi: 10.1038/ncomms3549

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 in patients with breast cancer. *Cancer Sci.* 108, 604–611. doi: 10.1111/cas.13180

Kharazmi, E., Fallah, M., Pukkala, E., Olsen, J. H., Tryggvadottir, L., Sundquist, K., et al. (2015). Risk of familial classical Hodgkin lymphoma by relationship, histology, age, and sex: a joint study from five Nordic countries. *Blood* 126, 1990–1995. doi: 10.1182/blood-2015-04-639781

Kircher, M., Witten, D. M., Jain, P., O'roak, B. J., Cooper, G. M., and Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* 46, 310–315. doi: 10.1038/ng.2892

Kopanos, C., Tsiolkas, V., Kouris, A., Chapple, C. E., Albarca Aguilera, M., Meyer, R., et al. (2018). VarSome: the human genomic variant search engine. *Bioinformatics* 35, 1978–1980. doi: 10.1093/bioinformatics/bty897

Kumar, A., Bandapalli, O. R., Paramasivam, N., Giangiobbe, S., Diquigiovanni, C., Bonora, E., et al. (2018). Familial cancer variant prioritization pipeline version 2 (FCVPPv2) applied to a papillary thyroid cancer family. *Sci. Rep.* 8:11635. doi: 10.1038/s41598-018-29952-z

Kuppers, R. (2009). The biology of Hodgkin's lymphoma. *Nat. Rev. Cancer* 9, 15–27. doi: 10.1038/nrc2542

Kushekhar, K., Van Den Berg, A., Nolte, I., Hepkema, B., Visser, L., and Diepstra, A. (2014). Genetic associations in classical hodgkin lymphoma: a systematic review and insights into susceptibility mechanisms. *Cancer Epidemiol. Biomarkers Prev.* 23:2737–2747. doi: 10.1158/1055-9965.EPI-14-0683

Lek, M., Karczewski, K. J., Minikel, E. V., Samocha, K. E., Banks, E., Fennell, T., et al. (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature* 536, 285–291. doi: 10.1038/nature19057

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* 27, 2987–2993. doi: 10.1093/bioinformatics/btr509

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25, 1754–1760. doi: 10.1093/bioinformatics/btp324

Liu, D., and Mamorska-Dyga, A. (2017). Syk inhibitors in clinical development for hematological malignancies. *J. Hematol. Oncol.* 10:145. doi: 10.1186/s13045-017-0512-1

Liu, X., Wu, C., Li, C., and Boerwinkle, E. (2016). dbNSFP v3.0: a one-stop database of functional predictions and annotations for human nonsynonymous and splice-site SNVs. *Hum. Mutat.* 37, 235–241. doi: 10.1002/humu.22932

Loveday, C., Turnbull, C., Ramsay, E., Hughes, D., Ruark, E., Frankum, J. R., et al. (2011). Germline mutations in RAD51D confer susceptibility to ovarian cancer. *Nat. Genet.* 43, 879–882. doi: 10.1038/ng.893

Ma, L. J., Wu, W. J., Wang, Y. H., Wu, T. F., Liang, P. I., Chang, I. W., et al. (2016). SPOCK1 overexpression confers a poor prognosis in urothelial carcinoma. *J. Cancer* 7, 467–476. doi: 10.7150/jca.13625

Mcmaster, M. L., Sun, C., Landi, M. T., Savage, S. A., Rotunno, M., Yang, X. R., et al. (2018). Germline mutations in protection of telomeres 1 in two families with Hodgkin lymphoma. *Br. J. Haematol.* 181, 372–377. doi: 10.1111/bjh.15203

Ngan, E., Kiepas, A., Brown, C. M., and Siegel, P. M. (2018). Emerging roles for LPP in metastatic cancer progression. *J. Cell Commun. Signal.* 12, 143–156. doi: 10.1007/s12079-017-0415-5

Palles, C., Cazier, J. B., Howarth, K. M., Domingo, E., Jones, A. M., Broderick, P., et al. (2013). Germline mutations affecting the proofreading domains of POLE and POLD1 predispose to colorectal adenomas and carcinomas. *Nat. Genet.* 45, 136–144. doi: 10.1038/ng.2503

Petrovski, S., Wang, Q., Heinzen, E. L., Allen, A. S., and Goldstein, D. B. (2013). Genic intolerance to functional variation and the interpretation of personal genomes. *PLoS Genet.* 9:e1003709. doi: 10.1371/annotation/32c8d343-9e1d-46c6-bfd4-b0cd3fb7a97e

Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R., and Siepel, A. (2010). Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.* 20, 110–121. doi: 10.1101/gr.097857.109

Rimmer, A., Phan, H., Mathieson, I., Iqbal, Z., Twigg, S. R. F., Consortium, W. G. S., et al. (2014). Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nat. Genet.* 46, 912–918. doi: 10.1038/ng.3036

Ristolainen, H., Kilpivaara, O., Kamper, P., Taskinen, M., Saarinen, S., Leppa, S., et al. (2015). Identification of homozygous deletion in ACAN and other candidate variants in familial classical Hodgkin lymphoma by exome sequencing. *Br. J. Haematol.* 170, 428–431. doi: 10.1111/bjh.13295

Rotunno, M., Mcmaster, M. L., Boland, J., Bass, S., Zhang, X., Burdett, L., et al. (2016). Whole exome sequencing in families at high risk for Hodgkin lymphoma: identification of a predisposing mutation in the KDR gene. *Haematologica* 101, 853–860. doi: 10.3324/haematol.2015.135475

Saarinen, S., Aavikko, M., Aittomaki, K., Launonen, V., Lehtonen, R., Franssila, K., et al. (2011). Exome sequencing reveals germline NPAT mutation as a candidate risk factor for Hodgkin lymphoma. *Blood* 118, 493–498. doi: 10.1182/blood-2011-03-341560

Salipante, S. J., Mealiffe, M. E., Wechsler, J., Krem, M. M., Liu, Y., Namkoong, S., et al. (2009). Mutations in a gene encoding a midbody kelch protein in familial and sporadic classical Hodgkin lymphoma lead to binucleated cells. *Proc. Natl. Acad. Sci. U.S.A.* 106, 14920–14925. doi: 10.1073/pnas.0904231106

Siepel, A., Bejerano, G., Pedersen, J. S., Hinrichs, A. S., Hou, M., Rosenbloom, K., et al. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* 15, 1034–1050. doi: 10.1101/gr.3715005

Smigielski, E. M., Sirotkin, K., Ward, M., and Sherry, S. T. (2000). dbSNP: a database of single nucleotide polymorphisms. *Nucleic Acids Res.* 28, 352–355. doi: 10.1093/nar/28.1.352

Srivastava, A., Kumar, A., Giangiobbe, S., Bonora, E., Hemminki, K., Forsti, A., et al. (2019). Whole genome sequencing of familial non-medullary thyroid cancer identifies germline alterations in MAPK/ERK and PI3K/AKT signaling pathways. *Biomolecules* 9:E605. doi: 10.3390/biom9100605

Tan, D. E. K., Foo, J. N., Bei, J.-X., Chang, J., Peng, R., Zheng, X., et al. (2013). Genome-wide association study of B cell non-Hodgkin lymphoma identifies 3q27 as a susceptibility locus in the Chinese population. *Nat. Genet.* 45:804. doi: 10.1038/ng.2666

The Genomes Project Consortium, Auton, A., Abecasis, G. R., Altshuler, D. M., Durbin, R. M., Abecasis, G. R., et al. (2015). A global reference for human genetic variation. *Nature* 526, 68–74. doi: 10.1038/nature15393

Valla, K., Flowers, C. R., and Koff, J. L. (2018). Targeting the B cell receptor pathway in non-Hodgkin lymphoma. *Expert Opin. Investig. Drugs* 27, 513–522. doi: 10.1080/13543784.2018.1482273

Wang, K., Li, M., and Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* 38:e164. doi: 10.1093/nar/gkq603

Ward, L. D., and Kellis, M. (2012). HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res.* 40, D930–D934. doi: 10.1093/nar/gkr917

Wong, K., Robles-Espinoza, C. D., Rodriguez, D., Rudat, S. S., Puig, S., Potrony, M., et al. (2019). Association of the POT1 germline missense variant p.I78T with familial melanoma. *JAMA Dermatol.* 155, 604–609. doi: 10.1001/jamadermatol.2018.3662

Zhang, J., Walsh, M. F., Wu, G., Edmonson, M. N., Gruber, T. A., Easton, J., et al. (2015). Germline mutations in predisposition genes in pediatric cancer. *N. Engl. J. Med.* 373, 2336–2346. doi: 10.1056/NEJMoa1508054

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 Bioengineering and Biotechnology

ORIGINAL RESEARCH
published: 06 March 2020
doi: 10.3389/fbioe.2020.00074

A Computational Approach for Mapping Heme Biology in the Context of Hemolytic Disorders

Farah Humayun[1,2†], Daniel Domingo-Fernández[2†], Ajay Abisheck Paul George[1], Marie-Thérèse Hopp[1], Benjamin F. Syllwasschy[1], Milena S. Detzel[1], Charles Tapley Hoyt[2], Martin Hofmann-Apitius[2*] and Diana Imhof[1*]*

[1] Pharmaceutical Biochemistry and Bioanalytics, Pharmaceutical Institute, University of Bonn, Bonn, Germany, [2] Department of Bioinformatics, Fraunhofer Institute for Algorithms and Scientific Computing (SCAI), Sankt Augustin, Germany

Heme is an iron ion-containing molecule found within hemoproteins such as hemoglobin and cytochromes that participates in diverse biological processes. Although excessive heme has been implicated in several diseases including malaria, sepsis, ischemia-reperfusion, and disseminated intravascular coagulation, little is known about its regulatory and signaling functions. Furthermore, the limited understanding of heme's role in regulatory and signaling functions is in part due to the lack of curated pathway resources for heme cell biology. Here, we present two resources aimed to exploit this unexplored information to model heme biology. The first resource is a terminology covering heme-specific terms not yet included in standard controlled vocabularies. Using this terminology, we curated and modeled the second resource, a mechanistic knowledge graph representing the heme's interactome based on a corpus of 46 scientific articles. Finally, we demonstrated the utility of these resources by investigating the role of heme in the Toll-like receptor signaling pathway. Our analysis proposed a series of crosstalk events that could explain the role of heme in activating the TLR4 signaling pathway. In summary, the presented work opens the door to the scientific community for exploring the published knowledge on heme biology.

Keywords: heme, hemolytic disorders, signaling pathways, knowledge graphs, biological expression language

OPEN ACCESS

Edited by:
Jyoti Sharma,
Institute of Bioinformatics (IOB), India

Reviewed by:
Yu Lin,
Center for Devices and Radiological Health (CDRH), United States
Obul Reddy Bandapalli,
Hopp Children's Cancer Center Heidelberg (KiTZ), Germany

****Correspondence:***
Daniel Domingo-Fernández
daniel.domingo.fernandez@scai.fraunhofer.de;
danieldomingofernandez@hotmail.com
Martin Hofmann-Apitius
martin.hofmann-apitius@scai.fraunhofer.de
Diana Imhof
dimhof@uni-bonn.de

† These authors have contributed equally to this work

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Bioengineering and Biotechnology

Received: *14 October 2019*
Accepted: *28 January 2020*
Published: *06 March 2020*

Citation:
Humayun F, Domingo-Fernández D, Paul George AA, Hopp M-T, Syllwasschy BF, Detzel MS, Hoyt CT, Hofmann-Apitius M and Imhof D (2020) A Computational Approach for Mapping Heme Biology in the Context of Hemolytic Disorders. Front. Bioeng. Biotechnol. 8:74. doi: 10.3389/fbioe.2020.00074

INTRODUCTION

Heme is an iron ion-coordinating porphyrin derivative essential to aerobic organisms (Zhang, 2011). It plays a crucial role as a prosthetic group in hemoproteins involved in several biological processes such as electron transport, oxygen transfer, and catalysis (Smith and Warren, 2009; Zhang, 2011; Kühl and Imhof, 2014; Poulos, 2014). Besides its indispensable role in hemoproteins, it can act as a damage-associated molecular pattern leading to oxidative injury, inflammation, and consequently, organ dysfunction (Jeney, 2002; Wagener et al., 2003; Dutra and Bozza, 2014). Plasma scavengers such as haptoglobin and hemopexin bind hemoglobin and heme, respectively, thus keeping the concentration of labile heme at low concentrations (Smith and McCulloh, 2015). However, at high concentrations of hemoglobin and, consequently heme, these scavenging proteins get saturated, resulting in the accumulation of biologically available heme (Soares and Bozza, 2016). With respect to hemolytic diseases, the formation of labile heme at harmful concentrations has been a subject of research for some years now (Roumenina et al., 2016; Soares and Bozza, 2016; Gouveia et al., 2017).

Biomedical literature is an immense source of heterogeneous data that are dispersed throughout hundreds of journals. Furthermore, the majority of the results are scattered and published as unstructured free-text, or at best, presented in tables and cartoons representing the experimental study or biological processes and pathways. These shortcomings, combined with the exponential growth of biomedical literature, prevent the healthcare community and individual researchers from being aware of all the available information and knowledge in the literature. With the introduction of new technologies and experimental techniques, researchers have made significant advances in heme-related research and its role in the pathogenesis of numerous hemolytic diseases such as sepsis (Larsen et al., 2010; Effenberger-Neidnicht and Hartmann, 2018), malaria (Ferreira et al., 2008; Dey et al., 2012), and β-thalassemia (Vinchi et al., 2013; Conran, 2014; Garcia-Santos et al., 2017). In these diseases, large amounts of heme are released from ruptured erythrocytes and can potentially wreak havoc (Tolosano et al., 2010). Thus, it is crucial to develop new strategies that capture and exploit the vast amount of literature knowledge surrounding heme to better understand its mechanistic role in hemolytic disorders.

Biological knowledge formalized as a network can be used by clinicians as research and information retrieval tools, by biologists to propose *in vitro* and *in vivo* experiments, and by bioinformaticians to analyze high throughput *-omics* experiments (Catlett et al., 2013; Ali et al., 2019). Further, they can be readily semantically integrated with databases and other systems biology resources to improve their ability to accomplish each of these tasks (Hoyt et al., 2018). However, enabling this semantic integration requires organizing and formalizing the knowledge using specific vocabularies and ontologies. Although this endeavor involves significant curation efforts, it is key to the success of the subsequent modeling steps. Therefore, in practice, knowledge-based disease modeling approaches have been conducted only for major disorders such as cancer (Kuperstein et al., 2015) or neurodegenerative disorders (Mizuno et al., 2012; Fujita et al., 2014). In summary, while the scarcity of mechanistic information and the necessary amount of curation often impede launching the aforementioned approaches, modeling and mining literature knowledge provide a holistic picture of the field of interest. Furthermore, the underlying models derived from such approaches have a broad range of applications including hypothesis generation, predictive modeling and drug discovery.

Here, we present two resources aimed at assembling mechanistic knowledge surrounding the metabolism, biological functions, and pathology of heme in the context of selected hemolytic disorders. The first resource is a terminology formalizing heme-specific terms that have until now not been covered by other standard controlled vocabularies. The second resource is a heme knowledge graph (HemeKG), that is, a network comprising more than 700 nodes and more than 3,000 interactions. It was generated from 46 selected articles as the first attempt of modeling the knowledge, which is available from more than 20,000 heme-related publications. Finally, we demonstrate both resources by analyzing the crosstalk between heme biology and the TLR4 signaling pathway. The results of this analysis suggest that the activation profile for labile heme as an extracellular signaling molecule through TLR4 induces cytokine and chemokine production. However, the underlying molecular mechanism and individual pathway effectors are not fully understood and need further exploration.

MATERIALS AND METHODS

This section describes the methodology used to generate the mechanistic knowledge graph and its supporting terminology. Subsequently, it outlines the approach followed to conduct the pathway crosstalk analysis. A schematic diagram of the methodology is presented in **Figure 1**.

Knowledge Modeling

In order to identify recently published articles (i.e., published in the last 10 years) describing the role of heme in hemolytic disorders, PubMed was queried with the following: ("heme" AND "hemolysis") OR ("heme" AND "thrombosis") OR ("heme" AND "inflammation") AND ("2009"[Date – Publication]: "3000"[Date – Publication]). The resulting 3,108 articles were manually filtered by removing articles that were deemed too general or lacked a biochemical focus, as judged by expert opinion. After this filtering step, 6 reviews and 40 original research articles were selected for knowledge extraction and modeling. Knowledge was manually extracted and curated from this selected corpus using the official Biological Expression Language (BEL) curation guidelines from http://openbel.org/language/version_2.0/bel_specification_version_2.0.html and http://language.bel.bio as well as additional guidelines from https://github.com/pharmacome/curation.

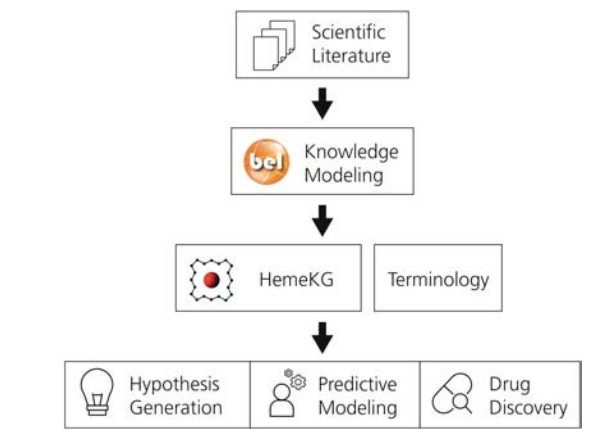


FIGURE 1 | The workflow used to generate the supporting terminology and HemeKG. The first step involves the selection of relevant scientific literature. Next, evidence from this selected corpus is extracted and translated into BEL to generate a computable knowledge assembly model, HemeKG. In parallel to the modeling task, a terminology to support knowledge extraction of articles about the heme molecule was built. Finally, HemeKG can be used for numerous tasks such as hypothesis generation, predictive modeling and drug discovery.

Evidence from the selected corpus was manually translated into BEL statements together with their contextual information (e.g., cell type, tissue and dosage information). For instance, the evidence "Heme/iron-mediated oxidative modification of LDL can cause endothelial cytotoxicity and – at sublethal doses – the expression of stress-response genes" (Nagy et al., 2010) corresponds to the following BEL statement:

SET Cell = "endothelial cell"
a(CHEBI:"oxidised LDL") pos bp(MESH:"Cytotoxicity, Immunologic").

Generation of a Supporting Terminology

During curation, a terminology was generated to support the standardization of domain-specific terminology encountered during the curation of articles related to the heme molecule. The aim of the terminology is to catalog and harmonize terms not present in other controlled vocabularies such as ChEBI (Degtyarenko et al., 2007) for chemicals, or Gene Ontology [GO; (Ashburner et al., 2000)] and Medical Subject Headings [MeSH; (Rogers, 1963)] for pathologies. Thus, each term was checked by two experts in the field assisted by the Ontology Lookup Service [OLS; (Cote et al., 2010)] to avoid duplicates with other terminologies or ontologies. Furthermore, we required that each entry included the following metadata: an identifier, a label, a definition, an example of usage in a sentence, and references to articles in which it was described. Furthermore, a list of synonyms was also curated in a separate file to facilitate the use of the terminology in annotation or text mining tasks. The supporting terminology is included in the **Supplementary Material** and can also be found at https://github.com/hemekg/terminology.

Analyzing Pathway Crosstalk Between Heme and the Toll-Like Receptor Signaling Pathway

Crosstalk analysis aims to study how two or more pathways communicate or influence each other. While there exist, numerous methodologies designed to investigate pathway crosstalk, the majority of these approaches exclusively quantify such crosstalk based on the overlap between a pair of pathways without delving into the nature of the crosstalk (Donato et al., 2013). In this section, we demonstrate how combining knowledge from HemeKG with a canonical pathway reveals mechanistic insights on the crosstalk between two different pathways.

Because of the amount of effort required to manually analyze crosstalk across multiple pathways, we conducted a pathway enrichment analysis on three pathway databases [i.e., KEGG Kanehisa et al., 2016), Reactome (Fabregat et al., 2017), WikiPathways (Slenter et al., 2017)] to identify pathways enriched with the gene set extracted from the entire Heme knowledge map. The enrichment analysis evaluated the overrepresentation of the genes present in HemeKG for each of the pathways in the three aforementioned databases using Fisher's exact test (Fisher, 1992). Furthermore, Benjamini–Yekutieli method under dependency was applied to correct for multiple testing (Yekutieli and Benjamini, 2001). Manual inspection of the enrichment analysis results revealed that the Toll-like receptor (TLR) signaling pathway was the most enriched pathway in Reactome and WikiPathways, and the third most enriched in KEGG (**Supplementary Table S1**). Therefore, this pathway was selected for study in the subsequent investigation.

First, the three different representations of this pathway were downloaded from each database and converted to BEL using PathMe (Domingo-Fernández et al., 2019). Next, the three BEL networks were combined with the HemeKG network highlighting their overlaps (**Supplementary Figures S1**, **S2**) in order to specifically analyze these parts of the combined network. Finally, five experts in the field reconstructed the hypothesized pathways from the combined network. The hypothesized pathways were depicted following the guidelines for scientific communication of biological networks outlined by Marai et al. (2019).

RESULTS

Building a Mechanistic Knowledge Graph Around Heme Biology in the Context of Hemolytic Disorders

We introduce the first knowledge graph made publicly available to the biomedical and bioinformatics community focused on heme biology (**Figure 2**). The presented heme knowledge graph was based on the selection of 40 original research articles and 6 review articles related to heme and its role in several pathways. These pathways include the tumor necrosis factor (TNF) and nuclear factor κ-light-chain-enhancer of activated B cells (NF-κB) signaling pathways, and the complement and coagulation cascades, through which heme plays a role in hemolysis, inflammation and thrombosis (Dutra and Bozza, 2014; L'Acqua and Hod, 2014; Roumenina et al., 2016; Martins and Knapp, 2018; Vogel and Thein, 2018). The focus of the review articles was chosen because of the relevance of these diseases and complications to large numbers of patients (L'Acqua and Hod, 2014; Litvinov and Weisel, 2016; Roumenina et al., 2016; Effenberger-Neidnicht and Hartmann, 2018). All of these pathologies are known to be interconnected and mapping them in relation to heme is promising for the discovery of yet overlooked links.

Following the guidelines outlined in the Methods section, knowledge was manually extracted and encoded from each of these 46 articles using BEL because of its ability to represent not only causal, but also correlative and associative relationships found in the literature, as well as corresponding provenance and experimental contextual information. This curation exercise resulted in HemeKG, a knowledge graph containing 775 nodes (**Table 1**) and 3,051 relations (**Table 2**), as well as contextual information ranging from cellular and anatomical localization to different states of the heme molecule (**Supplementary Figure S1**). Annotations, such as time point and concentration, enabled us to capture time dependencies between entities. By using this contextual information and the multiple biological scales presented in the model, we have not only been able to represent a part of heme's interactome (**Figure 2**), but also established

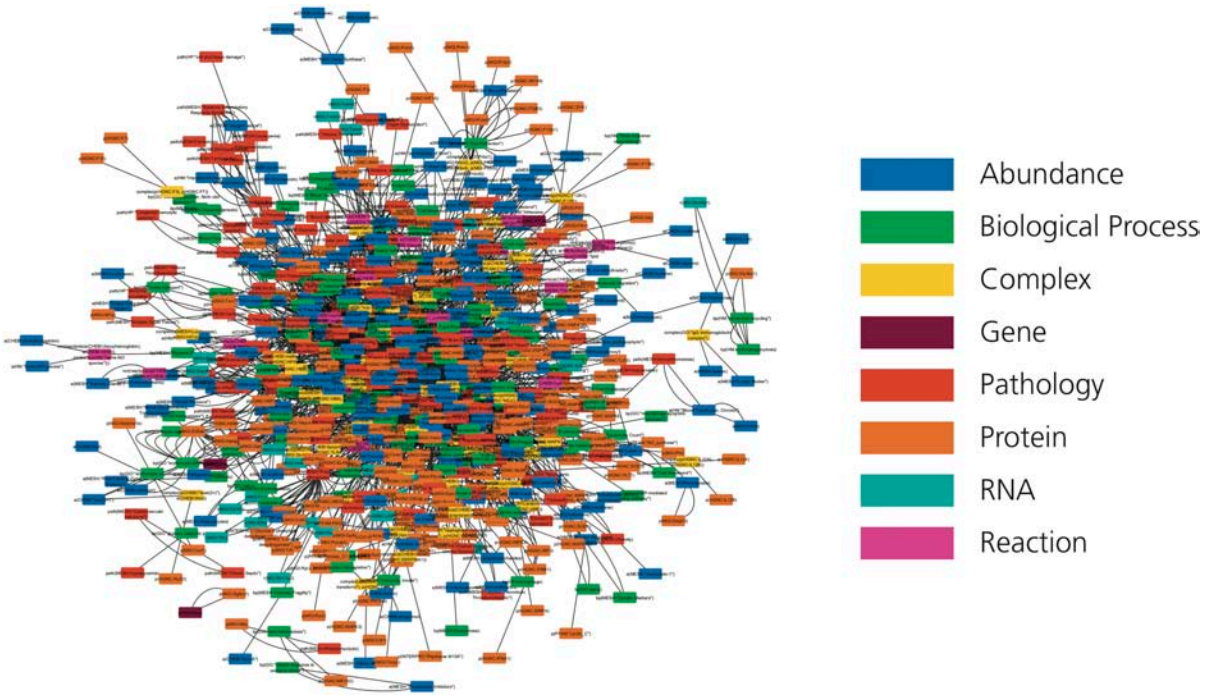


FIGURE 2 | The HemeKG network. Nodes are colored by their different functions in BEL (see legend).

several links to phenotypes and clinical endpoints. Both represent essential considerations for the design of future clinical studies of hemolytic conditions.

Finally, to facilitate the use of the curated content in this work, BEL documents are bundled with a dedicated Python package that enables direct access to the content, provides conversion utilities and allows for network exploration. Both the BEL documents and the Python package are available at https://github.com/hemekg/hemekg.

Curating a Supporting Heme Terminology

The specificity of our work, together with the lack of contextual terminologies related to heme biology, prompted us to generate a supporting terminology focused on heme. It contains more than 50 terms that delineate heme-related entities, such as biological processes, proteins, or pathologies that are not yet included in other standard resources such as (GO Ashburner et al., 2000). Building this terminology not only allowed us to describe entities with more expressiveness, but also facilitates text mining or annotation tasks related to the heme molecule in the future. The terminology is available at https://github.com/hemekg/terminology.

Dissection of the Crosstalk Between Heme and TLR Using HemeKG

The established heme knowledge graph can be used to study the crosstalk of heme biology with a pathway of interest. HemeKG is of special interest in the context of hemolytic disorders, such as malaria and sickle cell anemia, because these diseases are associated with the release of heme into circulation. Heme can then exert a detrimental role by regulating several proteins and signaling pathways (Kühl and Imhof, 2014). In order to select a pathway that highly overlaps with the generated network, we conducted pathway enrichment analysis using three major databases (i.e., KEGG, Kanehisa et al., 2016), Reactome (Fabregat et al., 2017), and WikiPathways (Slenter et al., 2017). The results of the enrichment analysis in the three databases pointed to TLR signaling as the most enriched pathway (**Supplementary Table S1**). Thus, we proceeded to analyze the crosstalk between this pathway and heme biology by exploring the overlap between HemeKG and the TLR pathways in the three aforementioned databases. Although heme has been linked to numerous (TLRs) including TLR2, TLR3, TLR4, TLR7, and TLR9 (Figueiredo et al., 2007; Lin et al., 2010; Dutra and Bozza, 2014; Min et al., 2017; Merle et al., 2019; Sudan et al., 2019), our analysis was prioritized on the most well-documented interaction, the one between heme and TLR4. Heme stimulates TLR4 to activate NF-κB secretion via myeloid differentiation primary response 88 (MyD88)-mediated activation of lκB (IKK) (see below). Activated IKK promotes the proteolytic degradation of NFKBIA. The phosphorylated IKK complex indirectly activates NF-κB and mitogen-activated protein kinases, such as JNK (C-Jun N-terminal kinase), ERK, and p38 leading to the secretion of TNF-α, interleukin 6 (IL6), IL1B, and keratinocyte-derived chemokine (Dutra and Bozza, 2014). This finally results in an activation of the innate immunity and the generation of proinflammatory factors, which reflects the relevance of heme in several disorders comprising inflammation and infection.

TABLE 1 | Summary of unique nodes for each entity class.

Abundances	Genes	RNAs	Proteins	Complexes	Reactions	Pathologies	Biological processes	Total
200	4	25	226	54	17	128	121	**775**

Each entity class corresponds to the terms formalized in BEL (more information at https://language.bel.bio).

TABLE 2 | Summary of relationship classes.

Increase	Decrease	Positive correlation	Negative correlation	Has component	Association	Causes No change	Ontological relations	Total
639	380	1,322	440	113	54	39	64	**3,051**

Each class corresponds to the relationships formalized in BEL (more information at https://language.bel.bio). The ontological relations class includes the following relationships: has reactant, has product, and has variant.

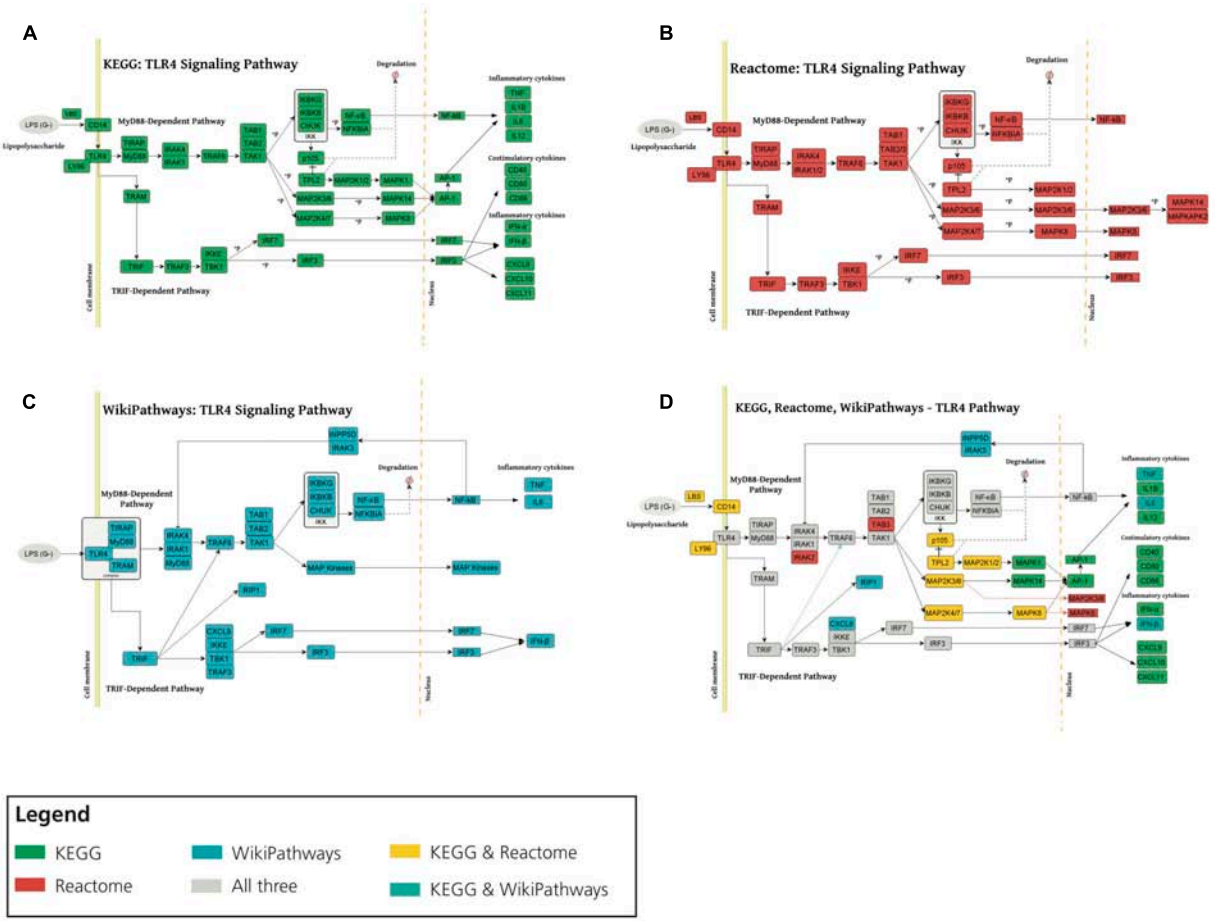


FIGURE 3 | Consensus around the TLR4 signaling pathway in three major pathway databases. TLR4 signaling pathway visualization of KEGG **(A)**, Reactome **(B)**, and WikiPathways **(C)**. **(D)** Superimposing TLR4 signaling pathway from KEGG, Reactome and WikiPathways. Each color corresponds to the presence of the given node in one or multiple databases (see Legend). MyD88, TAK1, IKK complex, MAP kinases, TNF, NF-κB, TRIF and IRF3 emerged in all three databases and also in HemeKG. KEGG and Reactome showed identical representations of the TLR4 pathway whereas WikiPathways was different in a way that nuclear NF-κB activates INPP5D-IRAK3 (inositol polyphosphate-5-phosphatase D IL1 receptor associated kinase 3) complex which inhibits the activity of IRAK1/IRAK4 (IL1 receptor associated kinase 1/4).

We first investigated the consensus of the three different representations of the TLR4 signaling pathway (**Figure 3A**). We observed that, overall, all three representations share a high degree of consensus as illustrated in **Figures 3B–D**. Here, we would like to point out that while KEGG and Reactome present practically identical representations, the WikiPathways

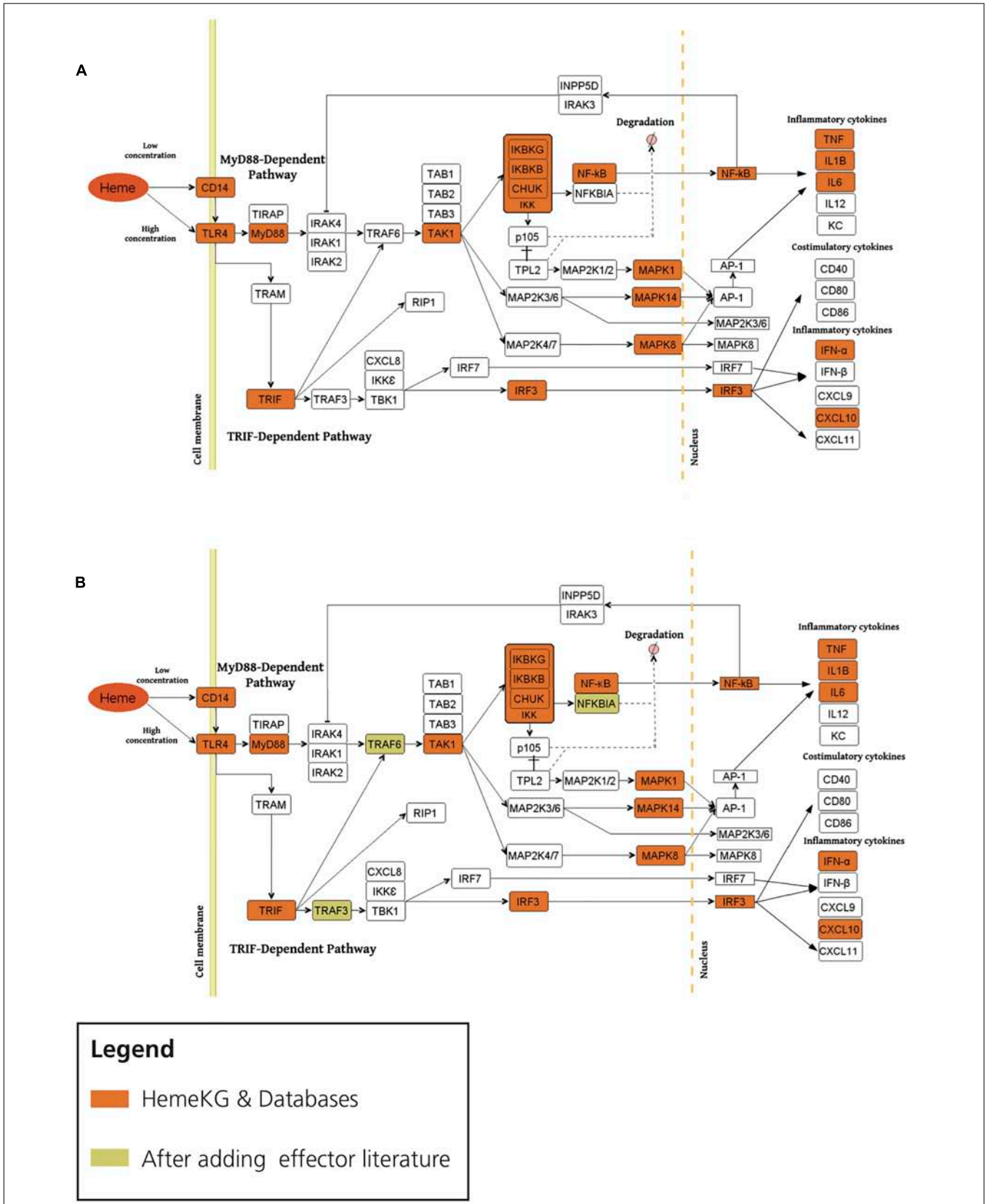


FIGURE 4 | Continued

FIGURE 4 | Overlaying the consensus TLR4 signaling pathway in databases with HemeKG (**A:** original overlaid network, **B:** overlaid network after inclusion of literature evidence for effectors). The orange colored boxes display the common effector molecules between the canonical TLR4 signaling pathway and induced TLR4 signaling pathway stimulated by labile heme. Heme/TLR4 activates the adaptor molecule MyD88. Activated MyD88 promotes the degradation of NFKBIA (NF-κB inhibitor α) through phosphorylation of the IKK complex (inhibitor of nuclear factor κB kinase complex), thus promoting NF-κB (nuclear factor κ-light-chain-enhancer of activated B cells) and MAPKs (mitogen-activated protein kinases) stimulation leading to the secretion of TNF-α, IL6, IL1B and KC (keratinocyte-derived chemokine) (Fortes et al., 2012; Dutra and Bozza, 2014). The TRIF (Toll-like receptor adaptor molecule 1) dependent pathway is activated upon signaling of heme through TLR4 leading to the activation of IRF3 (interferon regulatory factor 3) stimulating the secretion of interferons (i.e., IFN-α) and CXCL10 (C-X-C motif chemokine ligand 10) (Dickinson-Copeland et al., 2015). However, the activation profiles for IRAK1/2, TRAF6, TRAM, TRAF3, TBK1/IKK epsilon complex and IRF7 are not yet studied for heme-TLR4 signaling pathway.

representation exhibits slight differences. These differences and complementarities between pathways provide us with a more comprehensive view of the studied pathways, as illustrated by our previous work (Domingo-Fernández et al., 2019).

Second, in order to study the overlap between TLR4 signaling pathway and heme biology, we overlaid the consensus network of the pathway with HemeKG (**Figure 4**). Superimposing both networks revealed that MyD88, TAK1, IKK complex, MAP kinases, TNF, NF-κB, Toll-like receptor adaptor molecule 1 (TRIF), and interferon regulatory factor 3 (IRF3) were present in all three databases and in our model. However, several effector molecules, which were found in the three databases, were not found in our heme knowledge graph (HemeKG), for example, IL1 receptor-associated kinase proteins 1, 2, and 4 (IRAK1, IRAK2, and IRAK4, respectively); TNF receptor-associated factor 6 (TRAF6); TAB1-3; and others (**Figure 4A**). Thus, we specifically searched for literature reports of these effectors in the context of heme signaling, by entering the respective queries in PubMed, as this knowledge might not have been sufficiently covered by the 40 original research articles selected to establish HemeKG.

The activation profile for labile heme as an extracellular signaling molecule through TLR4 was suggested to be similar to the one established via Lipopolysaccharides (LPS) as signaling molecule from standard pathway databases (Pålsson-Mcdermott and O'Neill, 2004). This pathway begins with the induction of TIRAP (Mal)-associated MyD88 signaling on the one hand (Horng et al., 2002), and TRAM (TICAM-2)-associated TRIF (TICAM-1)-signaling, on the other hand (Seya et al., 2005), resulting in the upregulation of proinflammatory cytokines and chemokines (**Figure 4**). MyD88 protein as an adaptor has been shown to interact with IL1 receptor-associated kinase (IRAK) proteins 1, 2, and 4 to start the signaling cascade involving TRAF6, which is known to activate IKK in response to proinflammatory cytokines. However, in our heme knowledge graph the connections between IRAKs, TRAF6, and TAB proteins were missing (**Figure 4A**). By taking a closer look at these effectors in the context of heme, we found various information for example TRAF6 indicating both a direct and indirect link to heme-induced signaling via TLRs (Hama et al., 2012; IJssennagger et al., 2012; Park et al., 2014; Huang et al., 2015; Meng et al., 2017). In contrast, other effector molecules such as IRAK and TAB proteins (**Figure 4**) were not described in heme signaling so far. We then performed a PubMed search for these missing terms in combination with "heme." These findings led us to refine HemeKG so that only those signaling components for which no evidence was found manually still remain as white spots on the map (**Figure 4B**).

In addition, the preceding discussion has excluded parameters such as the concentration of labile heme available in the respective environment. This aspect will be particularly important, if heme-triggered signaling pathways are dependent on, or determined by the concentration of heme. At lower concentrations of heme, TLR4 signaling has been described to be CD14 dependent, whereas at high concentrations of heme, TLR4 activation does not require CD14 (Piazza et al., 2010; **Figure 4**). Also, there is a need to further investigate whether heme/TLR4 induction of the adapter molecule MyD88 is dependent or independent of TIRAP activation, similar to the LPS/TLR4 induced TIRAP-associated MyD88 signaling pathway. Furthermore, heme/TLR4 activates a pathway that leads to the activation of IRF3, resulting in the production of interferons for example, IFN-α (Dutra and Bozza, 2014) and overproduction of C-X-C motif chemokine 10 (CXCL10) (Lin et al., 2012; Dickinson-Copeland et al., 2015). In the literature, the molecular mechanism by which heme/TLR4-induced TRAF3 and IRF3/7 activation leads to the secretion of IFN-α and CXCL10 is not represented. It is therefore shown as a white box in the map (**Figure 4B**). Finally, the introduction of noncanonical pathways and receptor crosstalk-triggered cascades go beyond the scope of this work, but represent opportunities for future studies on heme signaling.

DISCUSSION

We have presented HemeKG, the first mechanistic model in the context of heme biology, as a viable solution to comprehensively summarize heme-related processes by bringing knowledge from disparate literature together. Furthermore, we have demonstrated how combining the knowledge from the heme knowledge graph with information available in pathway databases provides new insights into the network of interactions that regulate heme pathophysiology.

Because HemeKG was curated using standard vocabularies, its content can be linked to the majority of public databases. Therefore, enriching the HemeKG network with external data or incorporating its integrated knowledge into other resources is feasible. For example, the entire Bio2BEL framework[1] can be used to scale up this resource by enriching HemeKG with dozens of widely used biomedical databases. In order to make HemeKG accessible to a wider audience, we uploaded it to BEL Commons - a web application for curating, validating, and exploring knowledge assemblies encoded in BEL

[1]https://github.com/bio2bel

(Hoyt et al., 2018). Users can interactively explore the network, make modifications, integrate additional resources via Bio2BEL, and share those modifications using its versioning system. Furthermore, the variety of formats that our resource can be converted to also facilitates its use by other systems biology tools such as Cytoscape (Shannon, 2003) and NDEx (Pratt et al., 2015). In summary, the characteristics of HemeKG make this resource suitable not only for hypothesis generation as presented in our case scenario, but also for clinical decision support as previously demonstrated with other systems biology maps (Ostaszewski et al., 2018). For instance, computational mechanistic models are currently being used in combination with artificial intelligence methods for a variety of predictive applications (Khanna et al., 2018; Esteban-Medina et al., 2019; Çubuk et al., 2019). Instead of contextless canonical pathways as until now (i.e., pathways describing normal physiology), HemeKG could be used for predicting drug response and for drug repurposing in numerous related disorders such as malaria and sepsis. Finally, the supporting terminology built during this work could be used for a broad range of applications from data harmonization to natural language processing.

A potential limitation of this study is that it is constrained to a specific literature corpus as we are aware that the presented knowledge graph captures only a part of a much larger interaction network. This tends to be a common challenge when constructing contextualized maps and is further compounded by the difficulty in assessing the coverage of a network, explaining why some nodes are missing in HemeKG compared to the three pathway databases used in this study. Furthermore, the bias in the scientific community against publishing negative results must also be acknowledged. A clear example is how the hypotheses of our crosstalk analysis could be complemented by this knowledge gap that could reveal new interesting hypotheses. Thus, future updates in HemeKG, as in any work of this kind, will be required while prioritizing time and effort (Rodriguez-Esteban, 2015). Further, advanced network-based analyses (Catlett et al., 2013) could be used to rank heme-related pathways in the context of a given *-omics* data set.

Although numerous interactions between heme and TLRs have been described in the literature (Lin et al., 2010; Min et al., 2017), their downstream effects have not been contextualized (i.e., presented in a coherent/integrated manner like a knowledge model does). The analysis we have presented focusing on the crosstalk between heme biology and the TLR signaling pathway has shed some light on how this crosstalk could be related to heme biology. However, there also exist other well-known pathways related to heme, that could be investigated by conducting similar analyses in the future.

DATA AVAILABILITY STATEMENT

The data sets and scripts of this study can be found at https://github.com/hemekg.

AUTHOR CONTRIBUTIONS

DI, MH-A, and DD-F conceived and designed the study. FH curated the data and conducted the main analysis supervised by AP, DI, and DD-F. M-TH, BS, MD, and AP assisted in selecting the corpora and interpreting the results. CH designed the curation guidelines and implemented the Python package. DD-F, FH, CH, M-TH, BS, MD, and DI wrote and reviewed the manuscript.

FUNDING

This work was financially supported by the University of Bonn to DI and the Fraunhofer-Gesellschaft to MH-A is gratefully acknowledged.

ACKNOWLEDGMENTS

We would like to thank Sarah Mubeen for proofreading the article, and Amelie Wißbrock for useful scientific discussions. Finally, we would also like to thank the reviewers for their comments and suggestions.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fbioe.2020.00074/full#supplementary-material

REFERENCES

Ali, M., Hoyt, C. T., Domingo-Fernández, D., Lehmann, J., and Jabeen, H. (2019). BioKEEN: a library for learning and evaluating biological knowledge graph embeddings. *Bioinformatics* 35, 3538–3540. doi: 10.1093/bioinformatics/btz117

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. doi: 10.1038/75556

Catlett, N. L., Bargnesi, A. J., Ungerer, S., Seagaran, T., Ladd, W., Elliston, K. O., et al. (2013). Reverse causal reasoning: applying qualitative causal knowledge to the interpretation of high-throughput data. *BMC Bioinform* 14:340. doi: 10.1186/1471-2105-14-340

Conran, N. (2014). Intravascular hemolysis: a disease mechanism not to be ignored. *Acta Haematol.* 132, 97–99. doi: 10.1159/000356836

Cote, R., Reisinger, F., Martens, L., Barsnes, H., Vizcaino, J. A., and Hermjakob, H. (2010). The ontology lookup service: bigger and better. *Nucleic Acids Res.* 38, W155–W160. doi: 10.1093/nar/gkq331

Çubuk, C., Hidalgo, M. R., Amadoz, A., Rian, K., Salavert, F., Pujana, M. A., et al. (2019). Differential metabolic activity and discovery of therapeutic targets using summarized metabolic pathway models. *Npj Syst. Biol. Appl.* 5:7. doi: 10.1101/367334

Degtyarenko, K., De Matos, P., Ennis, M., Hastings, J., Zbinden, M., McNaught, A., et al. (2007). ChEBI: a database and ontology for chemical entities of biological interest. *Nucleic. Acids Res.* 36, D344–D350. doi: 10.1093/nar/gkm791

Dey, S., Bindu, S., Goyal, M., Pal, C., Alam, A., Iqbal, M. S., et al. (2012). Impact of intravascular hemolysis in malaria on liver dysfunction. *J. Biol. Chem.* 287, 26630–26646. doi: 10.1074/jbc.m112.341255

Dickinson-Copeland, C. M., Wilson, N. O., Liu, M., Driss, A., Salifu, H., Adjei, A. A., et al. (2015). Heme-mediated induction of CXCL10 and depletion of CD34+ progenitor cells is Toll-like receptor 4 dependent. *PLoS One* 10:e0142328. doi: 10.1371/journal.pone.0142328

Domingo-Fernández, D., Mubeen, S., Marín-Llaó, J., Hoyt, C. T., and Hofmann-Apitius, M. (2019). PathMe: merging and exploring mechanistic pathway knowledge. *BMC Bioinform* 20:243. doi: 10.1101/451625

Donato, M., Xu, Z., Tomoiaga, A., Granneman, J. G., MacKenzie, R. G., Bao, R., et al. (2013). Analysis and correction of crosstalk effects in pathway analysis. *Genome Res.* 23, 1885–1893. doi: 10.1101/gr.153551.112

Dutra, F. F., and Bozza, M. T. (2014). Heme on innate immunity and inflammation. *Front. Pharmacol.* 5:115. doi: 10.3389/fphar.2014.00115

Effenberger-Neidnicht, K., and Hartmann, M. (2018). Mechanisms of hemolysis during sepsis. *Inflammation* 41, 1569–1581. doi: 10.1007/s10753-018-0810-y

Esteban-Medina, M., Peña-Chilet, M., Loucera, C., and Dopazo, J. (2019). Exploring the druggable space around the fanconi anemia pathway using machine learning and mechanistic models. *BMC Bioinform* 20:370. doi: 10.1186/s12859-019-2969-0

Fabregat, A., Jupe, S., Matthews, L., Sidiropoulos, K., Gillespie, M., Garapati, P., et al. (2017). The Reactome pathway knowledgebase. *Nucleic Acids Res.* 42, D472–D477. doi: 10.1093/nar/gkx1132

Ferreira, A., Balla, J., Jeney, V., Balla, G., and Soares, M. P. (2008). A central role for free heme in the pathogenesis of severe malaria: the missing link? *J. Mol. Med.* 86, 1097–1111. doi: 10.1007/s00109-008-0368-5

Figueiredo, R. T., Fernández, P. L., Mourao-Sa, D. S., Porto, B. N., Dutra, F. F., Alves, L. S., et al. (2007). Characterization of heme as activator of Toll-like receptor 4. *J. Biol. Chem.* 282, 20221–20229. doi: 10.1074/jbc.m610737200

Fisher, R. A. (1992). "Statistical Methods for Research Workers, in *Breakthroughs in Statistics. Springer Series in Statistics (Perspectives in Statistics)*, eds S. Kotz and N. L. Johnson (New York, NY: Springer), 66–70.

Fortes, G. B., Alves, L. S., de Oliveira, R., Dutra, F. F., Rodrigues, D., Fernandez, P. L., et al. (2012). Heme induces programmed necrosis on macrophages through autocrine TNF and ROS production. *Blood* 119, 2368–2375. doi: 10.1182/blood-2011-08-375303

Fujita, K. A., Ostaszewski, M., Matsuoka, Y., Ghosh, S., Glaab, E., Trefois, C., et al. (2014). Integrating pathways of Parkinson's disease in a molecular interaction map. *Mol. Neurobiol.* 49, 88–102. doi: 10.1007/s12035-013-8489-4

Garcia-Santos, D., Hamdi, A., Saxova, Z., Fillebeen, C., Pantopoulos, K., Horvathova, M., et al. (2017). Inhibition of heme oxygenase ameliorates anemia and reduces iron overload in a thalassemia mouse model. *Blood* 131, 236–246. doi: 10.1182/blood-2017-07-798728

Gouveia, Z., Carlos, A. R., Yuan, X., da Silva, F. A., Stocker, R., Maghzal, G. J., et al. (2017). Characterization of plasma labile heme in hemolytic conditions. *FEBS J.* 284, 3278–3301. doi: 10.1111/febs.14192

Hama, M., Kirino, Y., Takeno, M., Takase, K., Miyazaki, T., Yoshimi, R., et al. (2012). Bach1 regulates osteoclastogenesis in a mouse model via both heme oxygenase 1–dependent and heme oxygenase 1–independent pathways. *Arthritis Rheum* 64, 1518–1528. doi: 10.1002/art.33497

Horng, T., Barton, G. M., Flavell, R. A., and Medzhitov, R. (2002). The adaptor molecule TIRAP provides signalling specificity for Toll-like receptors. *Nature* 420:329. doi: 10.1038/nature01180

Hoyt, C. T., Domingo-Fernández, D., and Hofmann-Apitius, M. (2018). BEL commons: an environment for exploration and analysis of networks encoded in biological expression language. *Database* 2018, 1–11. doi: 10.1093/database/bay126

Huang, H.-F., Zeng, Z., Wang, K. H., Zhang, H. Y., Wang, S., Zhou, W. X., et al. (2015). Heme oxygenase-1 protects rat liver against warm ischemia/reperfusion injury via TLR2/TLR4-triggered signaling pathways. *World J. Gastroenterol.* 21:2937. doi: 10.3748/wjg.v21.i10.2937

IJssennagger, N., Derrien, M., van, Doorn GM, Rijnierse, A., van, den Bogert B, Müller, M., et al. (2012). Dietary heme alters microbiota and mucosa of mouse colon without functional changes in host-microbe cross-talk. *PLoS One* 7:e49868. doi: 10.1371/journal.pone.0049868

Jeney, V. (2002). Pro-oxidant and cytotoxic effects of circulating heme. *Blood* 100, 879–887. doi: 10.1182/blood.v100.3.879

Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., and Morishima, K. (2016). KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.* 45, D353–D361. doi: 10.1093/nar/gkw1092

Khanna, S., Domingo-Fernández, D., Iyappan, A., Emon, M. A., Hofmann-Apitius, M., and Frohlich, H. (2018). Using multi-scale genetic, neuroimaging and clinical data for predicting alzheimer's disease and reconstruction of relevant biological mechanisms. *Sci. Rep.* 8:11173. doi: 10.1038/s41598-018-29433-3

Kühl, T., and Imhof, D. (2014). Regulatory FeII/III heme: the reconstruction of a molecule's biography. *ChemBioChem.* 15, 2024–2035. doi: 10.1002/cbic.201402218

Kuperstein, I., Bonnet, E., Nguyen, H. A., Cohen, D., Viara, E., Grieco, L., et al. (2015). Atlas of cancer signalling network: a systems biology resource for integrative analysis of cancer data with Google Maps. *Oncogenesis* 4:e160. doi: 10.1038/oncsis.2015.19

L'Acqua, C., and Hod, E. (2014). New perspectives on the thrombotic complications of haemolysis. *Br. J. Haematol.* 168, 175–185. doi: 10.1111/bjh.13183

Larsen, R., Gozzelino, R., Jeney, V., Tokaji, L., Bozza, F. A., Japiassu, A. M., et al. (2010). A central role for free heme in the pathogenesis of severe sepsis. *Sci. Transl. Med.* 2:51ra71. doi: 10.1126/scitranslmed.3001118

Lin, S., Yin, Q., Zhong, Q., Lv, F.-L., Zhou, Y., Li, J.-Q., et al. (2012). Heme activates TLR4-mediated inflammatory injury via MyD88/TRIF signaling pathway in intracerebral hemorrhage. *J. Neuroinflammation* 9:46. doi: 10.1186/1742-2094-9-46

Lin, T., Kwak, Y. H., Sammy, F., He, P., Thundivalappil, S., Sun, G., et al. (2010). Synergistic inflammation is induced by blood degradation products with microbial toll-like receptor agonists and is blocked by hemopexin. *J. Infect. Dis.* 202, 624–632. doi: 10.1086/654929

Litvinov, R. I., and Weisel, J. W. (2016). Role of red blood cells in haemostasis and thrombosis. *ISBT Sci. Ser.* 12, 176–183. doi: 10.1111/voxs.12331

Marai, E. G., Pinaud, B., Bühler, K., Lex, A., and Morris, J. H. (2019). Ten simple rules to create biological network figures for communication. *PLoS Comput. Biol.* 15:e1007244. doi: 10.1371/journal.pcbi.1007244

Martins, R., and Knapp, S. (2018). Heme and hemolysis in innate immunity: adding insult to injury. *Curr Opin Immunol.* 50, 14–20. doi: 10.1016/j.coi.2017.10.005

Meng, Z., Zhao, T., Zhou, K., Zhong, Q., Wang, Y., Xiong, X., et al. (2017). A20 ameliorates intracerebral hemorrhage–induced inflammatory injury by regulating TRAF6 Polyubiquitination. *J. Immunol.* 198, 820–831. doi: 10.4049/jimmunol.1600334

Merle, N. S., Paule, R., Leon, J., Daugan, M., Robe-Rybkine, T., Poillerat, V., et al. (2019). P-selectin drives complement attack on endothelium during intravascular hemolysis in TLR-4/heme-dependent manner. *Proc. Natl. Acad. Sci. U.S.A.* 116, 6280–6285. doi: 10.1073/pnas.1814797116

Min, H., Choi, B., Jang, Y. H., Cho, I.-H., and Lee, S. J. (2017). Heme molecule functions as an endogenous agonist of astrocyte TLR2 to contribute to secondary brain damage after intracerebral hemorrhage. *Mol. Brain* 10:27. doi: 10.1186/s13041-017-0305-z

Mizuno, S., Iijima, R., Ogishima, S., Kikuchi, M., Matsuoka, Y., Ghosh, S., et al. (2012). AlzPathway: a comprehensive map of signaling pathways of Alzheimer's disease. *BMC Syst. Biol.* 6:52. doi: 10.1186/1752-0509-6-52

Nagy, E., Eaton, J. W., Jeney, V., Soares, M. P., Varga, Z., Galajda, Z., et al. (2010). Red cells, hemoglobin, heme, iron, and atherogenesis. *Arter. Thromb. Vasc. Biol.* 30, 1347–1353. doi: 10.1161/ATVBAHA.110.206433

Ostaszewski, M., Gebel, S., Kuperstein, I., Mazein, A., Zinovyev, A., Dogrusoz, U., et al. (2018). Community-driven roadmap for integrated disease maps. *Brief. Bioinform* 20, 659–670. doi: 10.1093/bib/bby024

Pålsson-Mcdermott, E. M., and O'Neill, L. A. (2004). Signal transduction by the lipopolysaccharide receptor, Toll-like receptor 4. *Immunology* 113, 153–162. doi: 10.1111/j.1365-2567.2004.01976.x

Park, Y., Ryu, H. S., Lee, H. K., Kim, J. S., Yun, J., Kang, J. S., et al. (2014). Tussilagone inhibits dendritic cell functions via induction of heme oxygenase-1. *Int. Immunopharmacol* 22, 400–408. doi: 10.1016/j.intimp.2014.07.023

Piazza, M., Damore, G., Costa, B., Gioannini, T. L., Weiss, J. P., and Peri, F. (2010). Hemin and a metabolic derivative coprohemin modulate the TLR4 pathway differently through different molecular targets. *Innate Immun.* 17, 293–301. doi: 10.1177/1753425910369020

Poulos, T. L. (2014). Heme enzyme structure and function. *Chem. Rev.* 114, 3919–3962. doi: 10.1021/cr400415k

Pratt, D., Chen, J., Welker, D., Rivas, R., Pillich, R., Rynkov, V., et al. (2015). NDEx, the network data exchange. *Cell Syst.* 1, 302–305. doi: 10.1016/j.cels.2015.10.001

Rodriguez-Esteban, R. (2015). Biocuration with insufficient resources and fixed timelines. *Database* 2015:bav116. doi: 10.1093/database/bav116

Rogers, F. B. (1963). Medical subject headings. *Bull Med Libr Assoc.* 51, 114–116.

Roumenina, L. T., Rayes, J., Lacroix-Desmazes, S., and Dimitrov, J. D. (2016). Heme: Modulator of plasma systems in hemolytic diseases. *Trends. Mol. Med.* 22, 200–213. doi: 10.1016/j.molmed.2016.01.004

Seya, T., Oshiumi, H., Sasai, M., Akazawa, T., and Matsumoto, M. (2005). TICAM-1 and TICAM-2: toll-like receptor adapters that participate in induction of type 1 interferons. *Int. J. Biochem. Cell Biol.* 37, 524–529. doi: 10.1016/j.biocel.2004.07.018

Shannon, P. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* 13, 2498–2504. doi: 10.1101/gr.1239303

Slenter, D. N., Kutmon, M., Hanspers, K., Riutta, A., Windsor, J., Nunes, N., et al. (2017). WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research. *Nucleic Acids Res.* 46, D661–D667. doi: 10.1093/nar/gkx1064

Smith, A., and McCulloh, R. J. (2015). Hemopexin and haptoglobin: allies against heme toxicity from hemoglobin not contenders. *Front Physiol.* 6:187. doi: 10.3389/fphys.2015.00187

Smith, A., and Warren, M. (2009). *Tetrapyrroles: Birth, Life and Death*. New York, NY: Springer.

Soares, M. P., and Bozza, M. T. (2016). Red alert: labile heme is an alarmin. *Curr. Opin. Immunol.* 38, 94–100. doi: 10.1016/j.coi.2015.11.006

Sudan, K., Vijayan, V., Madyaningrana, K., Gueler, F., Igarashi, K., Foresti, R., et al. (2019). TLR4 activation alters labile heme levels to regulate BACH1 and heme oxygenase-1 expression in macrophages. *Free Rad. Biol. Med.* 137, 131–142. doi: 10.1016/j.freeradbiomed.2019.04.024

Tolosano, E., Fagoonee, S., Morello, N., Vinchi, F., and Fiorito, V. (2010). Heme scavenging and the other facets of hemopexin. *Antioxid. Redox Signal.* 12, 305–320. doi: 10.1089/ars.2009.2787

Vinchi, F., De Franceschi, L., Ghigo, A., Townes, T., Cimino, J., Silengo, L., et al. (2013). Hemopexin therapy improves cardiovascular function by preventing heme induced endothelial toxicity in mouse models of hemolytic diseases. *Circulation* 127, 1317–1329. doi: 10.1161/CIRCULATIONAHA.112.130179

Vogel, S., and Thein, S. L. (2018). Platelets at the crossroads of thrombosis, inflammation and haemolysis. *Br. J. Haematol.* 180, 761–767. doi: 10.1111/bjh.15117

Wagener, F. A. D. T. G., van Beurden, H. E., von den Hoff, J. W., Adema, G. J., and Figdor, C. G. (2003). The heme-heme oxygenase system: a molecular switch in wound healing. *Blood* 102, 521–528. doi: 10.1182/blood-2002-07-2248

Yekutieli, D., and Benjamini, Y. (2001). The control of the false discovery rate in multiple testing under dependency. *Ann. Stat.* 29, 1165–1188. doi: 10.1214/aos/1013699998

Zhang, L. (2011). *Heme Biology: the Secret Life of Heme in Regulating Diverse Biological Processes*. Singapore: World Scientific.

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 Bioengineering and Biotechnology

ORIGINAL RESEARCH
published: 20 March 2020
doi: 10.3389/fbioe.2020.00220

Comparative Study of Transcriptomics-Based Scoring Metrics for the Epithelial-Hybrid-Mesenchymal Spectrum

*Priyanka Chakraborty[1], Jason T. George[2,3], Shubham Tripathi[2,4,5], Herbert Levine[2,5,6] and Mohit Kumar Jolly[1]**

[1] Centre for BioSystems Science and Engineering, Indian Institute of Science, Bengaluru, India, [2] Center for Theoretical Biological Physics, Rice University, Houston, TX, United States, [3] Medical Scientist Training Program, Baylor College of Medicine, Houston, TX, United States, [4] Ph.D. Program in Systems, Synthetic, and Physical Biology, Rice University, Houston, TX, United States, [5] Department of Physics, College of Science, Northeastern University, Boston, MA, United States, [6] Department of Bioengineering, College of Engineering, Northeastern University, Boston, MA, United States

OPEN ACCESS

Edited by:
Jyoti Sharma, Institute of Bioinformatics (IOB), India

Reviewed by:
Tian Hong, The University of Tennessee, Knoxville, United States
Prashanth N. Suravajhala, Birla Institute of Scientific Research, India

***Correspondence:**
Mohit Kumar Jolly
mkjolly@iisc.ac.in; mkjolly.15@gmail.com

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Bioengineering and Biotechnology

Received: *02 January 2020*
Accepted: *04 March 2020*
Published: *20 March 2020*

Citation:
Chakraborty P, George JT, Tripathi S, Levine H and Jolly MK (2020) Comparative Study of Transcriptomics-Based Scoring Metrics for the Epithelial-Hybrid-Mesenchymal Spectrum. Front. Bioeng. Biotechnol. 8:220. doi: 10.3389/fbioe.2020.00220

The Epithelial–mesenchymal transition (EMT) is a cellular process implicated in embryonic development, wound healing, and pathological conditions such as cancer metastasis and fibrosis. Cancer cells undergoing EMT exhibit enhanced aggressive behavior characterized by drug resistance, tumor-initiation potential, and the ability to evade the immune system. Recent *in silico*, *in vitro*, and *in vivo* evidence indicates that EMT is not an all-or-none process; instead, cells can stably acquire one or more hybrid epithelial/mesenchymal (E/M) phenotypes which often can be more aggressive than purely E or M cell populations. Thus, the EMT status of cancer cells can prove to be a critical estimate of patient prognosis. Recent attempts have employed different transcriptomics signatures to quantify EMT status in cell lines and patient tumors. However, a comprehensive comparison of these methods, including their accuracy in identifying cells in the hybrid E/M phenotype(s), is lacking. Here, we compare three distinct metrics that score EMT on a continuum, based on the transcriptomics signature of individual samples. Our results demonstrate that these methods exhibit good concordance among themselves in quantifying the extent of EMT in a given sample. Moreover, scoring EMT using any of the three methods discerned that cells can undergo varying extents of EMT across tumor types. Separately, our analysis also identified tumor types with maximum variability in terms of EMT and associated an enrichment of hybrid E/M signatures in these samples. Moreover, we also found that the multinomial logistic regression (MLR)-based metric was capable of distinguishing between "pure" individual hybrid E/M vs. mixtures of E and M cells. Our results, thus, suggest that while any of the three methods can indicate a generic trend in the EMT status of a given cell, the MLR method has two additional advantages: (a) it uses a small number of predictors to calculate the EMT score and (b) it can predict from the transcriptomic signature of a population whether it is comprised of "pure" hybrid E/M cells at the single-cell level or is instead an ensemble of E and M cell subpopulations.

Keywords: EMT, MET, EMT score, EMT quantification, tumor heterogeneity, hybrid epithelial/mesenchymal

INTRODUCTION

The epithelial–mesenchymal transition (EMT) is a cell biological process crucial for various aspects of tumor aggressiveness – cancer metastasis (Jolly et al., 2017), resistance against cell death (Huang et al., 2013), metabolic reprogramming (Thomson et al., 2019), refractory response to chemotherapy and radiotherapy (Kurrey et al., 2009), tumor-initiation potential (Jolly et al., 2014), and immune evasion (Tripathi et al., 2016; Terry et al., 2017) – thus eventually affecting patient survival (Tan et al., 2014). EMT is a multidimensional, non-linear process that involves changes in a compendium of molecular and morphological traits, such as altered cell polarity, partial or complete loss of cell–cell adhesion, and increased migration and invasion. Cells may take different routes in this multidimensional landscape as effectively captured by recent high-throughput dynamic approaches (Karacosta et al., 2019; Watanabe et al., 2019). The trajectories taken by cancer cells in the EMT landscape may depend on the dosage and duration of the EMT induction signal (Stylianou et al., 2018; Katsuno et al., 2019; Tripathi et al., 2020), and thus may be associated with varying metastatic potency (Aiello et al., 2018) and varying degrees of resistance against different drugs (Biddle et al., 2016), thereby driving a context-specific association of patient survival with EMT (Chikaishi et al., 2011; Tan et al., 2014; Yan et al., 2016).

Initially thought of as binary, EMT is now considered as a complex process involving one or more hybrid epithelial/mesenchymal (E/M) states (Jolly and Celia-Terrassa, 2019). These hybrid E/M states can be more plastic and tumorigenic than "purely E" or "purely M" ones, thus constituting the "fittest" phenotype for metastasis (Grosse-Wilde et al., 2015; Bierie et al., 2017; Pastushenko et al., 2018; Kröger et al., 2019; Tripathi et al., 2020). Consequently, the presence and frequency of such hybrid E/M cells in primary tumors and in circulating tumor cells (CTCs) can be associated with poor patient survival (Jolly et al., 2019a; Saxena et al., 2019). Computational methods aimed at quantifying EMT on a continuous spectrum in order to enhance diagnostic, prognostic, and therapeutic intervention are therefore indispensable.

Various methods have been developed to obtain a quantitative measure of the extent of EMT (hereafter, referred to as EMT score) that cells in a given sample have undergone. Here we focus on methods accomplishing this task using the gene expression data. First, a 76-gene EMT signature (76GS; hereafter referred to as the 76GS method) was developed and validated using gene expression from non-small cell lung cancer (NSCLC) cell lines and patients treated in the BATTLE trial (Byers et al., 2013). This scoring method calculates EMT scores based on a weighted sum of the expression levels of 76 genes; the weight factor of a gene is the correlation coefficient between the expression level of that gene and that of CDH1 (E-cadherin) in that dataset; thus, the absolute EMT scores of E samples using the 76 GS method are relatively higher than those of M samples (Guo et al., 2019). Second, an EMT score separately for cell lines and tumors was developed based on a two-sample Kolmogorov–Smirnov test (KS; hereafter referred to as the KS method). This score varies on a scale of −1 to 1, with the higher scores corresponding to more M samples (Tan et al., 2014). Third, a multinomial logistic regression (MLR; hereafter referred to as the MLR method)-based model quantified the extent of EMT in a given sample on a scale of 0–2. This method particularly focuses on characterizing a hybrid E/M phenotype using the expression levels of 23 genes – 3 predictors and 20 normalizers – identified through NCI-60 gene expression data. It consequently calculates the probability that given sample belongs to E, M, or hybrid E/M categories. An EMT score is assigned based on those probabilities; the higher the score, the more M the sample is (George et al., 2017). A comparative analysis of these methods in terms of similarities, differences, strengths, and limitations, remains to be done.

Here, we present a comprehensive evaluation of these methods – 76GS, KS, and MLR – in terms of quantifying EMT and characterizing the hybrid E/M phenotype. First, we calculate the correlations observed across different *in vitro*, *in vivo*, and patient datasets, and observe good quantitative agreement among the scores calculated using these three methods. This analysis suggests that all of them, despite using varied gene lists and methods, concur in capturing a generic trend embedded in the multi-dimensional EMT gene expression landscape. Second, we identify which cancer types are more heterogeneous than others in terms of their EMT status; intriguingly, our results show that enrichment for a hybrid E/M phenotype contributes to heterogeneity. Third, we compare the ability of these methods to distinguish between "pure" individual hybrid E/M cells vs. mixtures of E and M cells that can exhibit an EMT score similar to that of hybrid E/M samples. Our results offer proof-of-principle that the MLR method can identify these differences. Overall, our results demonstrate the consistency of these EMT scoring metrics in quantifying the spectrum of EMT. Moreover, two advantages of MLR method are highlighted – namely, the use of a small number of predictors to calculate the EMT score, and the ability to characterize difference between admixtures of E and M cells vs. truly hybrid E/M cells.

MATERIALS AND METHODS

Software and Datasets

All computational and statistical analyses were performed using R (version 3.4.0) and Bioconductor (version 3.6). Microarray datasets were downloaded using *GEOquery* R Bioconductor package (Davis and Meltzer, 2007). TCGA datasets were obtained from the *UCSC xena tools* (Wang S. et al., 2019). CCLE and NCI60 datasets were downloaded from respective websites.

Preprocessing of Microarray Data Sets

All microarray datasets were preprocessed to obtain the gene-wise expression for each sample from probe-wise expression matrix. To map the probes to genes, relevant platform annotation files were utilized. If there were multiple probes mapping to one gene, then the mean expression of all the mapped probes was considered for that gene.

Calculation of EMT Scores

Epithelial–mesenchymal transition (EMT) scores were calculated for samples in a particular data set using all three methods. For

a particular microarray data set, expression of respective gene signatures was given as an input to calculate EMT score using all three different methods.

76GS

The EMT scores were calculated based on a 76-gene expression signature reported (Byers et al., 2013; **Supplementary Table S1**) and the metric mentioned based on that gene signature (Guo et al., 2019). For each sample, the score was calculated as a weighted sum of 76 gene expression levels and the scores were centered by subtracting the mean across all tumor samples so that the grand mean of the score was zero. Negative scores can be interpreted as M phenotype whereas the positive scores as E.

MLR

The ordinal MLR method predicts EMT status based on the order structure of categories and the principle that the hybrid E/M state falls in a region intermediary to E and M. Quantitative estimates of EMT spectrum were inferred based on the assumptions and equations mentioned (George et al., 2017; **Supplementary Table S2**). The samples are scored ranging from 0 (pure E) to 2 (pure M), with a score of 1 indicating a maximally hybrid phenotype. These scores are calculated based on the probability of a given sample being assigned to the E, E/M, and M phenotypes.

KS

The KS EMT scores were calculated as previously reported (Tan et al., 2014; **Supplementary Tables S3, S4**). This method compares cumulative distribution functions (CDFs) of E and M signatures. First, the distance between E and M signatures was calculated via the maximum distance between their CDFs as follows: For CDFs $F_E(x)$ and $F_M(x)$ representing the levels of transcript x for E and M signatures, respectively, the distance between signatures is assessed by using the uniform norm

$$||F_E - F_M|| \equiv \max_x |F_E(x) - F_M(x)| \quad (1)$$

This quantity represents the test statistic in the subsequent two-sample test used to calculate the EMT score. The score is determined by hypothesis testing of two alternative hypotheses as follows (with the null hypothesis being that there is no difference in CDFs of M and E signatures): (1) CDF of M signature is greater than CDF of E signature. (2) CDF of E is greater than CDF of M signature. Sample with a positive EMT score is M whereas negative EMT score is associated with E phenotype.

Correlation Analysis

Correlation between EMT scores was calculated by Pearson's correlation, unless otherwise mentioned.

Survival Analysis

All samples were segregated into $76GS^{high}$ and $76GS^{low}$, MLR^{high} and MLR^{low}, KS^{high} and KS^{low} groups based on the mean values of respective EMT score. Observed survival distributions are graphically depicted for each method with the above-mentioned two categories.

Mixture Curve Analysis

For each dataset analyzed using mixture curves, the most M (pure-M) and most E (pure-E) samples were identified by ordering samples based on MLR EMT score and selecting the top and bottom 35 samples, respectively. The mean or median was calculated for the pure-E and pure-M samples as a representative of the purified E or M state in the MLR predictor space. From this, the mixture curve is derived by taking all convex combinations of purified states. Individual samples within a given dataset were ranked based on their proximity to the mixture curve using the usual l_2-norm distance. The top 10, 20, 50, and 100 samples closest to, and furthest from, the mixture curve were used as representative mixtures of E and M populations and hybrid E/M signatures, respectively.

RESULTS

Concordance in Capturing EMT Response

We used three different EMT scoring methods to quantify the extent of EMT in given transcriptomics data; each method utilizes a distinct gene set as well as a different underlying algorithm. In the 76GS method, the higher the score, the more E a sample is, given that the method calculates as weighted sum of expression levels of 76 genes, with the weight factor being correlation coefficient with levels of the canonical E marker CDH1 (**Figure 1A**). This method has no specific predefined range of values, although the range of values obtained are bounded by the maximal possible value of gene expression detected by microarray. Unlike the 76GS method, the MLR and KS methods have predefined scales for EMT scores. MLR and KS score EMT on a spectrum of [0, 2] and [−1, 1] respectively, with higher scores indicating M signatures (**Figures 1B,C**). While MLR and KS methods are absolute, requiring a fixed transcript signature for EMT score calculation, the 76GS method of EMT scoring depends on the number and nature of samples analyzed in a given dataset. Consequently, a hybrid E/M sample may have a (pseudo) low 76GS score whenever the available dataset contains more M samples, or a (pseudo) high score for datasets enriched in E samples. Each scoring method also varies in the number of required gene transcripts: while the MLR method utilizes 23 entries, the 76GS method requires 76 entires. The KS method utilizes 315 and 218 transcripts for tumor samples and cell lines samples, respectively.

We first investigated the extent of concordance in EMT scores calculated via these three methods for well-studied cohorts of cancer cell lines: NCI-60 and CCLE (Shankavaram et al., 2009; Barretina et al., 2012). We expected to see a negative correlation between EMT scores calculated via 76GS and KS methods and that between EMT scores using 76GS and MLR methods, whereas a positive correlation should exist between EMT scores from the MLR and KS methods. Indeed, for both NCI-60 and CCLE datasets, the EMT scores calculated via different methods were found to be correlated significantly with a high absolute value of correlation coefficients in the expected direction, when compared

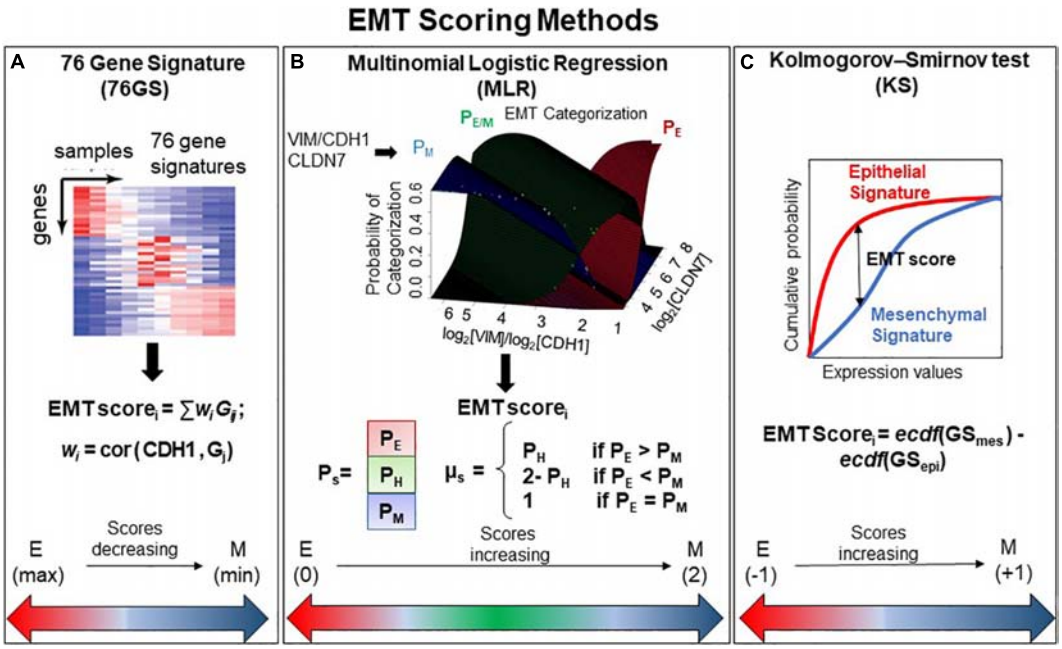


FIGURE 1 | General outline of all three EMT scoring methods. **(A)** 76GS score is calculated by weighted sum of 76 genes, where EMT score$_i$ is score for *i*th sample, w_j is correlation of *j*th gene (G_j) with CDH1 gene in that dataset to which the sample *i* belongs, G_{ij} is the *j*th gene's normalized expression in *i*th sample. **(B)** MLR utilizes log_2(VIM)/log_2(CDH1) and log_2(CLDN7) space to predict categorization of a sample into E, E/M, or M category. Where P_E, P_H, and P_M are the probabilities of a sample falling into each phenotype. EMT score$_i$ is the score for *i*th sample, which is defined in relation to P_E, P_H, and P_M. Figure adapted from George et al. (2017) with permission. **(C)** KS score is estimated by the empirical cumulative distributions of epithelial and mesenchymal gene set, denoted by ecdf (GS_{mes}) and ecdf (GS_{epi}), respectively. EMT score$_i$ is the maximum vertical distance between the ecdf (GS_{mes}) and ecdf (GS_{epi}) (given by Eq. 1 in the section "Materials and Methods") for a given sample *i*.

pairwise (**Figure 2** and **Supplementary Figure S1**). Given that the three scoring methods utilize very different metrics and varying number of genes to define and quantify EMT, it was remarkable that all three showed such high consistency in scoring EMT for these datasets that contained cell lines across various cancer types.

Next, we investigated whether this trend was also present in the TCGA patient samples of different tumor types. Again, the trend remained consistent across tumor types – a strongly positive significant correlation between scores via MLR and KS, and a strongly negative significant correlation between scores via 76GS and KS and those via 76GS and MLR methods (**Figures 3A–C** and **Supplementary Figure S2**). Among all tumor types in TCGA data, breast cancer exhibited the highest observed correlation coefficient across methods (**Figure 3C**). Thus, the association between EMT scores and patient survival was assessed using breast cancer patient samples. The samples were scored using all three methods and segregated into high and low groups based on the mean value of each EMT score. The $76GS^{low}$ subgroup can be thought of as similar to the MLR^{high} and/or KS^{high} ones, given their relatively strong M signature. The three EMT scoring methods showed consistent trends in predicting overall survival highlighting that patients with a strongly M phenotype had better survival probability (**Figure 3D**), endorsing the emerging notion that the predominance of EMT in primary tumors and/or CTCs need not always be correlated with worse patient survival (Tan et al., 2014; Saxena et al., 2019).

Epithelial–mesenchymal transition can be driven by diverse biomechanical and/or biochemical stimuli in tumor microenvironments. TGFβ is one of the best-studied drivers of EMT, and a recent study identified a signature specific to TGFβ-induced EMT (Foroutan et al., 2017). EMT scores calculated via any of the three methods – KS, MLR, and 76GS – correlated well with the scores calculated for TGFβ-induced EMT gene signature (**Supplementary Figure S3**), further endorsing the equivalence of these methods in identifying the onset of EMT.

After establishing this consistency using *in vitro* cell line datasets and TCGA patient samples, we focused on several publicly available microarray datasets including those of EMT induction or reversal, isolation of subpopulations, etc. Each dataset comprised a variety of samples in terms of different cell lines, conditions, and treatments. An analysis of different GEO datasets showed that EMT scores calculated via these three methods, when compared pairwise, were significantly correlated in the expected direction (**Figure 4A** and **Supplementary Table S5**). Out of 85 different datasets, a large percentage of them showed trends in the expected direction (62/85 in KS vs. 76GS; 64/85 in MLR vs. 76GS; 49/85 in MLR vs. KS) (**Figure 4B**). Strikingly, 43 datasets were found to be common across all three pairwise comparisons (**Figure 4C**), establishing a high degree of concordance among EMT scores calculated via these three EMT scoring methods.

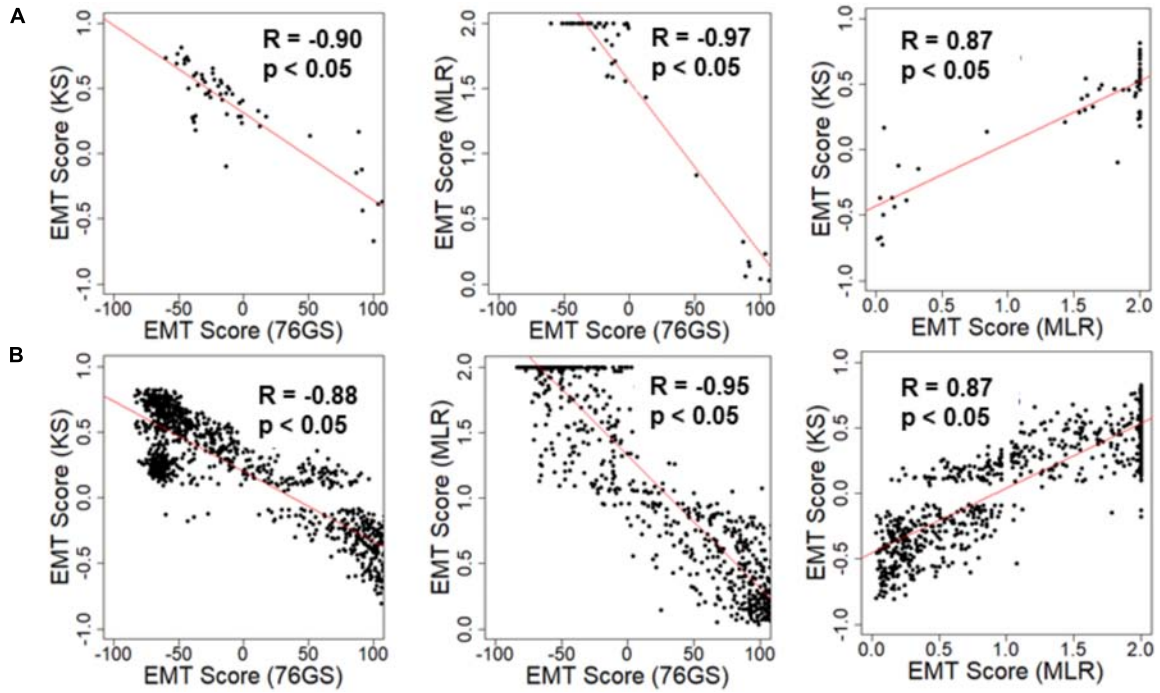


FIGURE 2 | Scatter plot depicting the correlation between the EMT scores of cancer cell line samples calculated via three methods. Each pairwise relation is estimated by a linear regression line (red), Pearson's correlation coefficient (*R*), and *p*-value (*p*) reported in each plot. **(A)** NCI60 dataset and **(B)** CCLE dataset.

Next, we investigated specific cases where EMT/MET was induced in various cell lines by different EMT/MET regulators. Lung cancer cell lines A549, HCC827, and H358 in which EMT was induced by TGFβ showed higher EMT scores using MLR and KS methods, but lower scores via 76GS method, compared to untreated ones (**Figure 5A**). Similarly, the E breast cancer cell line MCF-7 transfected to overexpress EMT-inducing transcription factor Snail exhibited a more M phenotype relative to the control, as identified via all three scoring methods (**Figure 5B**). Consistent trends were seen in EpRAS tumor cells treated with TGFβ (**Figure 5C**), and in human mammary E cells HMLE overexpressing one of the three EMT-inducing transcription factors (EMT-TFs) – SNAI1 (Snail), SNAI2 (Slug), and TWIST (**Figure 5D**). Interestingly, all three scoring methods suggested that EMT induced by Snail or Slug was stronger than that induced by Twist (**Figure 5D**). Further, inducing EMT via overexpression of EMT-TFs Twist, Snail, Goosecoid, or treatment with TGFβ or knockdown of E-cadherin was capable of altering the EMT scores of HMLE cells (**Supplementary Figure S4A**).

Additionally, these three methods also captured the reversal of EMT – M–E transition (MET) – induced by MET-inducing transcription factor GRHL2 in MDA-MB-231 cells (**Figure 5E**). Moreover, baseline differences in EMT status between two hepatocellular carcinoma cell lines identified experimentally (Van Zijl et al., 2011) were also recapitulated by all three scoring methods; while HCC-1.2 (referred to as 3p) showed more E features, HCC1.1 (referred to as 3sp) was relatively more M (**Figure 5F**). We also calculated the EMT scores for the dynamic EMT time series datasets (i.e., cases where more than two time points were available for EMT induction); all three methods were able to recapitulate the relevant trends in EMT scores as expected when EMT was induced in A549 and LNCAP cells (**Supplementary Figures S4B,D**). Further, all three EMT scoring methods captured the trend in the change of EMT status in prostate cancer E PC3 cells (PC3-Epi) and M cell lines derived from PC3 (PC-EMT) through interactions with macrophages (Roca et al., 2013). PC3-EMT cells transfected with ZEB1-shRNA vector (sh4), but not with the scrambled control (Scr), indicated an MET (**Supplementary Figure S4C**). Finally, we calculated EMT scores for a population of CTCs collected from breast cancer patients and *ex vivo* cancer models and observed heterogeneity in CTCs along the E-hybrid–M spectrum (**Supplementary Figures S4E,F**), reminiscent of similar observations based on immunohistochemical staining of a few canonical markers (Yu et al., 2013).

Variability in EMT Scores Measures Tumor Heterogeneity

Recent studies have emphasized that intra-tumor heterogeneity and inter-tumor heterogeneity can accelerate progression and metastasis (Lawson et al., 2018). Thus, we were interested in identifying which tumor types are more heterogeneous with regard to EMT scores calculated via the three methods. We grouped the CCLE samples by different tumor types and calculated the mean and variance of all EMT scores across a given tumor. The EMT scores, calculated across the three methods, showed less variation in the EMT scores of the tumor

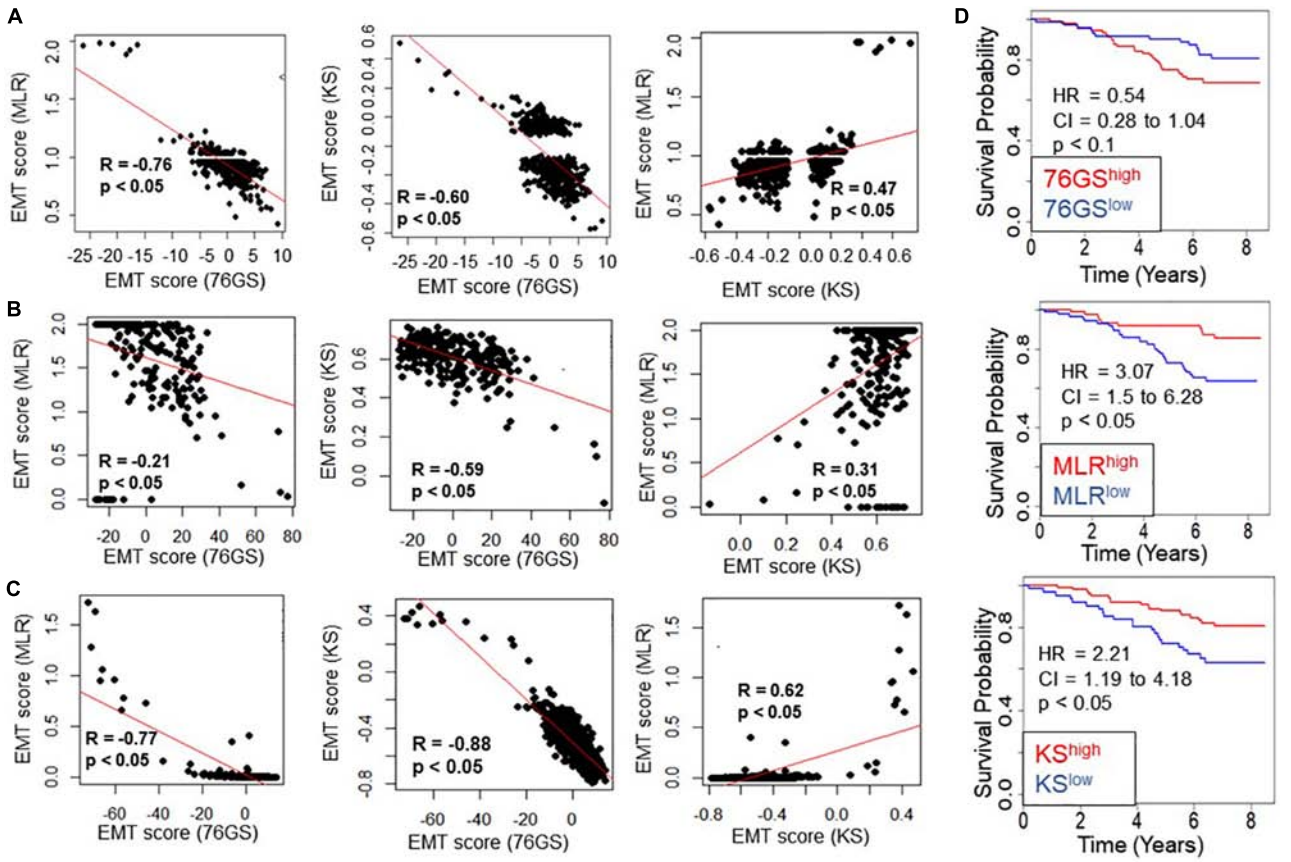


FIGURE 3 | Concordance across all three EMT scoring methods in quantification of EMT and survival prediction of tumor patients. Each pairwise relation is estimated by linear regression (red), Pearson's correlation coefficient (*R*), and *p*-value (*p*), reported in each plot. **(A)** TCGA ovarian cancer dataset, **(B)** TCGA sarcoma dataset, **(C)** TCGA breast cancer dataset. **(D)** Correlation between EMT score (high vs. low) and overall survival (OS) in breast cancer patients. Kaplan–Meier survival analysis is performed to estimate differences in survival of $76GS^{high}$, MLR^{high}, KS^{high} and $76GS^{low}$, MLR^{low}, KS^{low} groups, respectively, in GSE1456. *p*-values (*p*) reported are based on the log rank test. HR (hazard ratio) and confidence interval (95% CI) reported are estimated using cox regression.

types of M origin such as sarcoma and lymphoma, compared to that of the other tumor types such as breast cancer and lung cancer (**Figures 6A–C** and **Supplementary Table S7**). The most heterogeneous tumor types identified based on the variance in EMT scores largely overlapped for all methods: (a) breast cancer, (b) stomach cancer, (c) NSCLC, (d) bile duct cancer, and (e) urinary tract cancer (**Figures 6A–C**). We also calculated pairwise correlations of EMT scores across all the tumor types and observed consistently significant trends (**Supplementary Table S8**).

One of the proposed mechanisms underlying such heterogeneity in EMT status has been E–M plasticity, i.e., the proclivity of individual cells in a population to obtain and switch among multiple phenotypic states. Such plasticity is typically seen to be higher in cells in one or more hybrid E/M states (Pastushenko and Blanpain, 2019; Tripathi et al., 2019, 2020). Thus, we asked whether the frequency of hybrid E/M phenotype contributes to heterogeneity in terms of EMT scoring. One of the EMT scoring methods – MLR – calculates the probability of a given transcriptomic profile being associated with the E, hybrid E/M, or M state, thus enabling us to identify hybrid E/M samples specifically. First, we found that the variance of EMT scores was the highest in samples identified as hybrid E/M as compared to E and M samples (**Supplementary Table S6A**). Consistently, a high correlation coefficient value in EMT scores was maintained, when calculated separately for CCLE samples in E, E/M, and M categories (**Supplementary Table S6B**). Next, we checked the relative frequency and absolute number of hybrid E/M samples (as defined by MLR method) across tumor types, among the cases where EMT scores calculated via all three methods were significantly correlated. Indeed, the tumor types that met the three conditions – (a) total number of hybrid E/M samples being more than 10, (b) percentage of hybrid E/M samples being >20%, and (c) a good correlation among all three methods – were enriched in the most variable tumor types (**Figure 6D**), suggesting hybrid E/M phenotypes contribute maximally to E–M heterogeneity (**Supplementary Table S9**).

We also calculated the correlations in EMT scores obtained from each method, after segregating the cell line samples into E, E/M, and M, based on predictions from the MLR method. The correlation coefficients within the E, E/M, and M subgroups of a given tumor subtype were observed to be

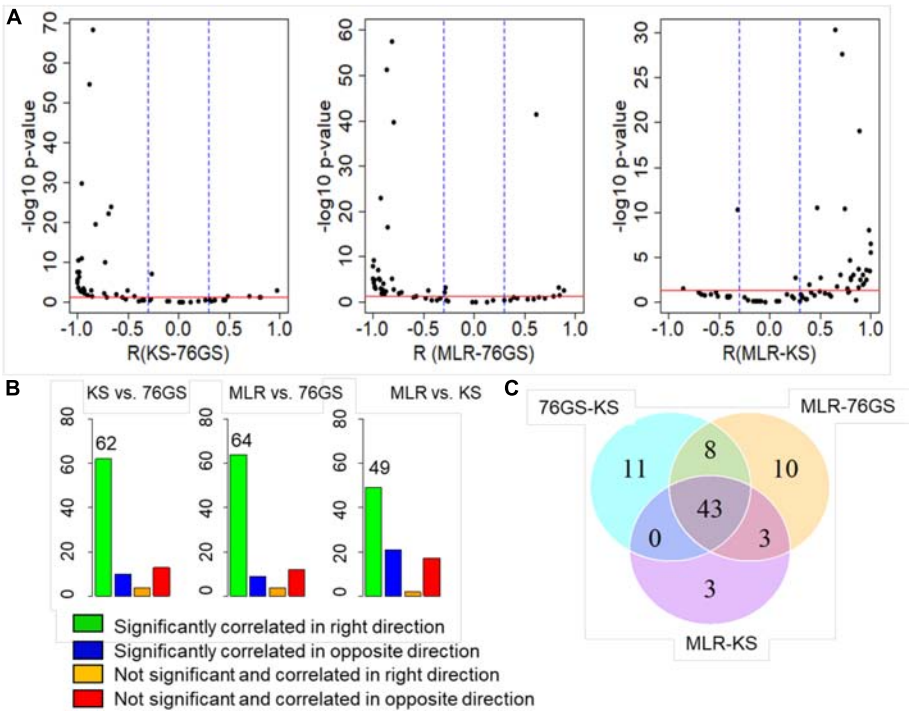


FIGURE 4 | Plots depicting pairwise comparisons of all three EMT scores. **(A)** Volcano plots showing the correlation of different EMT scoring methods across 85 different GEO microarray datasets along with the *p*-values for the respective correlation coefficient values. In each case, –log10(*p*-value) is plotted as a function of Pearson's correlation coefficient. Thresholds for correlation ($R < -0.3$ or $R > 0.3$; vertical blue lines) and *p*-values ($p < 0.05$; horizontal red line) are denoted. **(B)** Bar plots for different categories based on the correlation sign and statistical significance of all three pairwise comparisons across 85 datasets. $p < 0.05$ and $R < -0.3$ or $R > 0.3$. **(C)** Venn diagram showing the common GEO datasets across all pairwise comparisons that are significantly correlated in the expected direction.

somewhat different than those found for all tumor subtype samples without any partitioning into E, E/M, and M subgroups (**Supplementary Table S8**). These results suggest that while a generic trend in terms of EMT scores is seen across the three methods, the categorization in terms of E, E/M, and M may vary to some degree based on the EMT scoring method used. It should be noted that while the MLR method classifies samples into three broad categories (E, E/M, and M), it makes no assumption on the existence, the number, or the stability of sub-states within each category. In fact, the scores calculated using the MLR method use a continuous scale for EMT quantification, which measures the extent of EMT and thus, reflects, in principle, an entire range of different partial states of EMT.

Individual Hybrid E/M Samples Are Different From Hybrid Mixtures of E and M

A given transcriptomic profile may be classified as hybrid E/M for several reasons: (a) the sample contains individually hybrid E/M cells (hybrids), (b) the sample contains a mixture of E and M cells (mixtures), or (c) the sample contains a combination of hybrids and mixtures. We sought to distinguish true hybrids from mixtures based on an additional feature of MLR scoring – mixture curve analysis (Jia et al., 2019). This analysis quantifies the distance of a given sample from a "mixture curve" which connects the position of mean signatures of "pure" E and "pure" M samples. The farther a given sample is from the mixture curve, the higher the likelihood of that particular sample containing truly hybrid E/M cells.

First, we determined the mixture curves based on the CCLE samples. We ranked all cell lines in the CCLE dataset based on their EMT scores and identified the top 35 most E (i.e., lowest 35 in terms of MLR EMT scores) and top 35 most M samples (i.e., highest 35 in terms of EMT MLR scores). Then, the mixture curve was determined based on the convex combinations of mean signatures of these 35 "pure" E and 35 "pure" M reference samples. All the CCLE cell lines identified as hybrid E/M were then plotted alongside the mixture curve (**Figure 7A**) and their distances from the curve were calculated. While some samples fell close to the curve, many deviated substantially (**Figure 7B**). We subsequently picked the farthest and the closest 10, 20, 50, and 100 samples from the mixture curve and calculated their EMT scores. Intriguingly, the mean EMT score of samples farthest from the mixture curve was different than that of the closest samples as calculated using MLR, irrespective of the number of samples chosen (**Figure 7C**). Similarly, another "mixture curve" based on median of 35 "pure" E and "pure" M reference samples was obtained from CCLE dataset (**Supplementary Figure S5A**); the cell lines closest to either mixture curve tended to be more E than the ones farthest

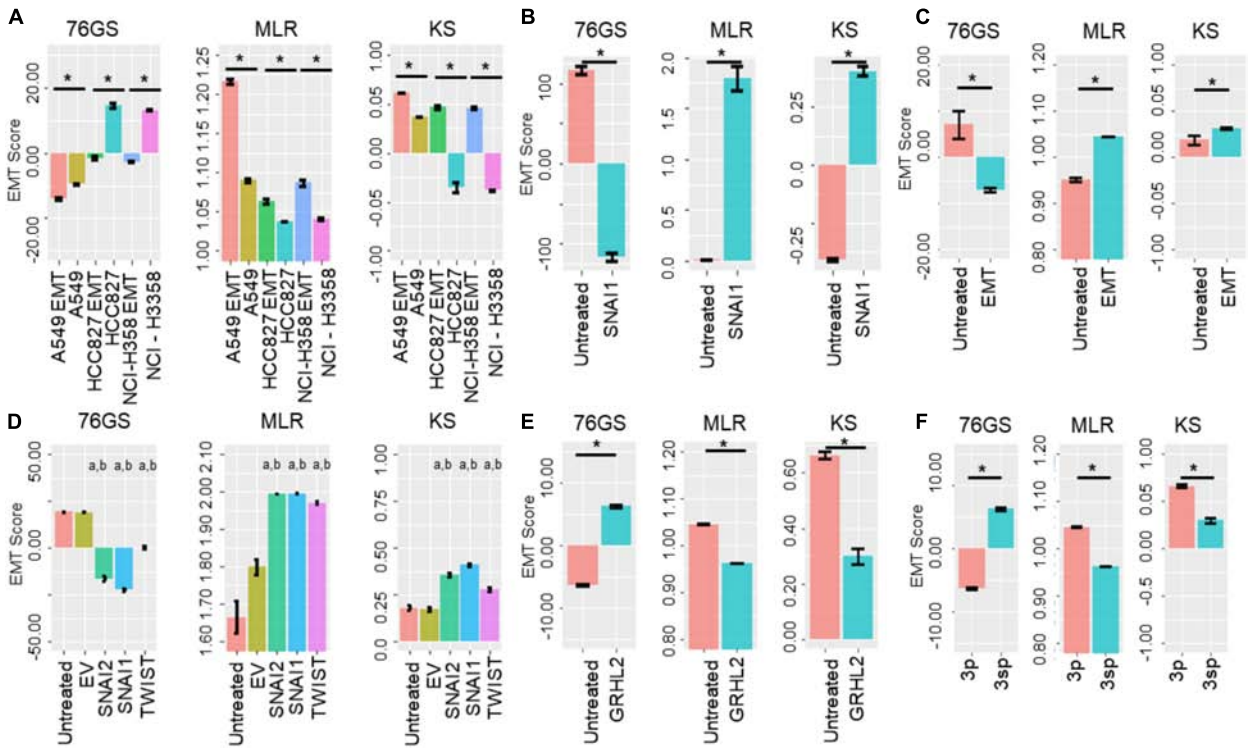


FIGURE 5 | Bar plots showing EMT scores of different cell lines calculated using the three EMT scoring methods. **(A)** EMT induction is shown in three cell lines – A549, HCC87, and NCIH358 (GSE49664). **(B)** EMT induction in MCF7 cell line (GSE58252). **(C)** EMT induction in EpRas cells (GSE59922). **(D)** EMT induction by different EMT-inducing transcription factors. "a" denotes statistical significant difference ($p < 0.05$, $n = 3$, two-tailed Student's t-test) for pairwise comparison of a given set with untreated (first column), "b" denotes the same when compared with empty vector (EV; second column) (GSE43495). **(E)** MET induction by GRHL2 in MDA-MB-231 cell line (GSE36081). **(F)** Two cell lines of hepatocellular carcinoma with varying EMT status (GSE26391). Each control case has been compared to EMT/MET induced case (*$p < 0.05$, $n = 3$, two-tailed Student's t-test; error bars represent standard deviation).

from the curve (**Figure 7C** and **Supplementary Figure S5B**). Qualitatively, speaking 76GS and KS methods showed similar results (**Supplementary Figures S5C–F**).

In order to distinguish the hybrid E/M samples from mixtures of pure E and pure M samples, we lastly characterized the composition of the closest and farthest hybrid E/M samples by estimating the percentage of M phenotype (%M) in each sample based on the convex combination "mixture curve" in the two-dimensional space (VIM/CDH1 expression; CLDN7 expression). While the difference in mean values of the composition (%M) of closest and farthest samples was marginal, but their overall distributions in terms of %M differed substantially (**Figure 7D**). This analysis demonstrates the possibility of a quantifiable compositional difference between truly hybrid E/M samples and mixtures of E and M cells.

DISCUSSION

Epithelial–mesenchymal transition is a reversible and dynamic process which has been shown to be activated during cancer progression. EMT involves a multitude of changes at both molecular and morphological levels. Various attempts to characterize the spectrum of EMT at molecular and/or morphological levels have been made recently, enabled by latest developments in multiplex imaging, single-cell RNA-seq and inducible systems (Mandal et al., 2016; Pastushenko et al., 2018; Stylianou et al., 2018; Cook and Vanderhyden, 2019; Devaraj and Bose, 2019; Karacosta et al., 2019; Wang W. et al., 2019; Watanabe et al., 2019; Lam et al., 2020). These approaches have highlighted the dynamical nature of EMT in driving cancer progression in patients (Jolly and Celia-Terrassa, 2019), and the heterogeneity in EMT status in cell lines and patient samples (Panchy et al., 2020; Shen et al., 2020). Further, various approaches to quantify the EMT spectrum of samples based on different signatures of tumor types have been made (Foroutan et al., 2017; Puram et al., 2017). Among all the methods available for EMT scoring, we have compared the ones that are more generalized – KS (Tan et al., 2014), MLR (George et al., 2017), and 76 GS (Byers et al., 2013; Guo et al., 2019). These three methods use different combinations of genes and metrics; however, they show a very good concordance among them in terms of identifying an empirical trend along the EMT axis.

Here, we compared the aforementioned EMT scoring metrics for their ability to identify the onset and extent of EMT/MET via calculating EMT scores for cell line cohorts NCI-60 and CCLE, TCGA cohorts from multiple subtypes, and datasets

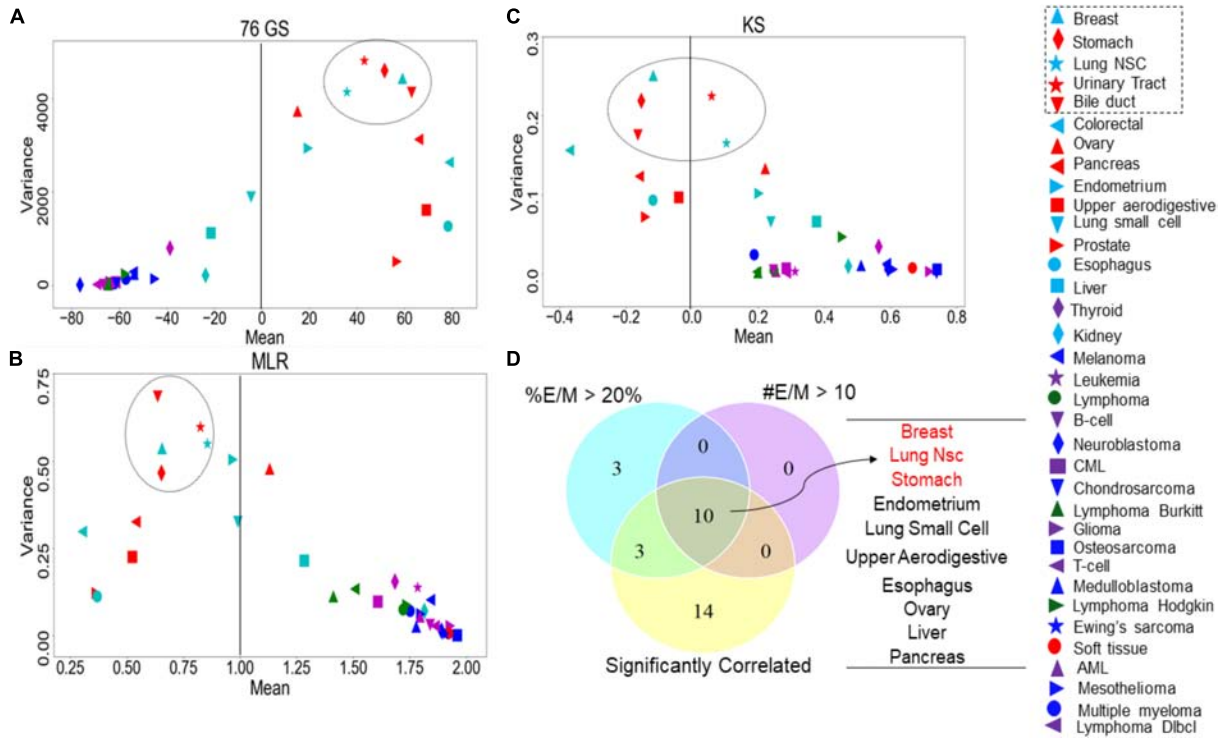


FIGURE 6 | Variance and mean of EMT scores in CCLE samples grouped by tumor subtype, highlighting the most variable tumor types (circled). **(A)** 76GS EMT scores, **(B)** MLR EMT scores, and **(C)** KS EMT scores. **(D)** Venn diagram showing the overlap between each tumor type based on the abundance of hybrid samples as defined by the MLR method, where #EM > 10 denote the cases where the absolute number of hybrid E/M samples in a tumor subtype is >10; %EM > 20 denote the cases where the percentage of cell lines identified as hybrid E/M in a given tumor subtype is >20%.

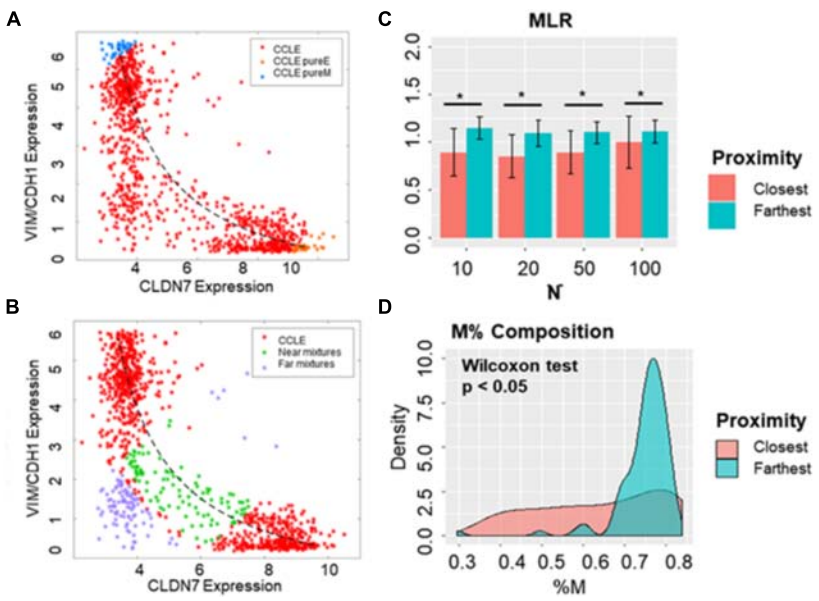


FIGURE 7 | Distinguishing between hybrid E/M cells vs. mixtures of E and M cells. **(A)** Scatter plot showing CCLE cell lines that display a hybrid E/M phenotype (red) on the mixture curve (dotted curve) determined by the mean of 35 pure E (orange) and pure M (blue) reference samples in CCLE dataset. **(B)** Scatter plot showing the 100 farthest (purple) and 100 closest (green) samples based on the distance from the mixture curve. **(C)** Bar plots showing EMT scores of N (10, 20, 50, and 100) closest and farthest hybrid E/M samples from mixture curve. **(D)** Mesenchymal proportion (%M) distribution of the 100 closest and farthest hybrid samples from mixture curve. *$p < 0.05$, N = 10, 20, 50 and 100, two-tailed Student's t-test; error bars represent standard deviation.

containing samples with overexpression and/or knockdown of many EMT/MET inducers such as TGFβ, Snail, Slug, Twist, E-cadherin, and GRHL2 (De Craene and Berx, 2013). The remarkable concordance among EMT scores calculated via the methods analyzed above suggests the existence of a macroscopic signal that can resolve the extent of EMT in a given sample amidst the complexity of EMT and the networks regulating it. It is plausible that within these regulatory networks, there exist key nodes forming one (or more) core circuit(s) which receive(s) a large number of inputs and may have diverse outputs, reminiscent of bow-tie structures seen in biological networks of cell-fate decision-making (Friedlander et al., 2015). This idea of core circuit(s) driving EMT is substantiated by transcriptomic meta-analysis identifying common signatures for EMT driven by distinct inducers (Taube et al., 2010; Liang et al., 2016). For instance, one network motif commonly found in core circuits regulating EMT and associated traits is a mutually inhibitory feedback loop between two "master regulators" driving opposing cell phenotypes (Hong et al., 2015; Huang et al., 2015; Saha et al., 2018); for instance, ZEB1 driving EMT and miR-200 driving MET (Jia et al., 2017). An intricate coupling among such feedback loops may give rise to a spectrum of EMT phenotypes as has been seen across cancer types in cell lines, CTCs, and primary tumor biopsies (Armstrong et al., 2011; Huang et al., 2013; Schliekelman et al., 2015; Andriani et al., 2016; Iyer et al., 2019; Markiewicz et al., 2019; Varankar et al., 2019).

In addition to EMT score concordance, the three methods showed excellent agreement in their ability to identify the most EMT-variable tumors. Most tumors of M lineage, including sarcoma samples, were shown to be least variable, as evidenced by the similarity among samples having M assignment in the CCLE dataset. This contrasts with breast cancer, NSCLC, bile duct cancer, urinary tract cancer, and stomach cancer, which exhibited the largest degree of variability in terms of their inherent EMT status in addition to being less M on average. The observations concerning the EMT status of sarcomas, breast cancer, and NSCLC are well-supported by existing experimental data (Blick et al., 2008; Schliekelman et al., 2015; Jolly et al., 2019b); however, the relationship between EMT status and heterogeneity among samples of a particular tumor type requires further investigation. Our results also demonstrate a link between the predominance of hybrid E/M status and heterogeneity patterns, possibly emerging due to relatively higher plasticity of cells in one or more hybrid E/M phenotypes (Pastushenko et al., 2018; Tripathi et al., 2020). Our findings are clinically relevant as tumor types having a greater number of hybrid E/M cells may require alternative treatment strategies compared to those containing predominantly E or predominantly M populations, necessitating future investigations into improved therapeutic design based on an analysis of EMT status and variability.

This comparative analysis of the three methods shows two key advantages of MLR method. First, it uses the least number of genes to calculate an EMT score – 23 genes required by MLR compared to 76 genes by 76GS, and 315 genes for tumor and 218 genes for cell lines by KS. This feature is important because 23 genes can be relatively easily measured experimentally without microarray or RNA-seq. Second, the MLR method, by virtue of its underlying theoretical framework, is capable of isolating hybrid E/M samples and has been expanded to identify whether the resultant gene expression is more likely to derive from "true" individual hybrid E/M samples or admixtures of E and M samples. While, in theory, other methods could adopt similar adaptations to address this issue in the future, the resolution of E, M, and hybrid E/M populations through those methods would require analyzing a higher dimensional subspace of the original predictors, given the large number of genes used by those methods to calculate EMT scores. This feature contrasts with that of MLR method, where the mixture analysis is performed directly on the two-dimensional EMT predictor space (CLDN7 and VIM/CDH1) utilized by this method. Distinguishing between these possibilities is critical because the behavior of mixtures of E and M samples vs. truly hybrid E/M samples can be strikingly different; a recent study showed that the presence of hybrid E/M cells is essential to form tumors in mice, a task which could not be achieved as efficiently by co-cultures of E and M cells alone (Kröger et al., 2019). Previously, multiple studies have implicated the role of hybrid E/M phenotype with worse survival (Grosse-Wilde et al., 2015; Grigore et al., 2016). To date, it has not been established whether it is pure hybrids or mixtures of E and M cells which correlate with clinically observed parameters. Our results highlight the utility of using the MLR method for effectively distinguishing between these two possibilities, and future work should address the relationship between the purity of hybrid E/M samples and clinical outcome.

Our analysis shown here suffers from following limitations. First, in terms of classifying hybrid E/M into "pure" hybrid E/M vs. mixtures of E and M subpopulations, we have considered mutually exclusive criteria: (a) a sample identified as hybrid E/M at a bulk level contains mixtures of E and M subpopulations, and (b) a sample identified as hybrid E/M at a bulk level contains all "true" hybrid E/M cells. However, many cell lines may contain cells in each of the three phenotypes in varying ratios (Ruscetti et al., 2016; George et al., 2017; Jia et al., 2019). Thus, future efforts should aim to identify the relative proportions of these three different phenotypes in a given sample. Second, although we show that among the samples identified to be lying closest vs. farthest from the "mixture curve" by MLR, all three EMT scoring metrics suggested that the ones lying closest to the curve are more E than the ones lying farthest from the same, we lack a clear biological interpretation of this observation. Future efforts will focus on comparing the morphological and functional behavior of the CCLE cell lines identified to be closest vs. farthest from the "mixture curve" generated based on the CCLE samples. Third, our current efforts focus on microarray data because the gene signatures utilized by all three methods were identified on this platform. Although the MLR method has been implemented on RNA-seq datasets by regressing the values obtained from microarray and RNA-seq analysis on a case-by-case basis (Kilinc et al., 2019; Lourenco et al., 2020), varying sensitivity of microarray and RNA-seq methods needs to be

incorporated for future efforts in assessing these EMT scoring methods systematically.

DATA AVAILABILITY STATEMENT

All codes used for the analysis in this article can be accessed through the following link: https://github.com/priyanka8993/EMT_score_calculation.

AUTHOR CONTRIBUTIONS

MJ conceived and oversaw the research. PC, ST, and JG conducted the research. All authors analyzed the data and wrote the manuscript.

FUNDING

This work was supported by the Ramanujan Fellowship (SB/S2/RJN-049/2018) awarded by the Science and Engineering Research Board (SERB), Department of Science and Technology (DST), Government of India awarded to MJ. HL was supported by the National Science Foundation (NSF) grants PHY-1427654 (Center for Theoretical Biological Physics) and PHY-1935762. JG was supported by the National Cancer Institute of NIH (F30CA213878).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fbioe.2020.00220/full#supplementary-material

FIGURE S1 | Scatter plot depicting the correlation between the EMT scores of cancer cell line samples, calculated via three EMT scoring methods. Each pairwise relation is estimated by a linear regression line (red), Spearman's correlation coefficient (*R*) and *p*-value (*p*) reported in each plot. **(A)** NCI60 dataset and **(B)** CCLE dataset.

FIGURE S2 | Scatter plot depicting the correlation between the EMT scores of different tumor types in TCGA dataset, calculated via three methods. Each pairwise relation is estimated by a linear regression line (red), Pearson's correlation coefficient (*R*), and *p*-value (*p*) reported in each plot. **(A)** Lung squamous cell cancer, **(B)** colon adenocarcinoma, and **(C)** colon and rectal adenocarcinoma.

FIGURE S3 | EMT score correlation with TGFβ-specific EMT scoring method in CCLE dataset. **(A)** Pearson's correlation coefficient and **(B)** Spearman's correlation. Correlation coefficient (*R*) and *p*-value (*p*) reported in each plot.

FIGURE S4 | EMT scores of different EMT time series datasets and CTCs. **(A)** GSE24202 – EMT induction by different EMT regulators. **(B)** GSE84002 – EMT and MET induction over time by GFP, SNAI1 and SNAI2. **(C)** GSE43489 – EMT/MET induction in PC3 cell line. **(D)** GSE17708 – EMT induction over time. **(E)** GSE55470 – CTCs from breast cancer patients. **(F)** GSE50991 – CTCs from *ex vivo* lung cancer model (*$p < 0.05$, $n = 3$, two-tailed Student's *t*-test; error bars represent standard deviation for $n = 3$). Graphs **(E)** and **(F)** represent kernel density plots.

FIGURE S5 | (A) Scatter plot showing 100 farthest and closest samples based on the distance from mixture curve defined by median of 35 most pure E and pure M CCLE samples. **(B)** MLR EMT score for *N* (10,20,50,100) closest and farthest hybrid samples from median mixture curve. Bar plots showing EMT scores of *N* (10, 20, 50, 100) closest and farthest hybrid samples from mean mixture curve. **(C)** 76GS EMT score and **(D)** KS EMT score. Bar plots showing EMT scores of *N* (10, 20, 50, 100) closest and farthest hybrid samples from median mixture curve. **(E)** 76GS EMT score **(F)** KS EMT score (*$p < 0.05$, $N = 10$, 20, 50, and 100, two-tailed Student's *t*-test; error bars represent standard deviation for the given value of *N*).

TABLE S1 | 76 gene signatures.

TABLE S2 | List of predictors and normalizers used for calculation of EMT using MLR method.

TABLE S3 | Epithelial and mesenchymal signature used in KS-statistic (tumor signature).

TABLE S4 | Epithelial and mesenchymal signature used in KS-statistic (cell line signature).

TABLE S5 | EMT score correlation in the list of 85 microarray GEO datasets.

TABLE S6 | EMT scores in E, E/M and M categories of CCLE samples, as defined by MLR EMT scores. **(A)** Mean and standard deviation of EMT scores in E, E/M and M samples. **(B)** Correlation between EMT scores across E, E/M and M categories.

TABLE S7 | Most variable and least variable tumor types based on the coefficient of variation of EMT scores.

TABLE S8 | Pairwise correlation between all three EMT scores in subcategories (E, E/M, and M) across all tumor types of CCLE data.

TABLE S9 | Abundance of hybrid E/M samples in different tumor types.

REFERENCES

Aiello, N. M., Maddipati, R., Norgard, R. J., Balli, D., Li, J., Yuan, S., et al. (2018). EMT subtype influences epithelial plasticity and mode of cell migration. *Dev. Cell* 45, 681.e4–695.e4. doi: 10.1016/J.DEVCEL.2018.05.027

Andriani, F., Bertolini, G., Facchinetti, F., Baldoli, E., Moro, M., Casalini, P., et al. (2016). Conversion to stem-cell state in response to microenvironmental cues is regulated by balance between epithelial and mesenchymal features in lung cancer cells. *Mol. Oncol.* 10, 253–271. doi: 10.1016/j.molonc.2015.10.002

Armstrong, A. J., Marengo, M. S., Oltean, S., Kemeny, G., Bitting, R. L., Turnbull, J. D., et al. (2011). Circulating tumor cells from patients with advanced prostate and breast cancer display both epithelial and mesenchymal markers. *Mol. Cancer Res.* 9, 997–1007. doi: 10.1158/1541-7786.MCR-10-0490

Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A. A., Kim, S., et al. (2012). The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* 483, 603–607. doi: 10.1038/nature11003

Biddle, A., Gammon, L., Liang, X., Costea, D. E., and Mackenzie, I. C. (2016). Phenotypic plasticity determines cancer stem cell therapeutic resistance in oral squamous cell carcinoma. *EBioMedicine* 4, 138–145. doi: 10.1016/j.ebiom.2016.01.007

Bierie, B., Pierce, S. E., Kroeger, C., Stover, D. G., Pattabiraman, D. R., Thiru, P., et al. (2017). Integrin-β4 identifies cancer stem cell-enriched populations of partially mesenchymal carcinoma cells. *Proc. Natl. Acad. Sci. U.S.A.* 114, E2337–E2346. doi: 10.1073/pnas.1618298114

Blick, T., Widodo, E., Hugo, H., Waltham, M., Lenburg, M. E., Neve, R. M., et al. (2008). Epithelial mesenchymal transition traits in human breast

cancer cell lines. *Clin. Exp. Metastasis* 25, 629–642. doi: 10.1007/s10585-008-9170-6

Byers, L. A., Diao, L., Wang, J., Saintigny, P., Girard, L., Peyton, M., et al. (2013). An epithelial-mesenchymal transition gene signature predicts resistance to EGFR and PI3K inhibitors and identifies Axl as a therapeutic target for overcoming EGFR inhibitor resistance. *Clin. Cancer Res.* 19, 279–290. doi: 10.1158/1078-0432.CCR-12-1558

Chikaishi, Y., Uramoto, H., and Tanaka, F. (2011). The EMT status in the primary tumor does not predict postoperative recurrence or disease-free survival in lung adenocarcinoma. *Anticancer Res.* 31, 4451–4456.

Cook, D. P., and Vanderhyden, B. C. (2019). Comparing transcriptional dynamics of the epithelial-mesenchymal transition. *bioRxiv* [Preprint]. doi: 10.1101/732412

Davis, S., and Meltzer, P. S. (2007). GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. *Bioinformatics* 23, 1846–1847. doi: 10.1093/bioinformatics/btm254

De Craene, B., and Berx, G. (2013). Regulatory networks defining EMT during cancer initiation and progression. *Nat. Rev. Cancer* 13, 97–110. doi: 10.1038/nrc3447

Devaraj, V., and Bose, B. (2019). Morphological state transition dynamics in EGF-induced epithelial to mesenchymal transition. *J. Clin. Med.* 8:911. doi: 10.3390/jcm8070911

Foroutan, M., Cursons, J., Hediyeh-Zadeh, S., Thompson, E. W., and Davis, M. J. (2017). A Transcriptional program for detecting TGFβ-induced EMT in cancer. *Mol. Cancer Res.* 15, 619–631.

Friedlander, T., Mayo, A. E., Tlusty, T., and Alon, U. (2015). Evolution of bow-tie architectures in biology. *PLoS Comput. Biol.* 11:e1004055. doi: 10.1371/journal.pcbi.1004055

George, J. T., Jolly, M. K., Xu, S., Somarelli, J. A., and Levine, H. (2017). Survival outcomes in cancer patients predicted by a partial EMT gene expression scoring metric. *Cancer Res.* 77, 6415–6428. doi: 10.1158/0008-5472.CAN-16-3521

Grigore, A., Jolly, M. K., Jia, D., Farach-Carson, M., and Levine, H. (2016). Tumor budding: the name is EMT. partial EMT. *J. Clin. Med.* 5:51. doi: 10.3390/jcm5050051

Grosse-Wilde, A., Fouquier d' Herouel, A., McIntosh, E., Ertaylan, G., Skupin, A., Kuestner, R. E., et al. (2015). Stemness of the hybrid epithelial/mesenchymal state in breast cancer and its association with poor survival. *PLoS One* 10:e0126522. doi: 10.1371/journal.pone.0126522

Guo, C. C., Majewski, T., Zhang, L., Yao, H., Bondaruk, J., Wang, Y., et al. (2019). Dysregulation of EMT drives the progression to clinically aggressive sarcomatoid bladder cancer. *Cell Rep.* 27, 1781.e4–1793.e4. doi: 10.1016/j.celrep.2019.04.048

Hong, T., Watanabe, K., Ta, C. H., Villarreal-Ponce, A., Nie, Q., and Dai, X. (2015). An Ovol2-Zeb1 mutual inhibitory circuit governs bidirectional and multi-step transition between epithelial and mesenchymal states. *PLoS Comput. Biol.* 11:e1004569. doi: 10.1371/journal.pcbi.1004569

Huang, B., Jolly, M. K., Lu, M., Tsarfaty, I., Ben-Jacob, E., and Onuchic, J. N. (2015). Modeling the transitions between collective and solitary migration phenotypes in cancer metastasis. *Sci. Rep.* 5:17379. doi: 10.1038/srep17379

Huang, R. Y.-J., Wong, M. K., Tan, T. Z., Kuay, K. T., Ng, A. H., Chung, V. Y., et al. (2013). An EMT spectrum defines an anoikis-resistant and spheroidogenic intermediate mesenchymal state that is sensitive to e-cadherin restoration by a src-kinase inhibitor, saracatinib (AZD0530). *Cell Death Dis.* 4:e915. doi: 10.1038/cddis.2013.442

Iyer, A., Gupta, K., Sharma, S., Hari, K., Lee, Y. F., Ramalingan, N., et al. (2019). Integrative analysis and machine learning based characterization of single circulating tumor cells. *bioRxiv* [Preprint]. doi: 10.1101/867200

Jia, D., George, J. T., Tripathi, S. C., Kundnani, D. L., Lu, M., Hanash, S. M., et al. (2019). Testing the gene expression classification of the EMT spectrum. *Phys. Biol.* 16:025002. doi: 10.1088/1478-3975/aaf8d4

Jia, D., Jolly, M. K., Tripathi, S. C., Den Hollander, P., Huang, B., Lu, M., et al. (2017). Distinguishing mechanisms underlying EMT tristability. *Cancer Converg.* 1:2. doi: 10.1101/098962

Jolly, M. K., and Celia-Terrassa, T. (2019). Dynamics of phenotypic heterogeneity associated with EMT and stemness during cancer progression. *J. Clin. Med.* 8:1452. doi: 10.3390/jcm8101542

Jolly, M. K., Huang, B., Lu, M., Mani, S. A., Levine, H., and Ben-Jacob, E. (2014). Towards elucidating the connection between epithelial-mesenchymal transitions and stemness. *J. R. Soc. Interface* 11:20140962. doi: 10.1098/rsif.2014.0962

Jolly, M. K., Somarelli, J. A., Sheth, M., Biddle, A., Tripathi, S. C., Armstrong, A. J., et al. (2019a). Hybrid epithelial/mesenchymal phenotypes promote metastasis and therapy resistance across carcinomas. *Pharmacol. Ther.* 194, 161–184. doi: 10.1016/j.pharmthera.2018.09.007

Jolly, M. K., Ware, K. E., Gilja, S., Somarelli, J. A., and Levine, H. (2017). EMT and MET: necessary or permissive for metastasis? *Mol. Oncol.* 11, 755–769. doi: 10.1002/1878-0261.12083

Jolly, M. K., Ware, K. E., Xu, S., Gilja, S., Shetler, S., Yang, Y., et al. (2019b). E-cadherin represses anchorage-independent growth in sarcomas through both signaling and mechanical mechanisms. *Mol. Cancer Res.* 17, 1391–1402. doi: 10.1158/1541-7786.MCR-18-0763

Karacosta, L. G., Anchang, B., Ignatiadis, N., Kimmey, S. C., Benson, J. A., Shrager, J. B., et al. (2019). Mapping lung cancer epithelial-mesenchymal transition states and trajectories with single-cell resolution. *Nat. Commun.* 10:5587. doi: 10.1101/570341

Katsuno, Y., Meyer, D. S., Zhang, Z., Shokat, K. M., Akhurst, R. J., Miyazono, K., et al. (2019). Chronic TGF-β exposure drives stabilized EMT, tumor stemness, and cancer drug resistance with vulnerability to bitopic mTOR inhibition. *Sci. Signal.* 12:eaau8544. doi: 10.1126/scisignal.aau8544

Kilinc, A. N., Sugiyama, N., Reddy Kalathur, R. K., Antoniadis, H., Birogul, H., Ishay-Ronen, D., et al. (2019). Histone deacetylases, Mbd3/NuRD, and Tet2 hydroxylase are crucial regulators of epithelial–mesenchymal plasticity and tumor metastasis. *Oncogene* 39, 1498–1513.

Kröger, C., Afeyan, A., Mraz, J., Eaton, E. N., Reinhardt, F., Khodor, Y. L., et al. (2019). Acquisition of a hybrid E/M state is essential for tumorigenicity of basal breast cancer cells. *Proc. Natl. Acad. Sci. U.S.A.* 116, 7353–7362. doi: 10.1073/pnas.1812876116

Kurrey, N. K., Ghanate, A. D., Chaskar, P. D., Doiphode, R. Y., and Bapat, S. A. (2009). Snail and slug mediate radioresistance and chemoresistance by antagonizing p53-mediated apoptosis and acquiring a stem-like phenotype in ovarian cancer cells. *Stem Cells* 27, 2059–2068. doi: 10.1002/stem.154

Lam, V., Nguyen, T., Bui, V., Chung, B. M., Chang, L. C., Nehmetallah, G., et al. (2020). Quantitative scoring of epithelial and mesenchymal qualities of cancer cells using machine learning and quantitative phase imaging. *J. Biomed. Opt.* 25, 1–17. doi: 10.1117/1.JBO.25.2.026002

Lawson, D. A., Kessenbrock, K., Davis, R. T., Pervolarakis, N., and Werb, Z. (2018). Tumour heterogeneity and metastasis at single-cell resolution. *Nat. Cell Biol.* 20, 1349–1360. doi: 10.1038/s41556-018-0236-7

Liang, L., Sun, H., Zhang, W., Zhang, M., Yang, X., Kuang, R., et al. (2016). Meta-analysis of EMT datasets reveals different types of EMT. *PLoS One* 11:e0156839. doi: 10.1371/journal.pone.0156839

Lourenco, A. R., Ban, Y., Crowley, M. J., Lee, S. B., Ramchandani, D., Du, W., et al. (2020). Differential contributions of pre- and post-EMT tumor cells in breast cancer metastasis. *Cancer Res.* 80, 163–169. doi: 10.1158/0008-5472.CAN-19-1427

Mandal, M., Ghosh, B., Anura, A., Mitra, P., Pathak, T., and Chatterjee, J. (2016). Modeling continuum of epithelial mesenchymal transition plasticity. *Integr. Biol.* 8, 167–176. doi: 10.1039/C5IB00219B

Markiewicz, A., Topa, J., Nagel, A., Skokowski, J., Seroczynska, B., Stokowy, T., et al. (2019). Spectrum of epithelial-mesenchymal transition phenotypes in circulating tumour cells from early breast cancer patients. *Cancers (Basel).* 11:E59. doi: 10.3390/cancers11010059

Panchy, N., Azeredo-Tseng, C., Luo, M., Randall, N., and Hong, T. (2020). Integrative transcriptomic analysis reveals a multiphasic epithelial–mesenchymal spectrum in cancer and non-tumorigenic cells. *Front. Oncol.* 9:1479. doi: 10.3389/fonc.2019.01479

Pastushenko, I., and Blanpain, C. (2019). EMT transition states during tumor progression and metastasis. *Trends Cell Biol.* 29, 212–226. doi: 10.1016/j.tcb.2018.12.001

Pastushenko, I., Brisebarre, A., Sifrim, A., Fioramonti, M., Revenco, T., Boumahdi, S., et al. (2018). Identification of the tumour transition states occurring during EMT. *Nature* 556, 463–468. doi: 10.1038/s41586-018-0040-3

Puram, S. V., Tirosh, I., Parikh, A. S., Patel, A. P., Yizhak, K., Gillespie, S., et al. (2017). Single-cell transcriptomic analysis of primary and metastatic tumor ecosystems in head and neck cancer. *Cell* 171, 1611–1624. doi: 10.1016/j.cell.2017.10.044

Roca, H., Hernandez, J., Weidner, S., McEachin, R. C., Fuller, D., Sud, S., et al. (2013). Transcription factors OVOL1 and OVOL2 induce the mesenchymal to epithelial transition in human cancer. *PLoS One* 8:e76773. doi: 10.1371/journal.pone.0076773

Ruscetti, M., Dadashian, E. L., Guo, W., Quach, B., Mulholland, D. J., Park, J. W., et al. (2016). HDAC inhibition impedes epithelial-mesenchymal plasticity and suppresses metastatic, castration-resistant prostate cancer. *Oncogene* 35, 3781–3795. doi: 10.1038/onc.2015.444

Saha, M., Kumar, S., Bukhari, S., Balaji, S., Kumar, P., Hindupur, S., et al. (2018). AMPK-Akt double-negative feedback loop in breast cancer cells regulates their adaptation to matrix deprivation. *Cancer Res.* 78, 1497–1510. doi: 10.1158/0008-5472.CAN-17-2090

Saxena, K., Subbalakshmi, A. R., and Jolly, M. K. (2019). Phenotypic heterogeneity in circulating tumor cells and its prognostic value in metastasis and overall survival. *EBioMedicine* 46, 4–5. doi: 10.1016/j.ebiom.2019.07.074

Schliekelman, M. J., Taguchi, A., Zhu, J., Dai, X., Rodriguez, J., Celiktas, M., et al. (2015). Molecular portraits of epithelial, mesenchymal, and hybrid states in lung adenocarcinoma and their relevance to survival. *Cancer Res.* 75, 1789–1800. doi: 10.1158/0008-5472.CAN-14-2535

Shankavaram, U. T., Varma, S., Kane, D., Sunshine, M., Chary, K. K., Reinhold, W. C., et al. (2009). CellMiner: a relational database and query tool for the NCI-60 cancer cell lines. *BMC Genomics* 10:277. doi: 10.1186/1471-2164-10-277

Shen, Y., Schmidt, B. U. S., Kubitschke, H., Morawetz, E. W., Wolf, B., Käs, J. A., et al. (2020). Detecting heterogeneity in and between breast cancer cell lines. *Cancer Converg* 4:1. doi: 10.1186/s41236-020-0010-1

Stylianou, N., Lehman, M. L., Wang, C., Fard, A. T., Rockstroh, A., Fazli, L., et al. (2018). A molecular portrait of epithelial–mesenchymal plasticity in prostate cancer associated with clinical outcome. *Oncogene* 38, 913–934.

Tan, T. Z., Miow, Q. H., Miki, Y., Noda, T., Mori, S., Huang, R. Y., et al. (2014). Epithelial-mesenchymal transition spectrum quantification and its efficacy in deciphering survival and drug responses of cancer patients. *EMBO Mol. Med.* 6, 1279–1293. doi: 10.15252/emmm.201404208

Taube, J. H., Herschkowitz, J. I., Komurov, K., Zhou, A. Y., Gupta, S., Yang, J., et al. (2010). Core epithelial-to-mesenchymal transition interactome gene-expression signature is associated with claudin-low and metaplastic breast cancer subtypes. *Proc. Natl. Acad. Sci. U.S.A.* 107, 15449–15454. doi: 10.1073/pnas.1004900107

Terry, S., Savagner, P., Ortiz-Cuaran, S., Mahjoubi, L., Saintigny, P., Thiery, J. P., et al. (2017). New insights into the role of EMT in tumor immune escape. *Mol. Oncol.* 11, 824–846. doi: 10.1002/1878-0261.12093

Thomson, T. M., Balcells, C., and Cascante, M. (2019). Metabolic plasticity and epithelial-mesenchymal transition. *J. Clin. Med.* 8:967. doi: 10.3390/jcm8070967

Tripathi, S., Chakraborty, P., Levine, H., and Jolly, M. K. (2020). A mechanism for epithelial-mesenchymal heterogeneity in a population of cancer cells. *PLoS Comput. Biol.* 16:e1007619. doi: 10.1101/592691

Tripathi, S., Kessler, D. A., and Levine, H. (2019). Biological regulatory networks are minimally frustrated. *arXiv* [Preprint]. Available online at: https://arxiv.org/abs/1911.10252 (accessed September 15, 2019).

Tripathi, S. C., Peters, H. L., Taguchi, A., Katayama, H., Wang, H., Momin, A., et al. (2016). Immunoproteasome deficiency is a feature of non-small cell lung cancer with a mesenchymal phenotype and is associated with a poor outcome. *Proc. Natl. Acad. Sci. U.S.A.* 113, E1555–E1564. doi: 10.1073/pnas.1521812113

Van Zijl, F., Mall, S., Machat, G., Pirker, C., Zeillinger, R., Weinhaeusel, A., et al. (2011). A human model of epithelial to mesenchymal transition to monitor drug efficacy in hepatocellular carcinoma progression. *Mol. Cancer Ther.* 10, 850–860. doi: 10.1158/1535-7163.MCT-10-0917

Varankar, S. S., Kamble, S. S., Mali, A. M., More, M. M., Abraham, A., Kumar, B., et al. (2019). Functional balance between TCF21-Slug defines cellular plasticity and sub-classes in high-grade serous ovarian cancer. *Carcinogenesis* doi: 10.1093/carcin/bgz119 [Epub ahead of print].

Wang, S., Zhang, J., He, Z., Wu, K., and Liu, X. S. (2019). The predictive power of tumor mutational burden in lung cancer immunotherapy response is influenced by patients' sex. *Int. J. Cancer* 145, 2840–2849. doi: 10.1002/ijc.32327

Wang, W., Douglas, D., Zhang, J., Chen, Y.-J., Cheng, Y.-Y., Kumari, S., et al. (2019). M-TRACK: a platform for live cell multiplex imaging reveals cell phenotypic transition dynamics inherently missing in snapshot data. *bioRxiv* [Preprint]. doi: 10.1101/2019.12.12.874248

Watanabe, K., Panchy, N., Noguchi, S., Suzuki, H., and Hong, T. (2019). Combinatorial perturbation analysis reveals divergent regulations of mesenchymal genes during epithelial-to-mesenchymal transition. *npj Syst. Biol. Appl.* 5:21. doi: 10.1038/s41540-019-0097-0

Yan, S., Holderness, B. M., Li, Z., Seidel, G. D., Gui, J., Fisher, J. L., et al. (2016). Epithelial-mesenchymal expression phenotype of primary melanoma and matched metastases and relationship with overall survival. *Anticancer Res.* 36, 6449–6456. doi: 10.21873/anticanres.11243

Yu, M., Bardia, A., Wittner, B. S., Stott, S. L., Smas, M. E., Ting, D. T., et al. (2013). Circulating breast tumor cells exhibit dynamic changes in epithelial and mesenchymal composition. *Science* 339, 580–584. doi: 10.1126/science.1228522

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 Bioengineering and Biotechnology

ORIGINAL RESEARCH
published: 30 April 2020
doi: 10.3389/fbioe.2020.00276

Dysregulation of Signaling Pathways Due to Differentially Expressed Genes From the B-Cell Transcriptomes of Systemic Lupus Erythematosus Patients – A Bioinformatics Approach

S. Udhaya Kumar[1], D. Thirumal Kumar[1], R. Siva[1], C. George Priya Doss[1], Salma Younes[2], Nadin Younes[2], Mariem Sidenna[2] and Hatem Zayed[2]**

[1] School of Biosciences and Technology, Vellore Institute of Technology, Vellore, India, [2] Department of Biomedical Sciences, College of Health and Sciences, QU Health, Qatar University, Doha, Qatar

OPEN ACCESS

Edited by:
Manoj Kumar Kashyap,
Amity University Gurgaon, India

Reviewed by:
Jian Yu,
Beihang University, China
Noor Ahmad Shaik,
King Abdulaziz University,
Saudi Arabia

***Correspondence:**
C. George Priya Doss
georgepriyadoss@vit.ac.in
Hatem Zayed
hatem.zayed@qu.edu.qa

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Bioengineering and Biotechnology

Received: 06 January 2020
Accepted: 16 March 2020
Published: 30 April 2020

Citation:
Udhaya Kumar S, Thirumal Kumar D, Siva R, George Priya Doss C, Younes S, Younes N, Sidenna M and Zayed H (2020) Dysregulation of Signaling Pathways Due to Differentially Expressed Genes From the B-Cell Transcriptomes of Systemic Lupus Erythematosus Patients – A Bioinformatics Approach. Front. Bioeng. Biotechnol. 8:276. doi: 10.3389/fbioe.2020.00276

Systemic lupus erythematosus (SLE) is an autoimmune inflammatory disorder that is clinically complex and has increased production of autoantibodies. Via emerging technologies, researchers have identified genetic variants, expression profiling of genes, animal models, and epigenetic findings that have paved the way for a better understanding of the molecular and genetic mechanisms of SLE. Our current study aimed to illustrate the essential genes and molecular pathways that are potentially involved in the pathogenesis of SLE. This study incorporates the gene expression profiling data of the microarray dataset GSE30153 from the Gene Expression Omnibus (GEO) database, and differentially expressed genes (DEGs) between the B-cell transcriptomes of SLE patients and healthy controls were screened using the GEO2R web tool. The identified DEGs were subjected to STRING analysis and Cytoscape to explore the protein–protein interaction (PPI) networks between them. The MCODE (Molecular Complex Detection) plugin of Cytoscape was used to screen the cluster subnetworks that are highly interlinked between the DEGs. Subsequently, the clustered DEGs were subjected to functional annotation with ClueGO/CluePedia to identify the significant pathways that were enriched. For integrative analysis, we used GeneGo Metacore™, a Cortellis Solution software, to exhibit the Gene Ontology (GO) and enriched pathways between the datasets. Our study identified 4 upregulated and 13 downregulated genes. Analysis of GO and functional enrichment using ClueGO revealed the pathways that were statistically significant, including pathways involving T-cell costimulation, lymphocyte costimulation, negative regulation of vascular permeability, and B-cell receptor signaling. The DEGs were mainly enriched in metabolic networks such as the phosphatidylinositol-3,4,5-triphosphate pathway and the carnitine pathway. Additionally, potentially enriched pathways, such as the signaling pathways induced by oxidative stress and reactive oxygen species (ROS), chemotaxis and lysophosphatidic acid signaling induced via G protein-coupled receptors (GPCRs), and the androgen

receptor activation pathway, were identified from the DEGs that were mainly associated with the immune system. Four genes (*EGR1*, *CD38*, *CAV1*, and *AKT1*) were identified to be strongly associated with SLE. Our integrative analysis using a multitude of bioinformatics tools might promote an understanding of the dysregulated pathways that are associated with SLE development and progression. The four DEGs in SLE patients might shed light on the pathogenesis of SLE and might serve as potential biomarkers in early diagnosis and as therapeutic targets for SLE.

Keywords: systemic lupus erythematosus, protein–protein interactions, Metacore, microarray and bioinformatics, expression profiling data, biomarkers, functional enrichment analysis

INTRODUCTION

Systemic lupus erythematosus (SLE), also known as lupus, is a rare systemic autoimmune disease that mostly affects middle-aged women, mainly of Asian, African, American, and Hispanic origin (Costa-Reis and Sullivan, 2013; Cui et al., 2013; Gurevitz et al., 2013). SLE affects an estimated 5 million people across the world, with an incidence of 1–10 per 100,000 person-years (Pons-Estel et al., 2010). SLE is characterized by a wide range of different autoantibodies, deposition of immune complexes, and immune system infiltration and inflammation within damaged organs. SLE autoantibodies invade the patient's kidneys, heart, skin, joints, and brain, leading to various typical clinical symptoms. The most common clinical symptoms of lupus are rash, arthritis, and fatigue. Severe complications of SLE lead to nephritis, anemia, neurological symptoms, and thrombocytopenia, eventually leading to severe morbidity and mortality.

SLE is characterized by its clinical heterogeneity, with a wide range of clinical manifestations reflecting its complex etiopathogenesis (Tan et al., 1982). The clinical heterogeneity of SLE highlights the contribution of genetic and environmental factors to the susceptibility to the disease (Prokunina and Alarcon-Riquelme, 2004; Harley et al., 2009; Yang and Lau, 2015; Dang et al., 2016; Wang et al., 2017). To date, the reason for phenotypic variation in SLE is unknown. Understanding the molecular mechanisms behind the pathogenesis of SLE phenotypes could help in developing more efficient therapeutic approaches and preventive strategies.

With the extensive use of gene detection methods, high-throughput sequencing and extensive microarray data profiling studies on SLE have been conducted, and several differentially expressed genes (DEGs) and cellular pathways in SLE have been identified (Borrebaeck et al., 2014; Zhu et al., 2015). Nevertheless, until now, no particular gene has been recognized to act as a potential marker for the diagnosis of SLE. In addition, a large amount of data obtained from microarray technology and high-throughput sequencing have not been fully used. Ducreux et al. (2016) collected blood samples from SLE patients and healthy volunteers to identify differentially expressed genes (Ducreux et al., 2016). However, the interactions among differentially expressed genes and key genes involved in the signaling pathways of SLE remain to be elucidated. In addition, previous studies of genetic factors primarily focused on single genes; nevertheless, interactions among multiple genes may result in the multisystem invasion characteristics observed in SLE (Smith et al., 2017). Remarkably, studies have shown that disease—associated gene expression networks have a potential role in the immune response, which highlights their mechanism and therapeutic value for SLE (Deng and Tsao, 2010; Bentham et al., 2015).

Integrating and reanalyzing the data using bioinformatics methods may help in identifying gene regulatory pathways, essential genes, and their associated networks in SLE disease, which can provide new and valuable ideas for understanding the molecular mechanisms and identifying reliable diagnostic and therapeutic targets of SLE. Therefore, in this study, we first conducted a comprehensive collection of genes associated with SLE from the GEO dataset with ID GSE30153. Then, we performed a bioinformatics analysis of these genes with the MCODE (Molecular Complex Detection), GeneGo, and ClueGO tools. To further explore the pathogenesis of SLE in a more specific manner, functions and pathways identified by the modules were used to indicate the biological processes and biochemical pathways related to the immune system. Finally, the genes potentially associated with arthritis, pleurisy, and myocarditis, which are the common complications of SLE, were compared with SLE-related genes to identify common genes that participated in the development of SLE. To interpret the biological relevance of these changes in gene expression, we analyzed the microarray data via an integrated bioinformatic analysis expanding on traditional microarray analysis methods, namely, Gene Ontology (GO) and pathway analysis, thereby allowing the construction of interaction networks that might identify novel prognostic markers and therapeutic targets.

MATERIALS AND METHODS

Acquisition of Array Data and Processing

Gene expression profiling data from microarray array analysis of the GSE30153 dataset were downloaded from the NCBI GEO database (Gene Expression Omnibus database)[1]. The database accommodates gene expression datasets from a variety of experiments, such as DNA-seq, ChIPs, RNA-seq, microarray, and high-throughput hybridization array (Edgar et al., 2002; Barrett et al., 2013). GSE30153 contains 26 samples, including 17 patients

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

with SLE and 9 healthy controls of human sorted B-cells obtained by using the platform GPL570 (HG-U133_Plus_2) Affymetrix Human Genome U133 Plus 2.0 Array (Garaud et al., 2011). The downloaded gene expression profiling data are freely available in the public database, and there were no human or animal experiments conducted by any of the authors in this study.

Preprocessing of Data and DEG Identification

Using the robust multiarray standard model, the initial information from the dataset was subjected to quantile normalization, background correction, and log transition (Irizarry et al., 2003). Preprocessing included changing to gene symbols from probe IDs using the Gene ID converter from Entrez (Alibés et al., 2007). The statistical online tool GEO2R uses the R/Bioconductor, and limma package v3.26.8 was used to screen the raw gene expression data (Smyth, 2005; Barrett et al., 2013; Ritchie et al., 2015). We performed a Benjamini–Hochberg test (to determine the false discovery rate) and *T*-tests to compute the false discovery rate (FDR) and *p*-values to identify the DEGs between SLE patients and healthy control human sorted B-cells (Benjamini and Hochberg, 1995; Aubert et al., 2004). We set the primary criteria of $|\log(2\text{ fold change})| > 1$ and $p < 0.05$ to obtain significant DEGs from the dataset, whereas cutoffs of $\log2FC \geq 1$ and $\log2FC \leq -1$ were used to denote upregulated and downregulated DEGs, respectively. For high-throughput sequencing, a logarithm to base 2 is widely used and in the initial scaling, the doubling is equivalent to a log2FC of 1 (Love et al., 2014). A volcano plot was constructed using a web-based tool[2]. The resulting DEGs were used for further analysis.

Constructing PPI Networks

To assess the relationships between the DEGs from the GSE30153 dataset, we constructed a protein–protein interaction (PPI) network by using Search Tool for the Retrieval of Interacting Genes (STRING v11.0)[3] (Szklarczyk et al., 2017, 2019). The cutoff criterion was set to a high confident interaction score of ≥0.7 to eliminate inconsistent PPIs from the dataset. We then incorporated the results from the STRING database into Cytoscape software (v3.7.2)[4] to envisage the PPIs within the statistically relevant DEGs (Shannon et al., 2003). The MCODE plugin from Cytoscape was utilized to identify the interconnected regions or clusters from the PPI network. The cluster finding parameters were adopted, such as a degree cutoff of 2, a node score cutoff of 0.2, a kappa score (K-core) of 5, and a max depth of 100, which limits the cluster size for coexpressing networks (Bader and Hogue, 2003). The top clusters from MCODE were subjected to ClueGO v2.5.5/CluePedia v1.5.5 analysis to obtain comprehensive GO and pathway results from the PPI network. ClueGO combines GO and pathway analyses from KEGG and BioCarta and provides a fundamentally structured GO or pathway network from the PPI network (Bindea et al., 2009).

[2]https://paolo.shinyapps.io/ShinyVolcanoPlot/

[3]http://www.string-db.org/

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

Metacore GeneGo Analysis of DEGs

Metacore, a Cortellis Solution software (Clarivate Analytics, London, United Kingdom)[5], was used to perform curated pathway enrichment analysis and GO analysis. GeneGo facilitates the rapid assessment of metabolic pathways, protein biological networks, and pathway maps from high-throughput experimental data (MetaCoreLogin | Clarivate Analytics). Based on a significance threshold of $p < 0.05$, a pictorial representation of the molecular interactions of DEGs from the study groups is generated. Determination of a hypergeometric *p*-value enables the estimation of the chance that an intersection between DEGs and ontological elements is random. An FDR < 0.05 was used as a criterion to calculate if statistically significant DEGs constituted a processor pathway.

RESULTS

Identification of DEGs From the Dataset

Our study contained the gene expression profiles of the GSE30153 dataset from the GEO database, which were submitted by Garaud et al. (2011) based on analysis with the GPL570 platform (Affymetrix Human Genome U133 Plus 2.0 Array) (Garaud et al., 2011). The dataset encompasses 26 samples, including 17 patients with SLE and 9 healthy controls (**Table 1**). By utilizing the GEO2R online tool, we obtained the differentially expressed genes (DEGs) from the GSE30153 dataset by comparing the SLE samples with control samples. By calculating *p*-values and | log2FC | values, the top 250 DEGs were identified. A volcano plot was constructed using the Rstudio web server ShinyVolcanoPlot to identify DEGs by comparing the SLE and control groups from the dataset. The volcano plot in **Figure 1** depicts all the DEGs with a log2FC against the –log10 (*p*-value) between the two groups. With cutoffs of $p < 0.05$ and $\log2FC \geq 1.0$ or ≤ -1, we found 4 and 13 genes that were upregulated and downregulated, respectively, between the two groups (**Table 2**). The genes that were differentially expressed between the two groups are shown in **Supplementary Table S1**.

Screening of Module and Construction of Interlinking PPI Network

To assess the protein–protein connections among the DEGs, we used the STRING tool to compute the protein interactions and plotted them using Cytoscape v3.7.2. **Figure 2** depicts the PPI network with 103 nodes and 201 edges. The DEGs are represented as nodes, and the edges are interactions between the DEGs. A combined node score of >0.4 was considered to be significant. MCODE plugin v1.5.1 from Cytoscape was utilized to identify the densely interlinked regions within the protein network. As a result, we obtained the top two significant clusters from the DEGs protein network with MCODE scores of 5.043 and 3.625. A graphical representation of these clusters is shown in **Figures 3A,B**. The subnetworks, scores, number of nodes and edges, and node IDs are tabulated in **Table 3**.

[5]https://clarivate.com/products/metacore/

TABLE 1 | The primary characteristics of 26 studies in GSE30153 procured from the Gene Omnibus Expression database.

Group	Accession	Title	Organism	Disease state	Tissue	Cell type
Patient	GSM746726	Patient 1	Homo sapiens	Systemic lupus erythematosus (SLE)	Blood	Human sorted B cell
	GSM746727	Patient 2	Homo sapiens	Systemic lupus erythematosus (SLE)	Blood	Human sorted B cell
	GSM746728	Patient 3	Homo sapiens	Systemic lupus erythematosus (SLE)	Blood	Human sorted B cell
	GSM746729	Patient 4	Homo sapiens	Systemic lupus erythematosus (SLE)	Blood	Human sorted B cell
	GSM746730	Patient 5	Homo sapiens	Systemic lupus erythematosus (SLE)	Blood	Human sorted B cell
	GSM746731	Patient 6	Homo sapiens	Systemic lupus erythematosus (SLE)	Blood	Human sorted B cell
	GSM746732	Patient 7	Homo sapiens	Systemic lupus erythematosus (SLE)	Blood	Human sorted B cell
	GSM746733	Patient 8	Homo sapiens	Systemic lupus erythematosus (SLE)	Blood	Human sorted B cell
	GSM746734	Patient 9	Homo sapiens	Systemic lupus erythematosus (SLE)	Blood	Human sorted B cell
	GSM746735	Patient 10	Homo sapiens	Systemic lupus erythematosus (SLE)	Blood	Human sorted B cell
	GSM746736	Patient 11	Homo sapiens	Systemic lupus erythematosus (SLE)	Blood	Human sorted B cell
	GSM746737	Patient 12	Homo sapiens	Systemic lupus erythematosus (SLE)	Blood	Human sorted B cell
	GSM746738	Patient 13	Homo sapiens	Systemic lupus erythematosus (SLE)	Blood	Human sorted B cell
	GSM746739	Patient 14	Homo sapiens	Systemic lupus erythematosus (SLE)	Blood	Human sorted B cell
	GSM746740	Patient 15	Homo sapiens	Systemic lupus erythematosus (SLE)	Blood	Human sorted B cell
	GSM746741	Patient 16	Homo sapiens	Systemic lupus erythematosus (SLE)	Blood	Human sorted B cell
	GSM746742	Patient 17	Homo sapiens	Systemic lupus erythematosus (SLE)	Blood	Human sorted B cell
Control	GSM746743	Control 1	Homo sapiens	Control	Blood	Human sorted B cell
	GSM746744	Control 2	Homo sapiens	Control	Blood	Human sorted B cell
	GSM746745	Control 3	Homo sapiens	Control	Blood	Human sorted B cell
	GSM746746	Control 4	Homo sapiens	Control	Blood	Human sorted B cell
	GSM746747	Control 5	Homo sapiens	Control	Blood	Human sorted B cell
	GSM746748	Control 6	Homo sapiens	Control	Blood	Human sorted B cell
	GSM746750	Control 8	Homo sapiens	Control	Blood	Human sorted B cell
	GSM746751	Control 9	Homo sapiens	Control	Blood	Human sorted B cell
	GSM746752	Control 10	Homo sapiens	Control	Blood	Human sorted B cell

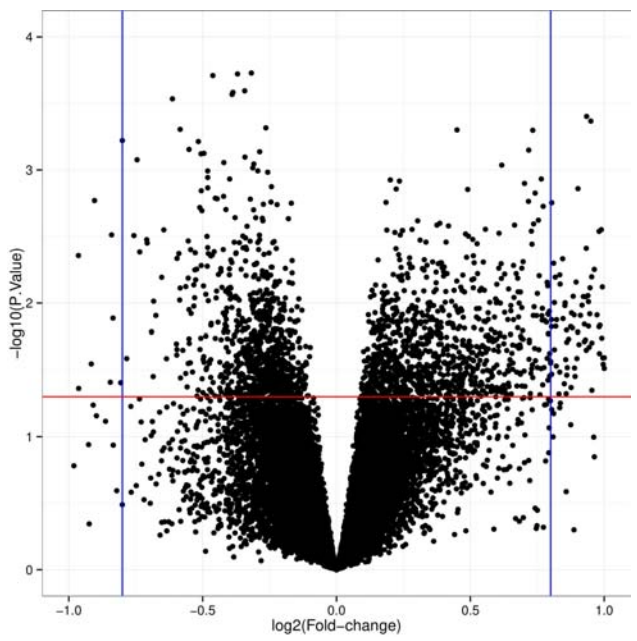


FIGURE 1 | Pictorial representation of volcano plot for differentially expressed genes (DEGs) in systemic lupus erythematosus (SLE) compared to controls from the GSE30153 dataset. The X-axis represents Log2FC, large magnitude fold changes; Y-axis represents −log10 of a *p*-value, high statistical significance. Each black dot represents one gene. Black dots above red and beside blue line (left-sided and right-sided) are log2FC ≥ 1 and *p*-value <0.05, representing SLE related DEGs.

TABLE 2 | Significantly upregulated and downregulated DEGs between two groups from GSE30153 dataset are tabulated.

GENE SYMBOL	log2FC	*p*-value
Upregulating Genes		
EGR1	1.22	0.00074
DSE	1.125	0.00291
CD1C	1.068	0.00053
GPM6A	1.052	0.00097
*GPM6A**	1.043	0.002981
Downregulating Genes		
RRM2	−2.406	0.0027527
*RRM2**	−2.152	0.0030096
TYMS	−1.923	0.0032032
CD38	−1.702	0.0031747
CAV1	−1.516	0.0048324
MIR7110	−1.4	0.0027212
ELL2	−1.354	0.0035256
SLC44A1	−1.219	0.0035298
SAR1B	−1.176	0.0049953
MAN1A1	−1.111	0.004425
CHAC2	−1.071	0.004354
ERAP1	−1.047	0.005321
ARF4	−1.044	0.0058274
PDIA4	−1.014	0.0041988

**The asterisk denotes the DEGs with two different probes from the dataset.*

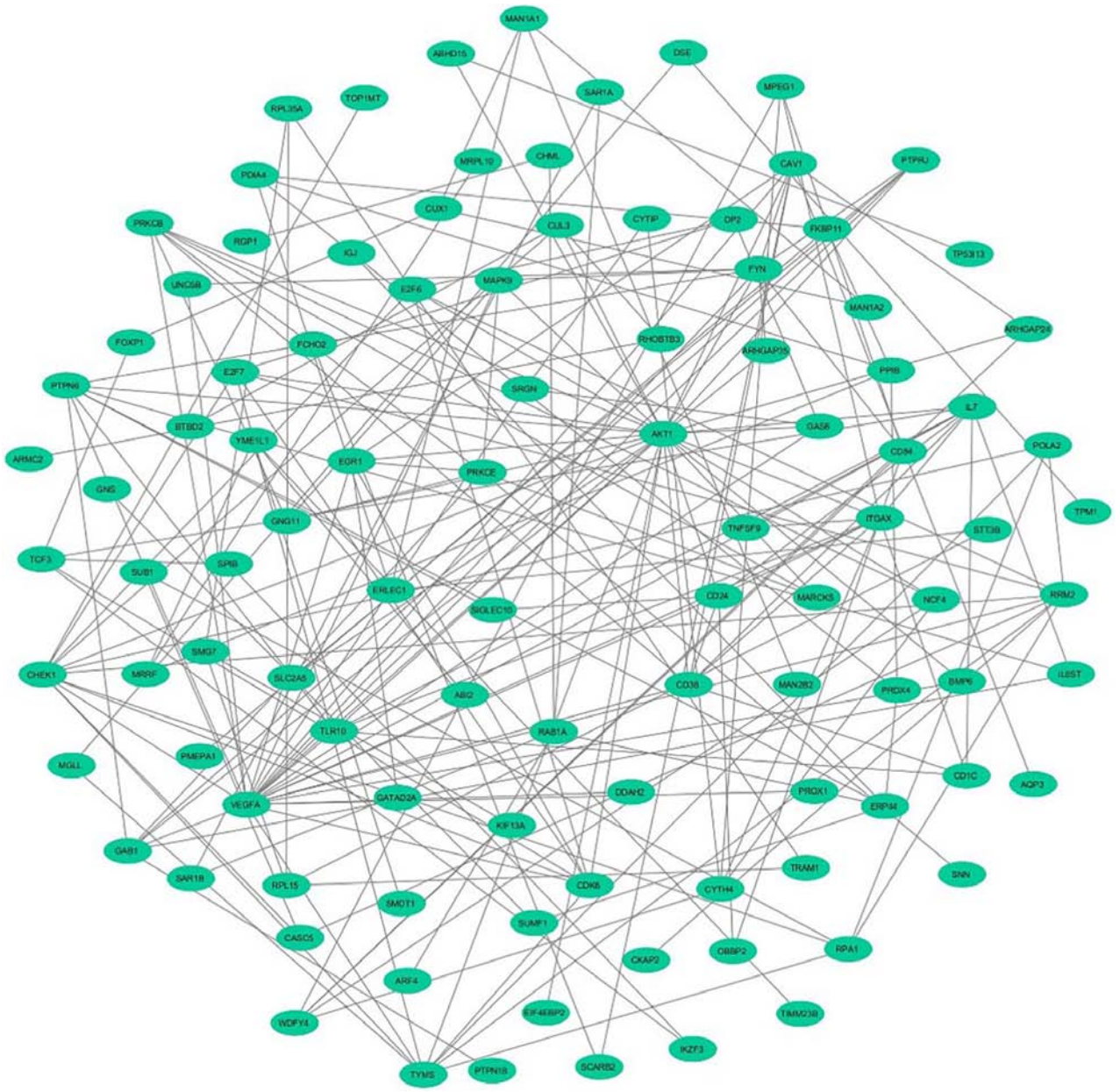


FIGURE 2 | The network demonstrates the protein–protein interaction between the DEGs identified from GSE30153 using Cytoscape. The nodes represented as ellipse (robin's blue) and edges as lines (gray).

Enrichment Analysis by ClueGO

The top two subnetworks from MCODE were used as an input for analyzing the functional enrichment of PPI subnetworks using the ClueGO/CluePedia plugin from Cytoscape. In a biologically clustered subnetwork, ClueGO helps to visualize the biological terms of broad gene clusters. The subnetwork enrichment analyses of MCODE cluster 1 and cluster 2 are depicted in **Figures 4A,B**. For functional enrichment analysis, we set the statistical options based on a two-sided hypergeometric test with a Benjamini–Hochberg correction, $p \leq 0.05$, and kappa scores ≥ 0.4 as criteria. The DEGs from cluster 1 were shown to be enriched mostly in T-cell costimulation (GO: 0031295), lymphocyte costimulation (GO: 0031294), negative regulation of vascular permeability (GO: 0043116), the metaphase/anaphase transition of the mitotic cell cycle (GO: 0007091), regulation of the transcription involved in the G1/S transition of the mitotic cell cycle (GO: 0000083), negative regulation of signal transduction in the absence of ligand (GO: 1901099), and KEGG pathways such as hematopoietic cell lineage (KEGG: 04640), B-cell receptor signaling pathway (KEGG: 04662), ErbB signaling pathway (KEGG: 04012), and AGE-RAGE (advanced glycation end products and receptor for AGE) signaling pathway in diabetic

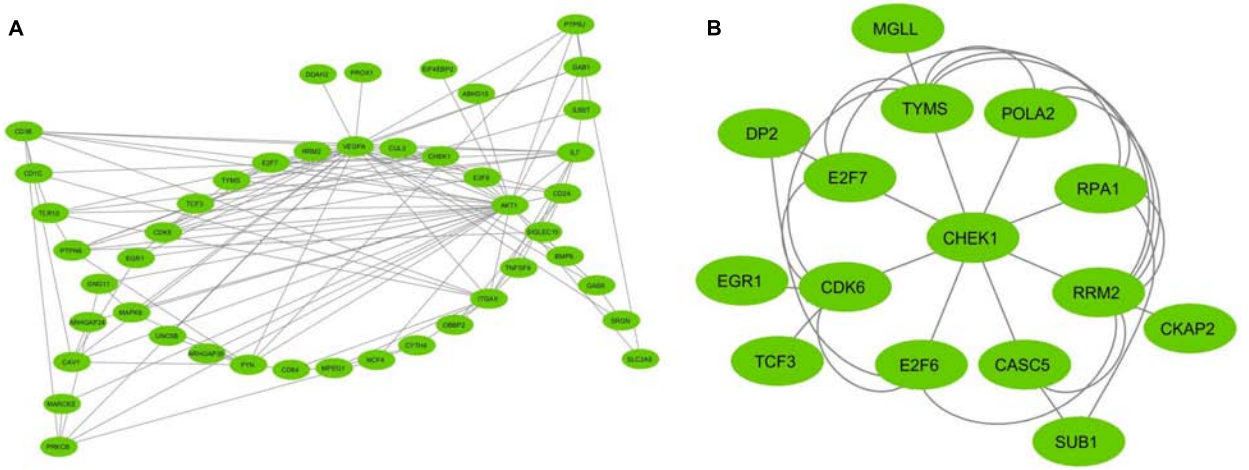


FIGURE 3 | The MCODE (Molecular Complex Detection) plugin from Cytoscape analyzed the top two clusters derived from the network of interactions between protein and protein. **(A)** Cluster 1; **(B)** Cluster 2. The MCODE cluster score > 3. The nodes represented as ellipse (green) and edges as lines (gray).

TABLE 3 | The interconnected regions are clustered from the GSE30153 dataset using MCODE plugin in Cytoscape.

Cluster	Score (density × No. of nodes)	Nodes	Edges	Node IDs
1	5.043	45	116	*OBBP2, IL6ST, CD1C, ABHD15, PTPRJ, ITGAX, UNC5B, TLR10, CD38, GAS6, NCF4, MAPK9, DDAH2, PTPN6, GAB1, ARHGAP24, CUL3, PROX1, CYTH4, E2F6, TNFSF9, VEGFA, TYMS, IL7, RRM2, PRKCB, MPEG1, MARCKS, SLC2A5, ARHGAP35, BMP6, TCF3, AKT1, EIF4EBP2, GNG11, CAV1, FYN, EGR1, SIGLEC10, CD24, CHEK1, E2F7, CD84, CDK6, SRGN*
2	3.625	15	29	*RRM2, CKAP2, MGLL, TCF3, SUB1, EGR1, POLA2, RPA1, CHEK1, E2F7, CASC5, DP2, E2F6, CDK6, TYMS*

complications (**Figure 4A**). The DEGs from cluster 2 were mainly enriched in the regulation of the transcription involved in the G1/S transition of the mitotic cell cycle (GO: 0000083), the negative regulation of the G0 and G1 transitions (GO: 0070317), and the p53 signaling pathway (KEGG: 04115) (**Figure 4B**). The pathways that were activated in the enrichment analysis were highly related to B-cell pathophysiology, resulting in events associated with the immune system, vasculopathy, and kidney.

Metacore^{TM} GeneGo for Enrichment Analysis of DEGs

Further functional enrichment analysis was carried out using Metacore^{TM} GeneGo software from Clarivate Analytics to comprehensively dissect the pathways associated with the DEGs. Using the functional ontology feature in GeneGo, the IDs of potential genes that were involved in the target pathways were identified. Based on hypergeometric *p*-values, the probability that the intersection of a gene set and associated ontological objects was random was evaluated. A decreased *p*-value indicated that the entity would be more significant to the DEGs, suggesting a better score. The functional enrichment analysis of the DEGs defined the top 10 metabolic networks, and canonical pathway maps are depicted in **Figures 5A,B**. For each classification, the significant statistical data rely on a low *p*-value. The pathway maps with the lowest *p*-value are shown in **Figures 6A–C**. These are the top-scoring signaling pathways based on the gene enrichment distribution, which emphasizes that the DEGs from human sorted B-cells are triggered via oxidative stress and ROS-induced cellular signaling (**Figure 6A**), chemotaxis and lysophosphatidic acid signaling via GPCRs (**Figure 6B**), and androgen receptor activation and downstream signaling in prostate cancer (**Figure 6C**). The well-distinguished proteins and complexes of proteins are shown as specific symbols[6]; all experimental data are displayed and have corresponding thermometer-like symbols on all the maps. The upregulated genes are indicated by a red thermometer facing upwards.

[6]https://portal.genego.com/help/MC_legend.pdf

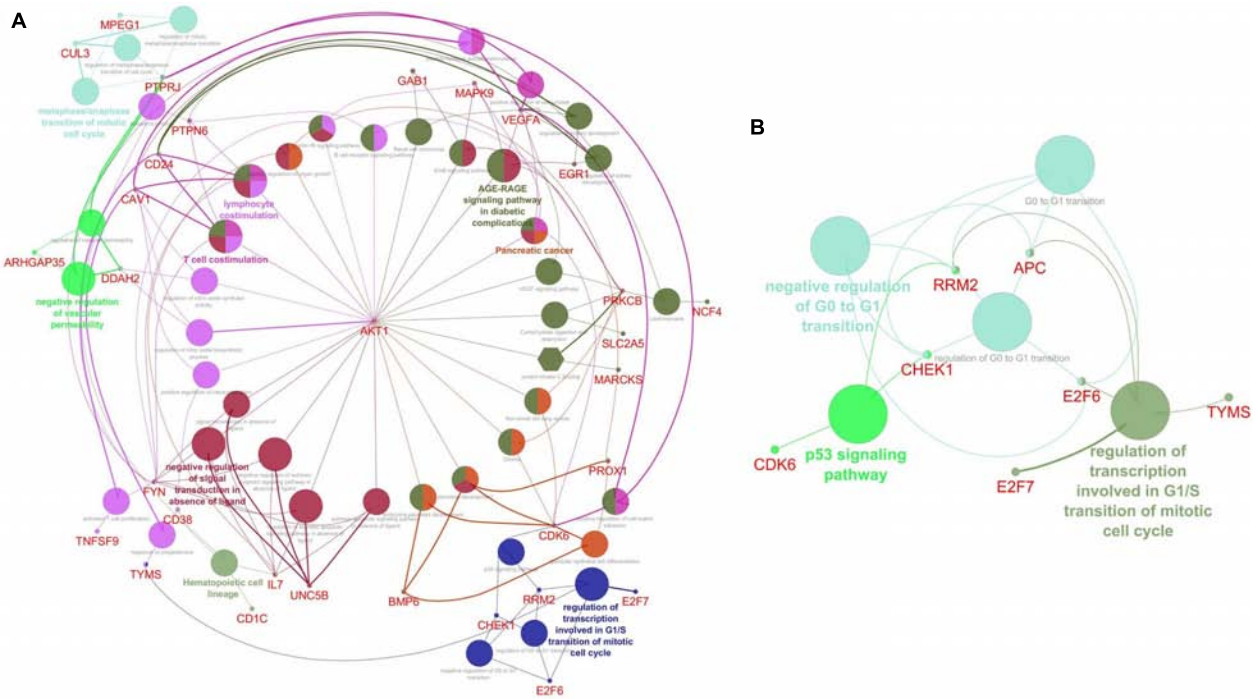


FIGURE 4 | Visualization of Gene Ontology (GO) enrichment profiles from DEGs using Cytoscape software based on network analysis of ClueGO/CluePedia inferred from MCODE cluster 1 **(A)** and cluster 2 **(B)**. The plugin provides a combined enrichment analysis of clusters, including the GO biological process, molecular function, and pathway from KEGG. The GO term/pathway network connectivity defined by edges and functional clusters on genes shared between terms (kappa score ≥ 0.4) and displaying pathways only with $p \leq 0.05$. The size of the node indicates the p-value. The color code of nodes represents the functional group that they belong to. The most important functional terms specify the pathway names within each class are indicated in bold colored characters. **(A)** The network enrichment analysis of cluster 1. Each node constitutes a precise term for cluster 1; **(B)** The network enrichment analysis of cluster 2. Each node constitutes a precise term for cluster 2.

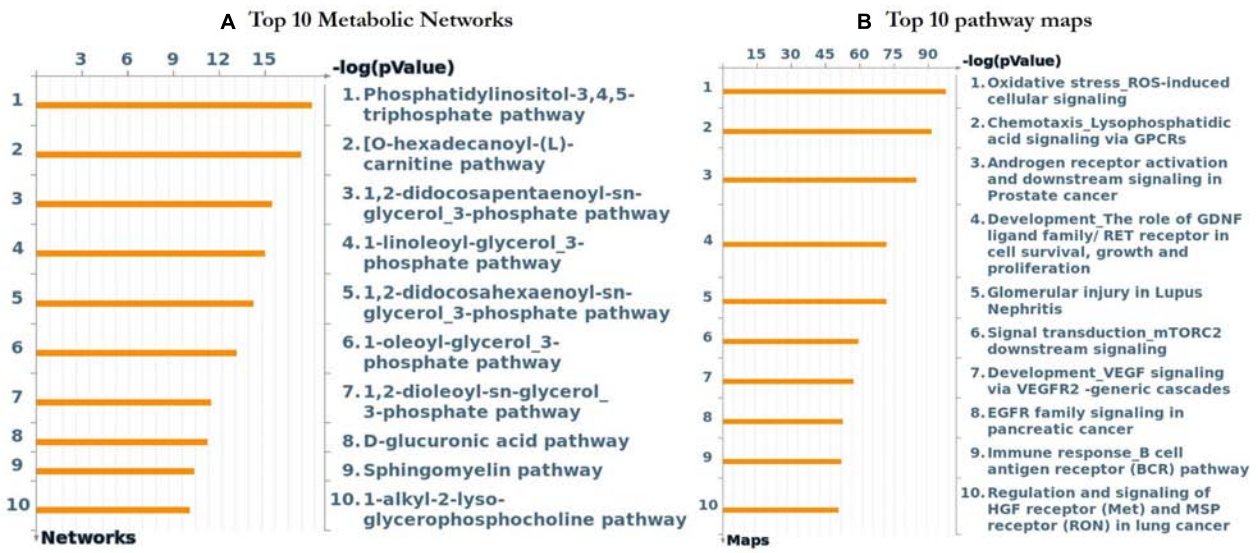


FIGURE 5 | The top 10 metabolic networks and pathway maps were annotated using GeneGo enrichment analysis for the genes that are differentially expressed from SLE patients vs. healthy controls, respectively. **(A)** The content of these metabolic networks was annotated and defined by GeneGo Cortellis Solution software. Each process represents a pre-set network of protein interactions characteristic for the process, and sorting was performed for the metabolic networks that are statistically significant. **(B)** The pathway maps (canonical) of GeneGo display a series of signals and metabolic charts that cover human in a structured manner. The significant expression of a gene/protein represented in histogram height.

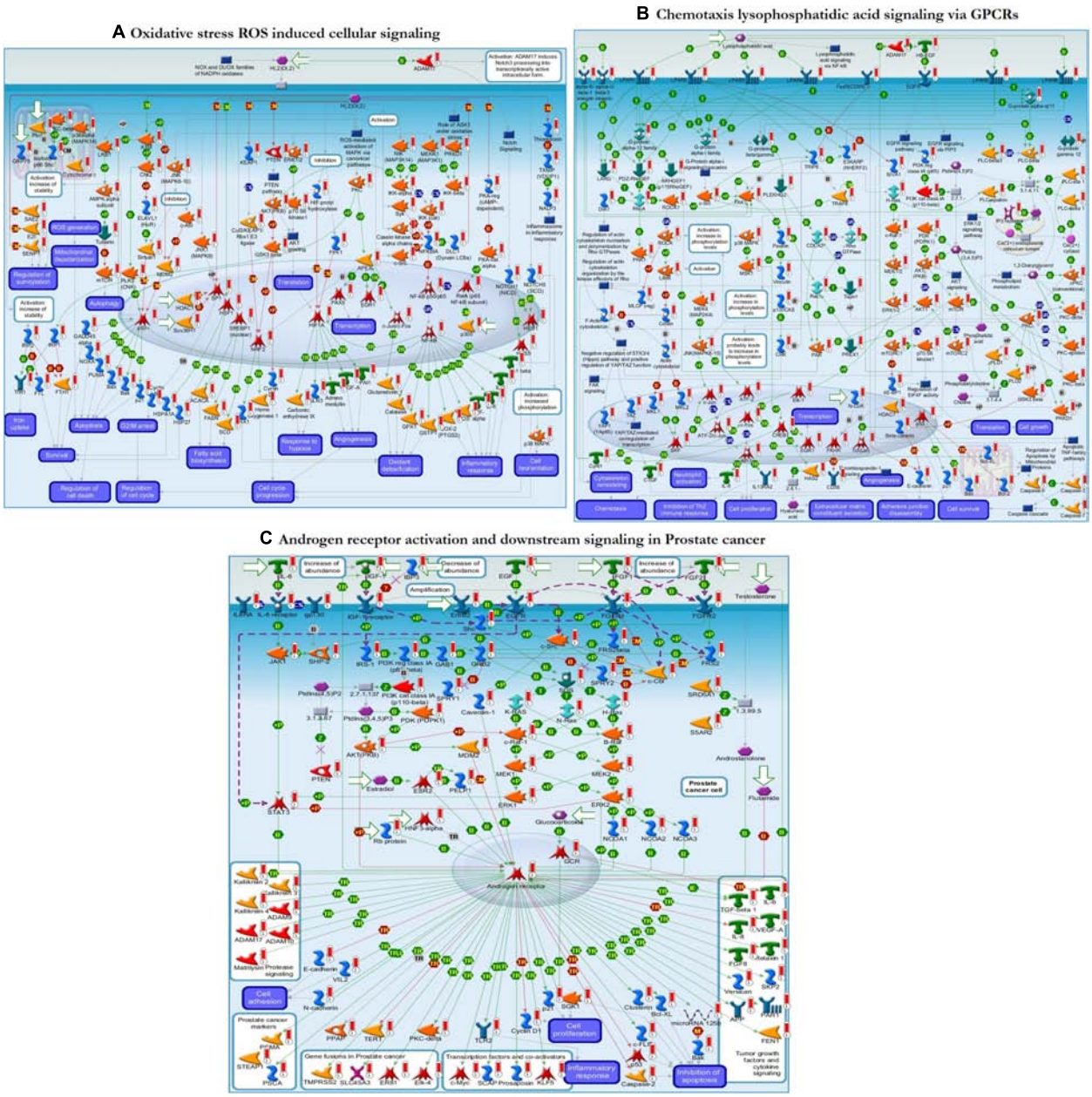


FIGURE 6 | The enrichment analysis from GeneGo showed three regulated pathways with the highest score that are triggered in the SLE human sorted B-cells. **(A)** Oxidative stress ROS induced cellular signaling. **(B)** Chemotaxis lysophosphatidic acid signaling via G protein-coupled receptors (GPCRs). **(C)** Androgen receptor activation and downstream signaling in prostate cancer. The image depicts the protein and protein complexes that are well characterized as a specific symbol; laboratory data from all reports are correlated and shown on the maps as thermometer-like indicators. The red or blue color upward/downward thermometers indicate gene transcripts with upregulation/downregulation, respectively. The proteins connected by arrows demonstrate the stimulating and inhibitory effect of the protein. Further details are given at https://portal.genego.com/help/MC_legend.pdf.

Pathway Map Interaction Results From Clarivate

From the Metacore™ results, we extracted the key genes from the enriched pathways that were differentially expressed, such as *EGR1*, *CD38*, *CAV1*, and *AKT1*. The differential expression of these genes was involved in the activation or inhibition of specific protein complexes in the enriched pathway maps (**Figure 6** and **Table 4**). Early growth response 1 (EGR1) is a transcription factor that interacts with the *IGF-2*, *APEX*, *SRD5A1*, *CD44*, and *EGFR* genes and activates them through transcriptional regulation. The cyclic adenosine diphosphate (ADP) ribose hydrolase CD38 is an enzyme involved in the activation of the genes *Semaphorin 4D, CD19,* and *c-Cbl* through physical interactions. Caveolin 1 (CAV1) is a binding

TABLE 4 | The interaction reports of key genes from pathway maps by Clarivate Analytics.

Network object "from"	Object type	Network object "to"	Object type	Effect	Mechanism	Link info	Input IDs	Signal	*P*-value	PMID
EGR1	Transcription factor	*IGF-2, APEX, SRD5A1, CD44, EGFR*	Receptor ligand, generic enzyme, generic enzyme, generic receptor, receptor with enzyme activity	Activation	Transcription regulation	EGR1 increases IGF II expression, EGR1 binds to gene APEX promoter and activates APEX expression, Egr-1 trans-activates the 5alpha-R1 promoter via the Egr-1-binding site at position −60/−54, Putative EGR1 binding site is found in gene CD44 promoter, EGR1 binds to gene EGFR promoter and activates EGFR expression.	*EGR1*	1	0.0032807	8584025; 9925986; 10606246; 11336542; 16043101; 19276347; 29092905; 29170465; 15788231; 15936112; 17194527; 18215136; 8628295; 9300687; 12670907; 15155664; 15923644; 19195913; 20357818; 25673149; 1417865; 11830539; 16750517; 17230532; 19032775; 20190820; 23763269
CD38	Generic enzyme	*SEMA4D, CD19, c-Cbl*	Generic receptor, generic binding protein, generic enzyme	Activation	Unspecified, Binding	CD31-induced activation of CD38 up-regulates Semaphorin 4D cell-surface expression in B cells, CD19/CD81 complex interacts with CD38 but this interaction is not required to induce proliferation in mouse B-lymphocytes, Fluorescence resource energy transfer and coimmunoprecipitation showed that c-Cbl and CD38 bind each other.	*CD38*	1	0.0031747	15613544; 17327405; 20570673; 22564057; 8695807; 18974118; 19635790
CAV1	Generic binding protein	*ErbB2, MDR1, HTR2A, Androgen receptor*	Receptor with enzyme activity, transporter, GPCR, transcription factor	Unspecified, Inhibition, activation	Binding	HER2 physically interacts with caveolin-1, Caveolin-1 interacts with p-gp, Down-regulation of caveolin-1 by siRNA reduced the interaction between p-gp and caveolin-1, followed by a decrease in [3H]-Taxol and [3H]-Vinblastine accumulation in RBE4 cells, Caveolin-1 physically interacts with HTR2A and increases its activity, Highly conserved 9 amino acid motif in the ligand binding domains (E domains) was identified in human/mouse ER alpha and ER beta, progesterone receptors A and B, and the androgen receptor. The localization sequence mediated palmitoylation of each SR, which facilitated caveolin-1 association, subsequent membrane localization, and steroid signaling.	*CAV1*	1	0.0048324	9374534; 9685399; 11697880; 22389470; 14622130; 15239129; 15498565; 17326770; 18485890; 19099191; 22389470; 25788263; 15190056; 8703009; 11278309; 17535799; 17940184; 18786521; 19931639; 22771325; 24375805

(Continued)

TABLE 4 | Continued

Network object "from"	Object type	Network object "to"	Object type	Effect	Mechanism	Link info	Input IDs	Signal	*P*-value	PMID
AKT1, AKT (PKB)	Protein kinase	*FKHR, mTOR, Bcl-10, FOXO3A, HNF3-beta, GSK3 beta*	Transcription factor, protein kinase, generic binding protein	Inhibition, activation	Phosphorylation	AKT1 phosphorylates FKHR1 and decreases its activity, Increased AKT phosphorylation regulates different metabolic pathways in liver, including increases in protein synthesis through activation of mTOR/p70 (S6kinase), AKT1 phosphorylates Bcl-10 and increases its activity, AKT1 phosphorylates FOXO3A and decreases its activity, AKT1 phosphorylates HNF3-beta and decreases its activity, AKT (PKB) inhibits GSK3 alpha by phosphorylation at Ser-9.	*AKT1*	1	0.0010146	10102273; 10358014; 10358075; 10377430; 11030146; 12393870; 16076959; 16099987; 16230533; 16603397; 17186497; 18388859; 18391970; 18420577; 18687691; 18786403; 19703413; 20940043; 21106439; 21157483; 21238503; 21407213; 21440577; 21708191; 21779512; 26053093; 27966458; 30413788; 10567225; 10910062; 11357143; 11438723; 12767043; 14970221; 15208671; 15549092; 16818631; 17660512; 18505677; 18566586; 18566587; 18678273; 21097843; 21177249; 21302298; 21343617; 22084251; 22595285; 23686889; 23872070; 26958938; 29221131; 16280327; 10102273; 12130673; 12767043; 17570479; 17577629; 17957242; 17960591; 18391970; 18687691; 19703413; 20223831; 20399660; 21106439; 21157483; 21440577; 21621563; 21708191; 21775285; 21779512; 24518891; 27966458; 14500912; 11584303; 11701324; 12124352; 12750378; 12808085; 14966899; 14985354; 15016802; 15297258

protein shown to interact with the *ErbB2, MDR1, HTR2A,* and *androgen receptor* genes, and inhibition or activation is followed by specific binding to its corresponding proteins. RAC-alpha serine/threonine-protein kinase (AKT1) is a protein kinase that interacts with the *FKHR, mTOR, Bcl-10, FOXO3A, HNF3-beta,* and *GSK3 beta* genes via phosphorylation, resulting in inhibition or activation. These genes were differentially expressed between sorted B-cells from controls and sorted B-cells from SLE patients and result in transcriptional regulation and inhibition of genes/proteins within the top-scored pathway maps.

DISCUSSION

DNA microarrays and next generation sequencing (NGS) approaches are high-throughput technologies that have resulted in the emergence of new biomedical discoveries. Data from microarray and gene expression profiles have enabled a deeper understanding of the intrinsic molecular pathways of complex mechanisms of biological systems and their responses (Russo et al., 2003; Babu, 2004; Perez-Diez et al., 2013; Kumar et al., 2019). It is therefore highly relevant to examine the peripheral B-cell transcriptomes of SLE patients and healthy controls to determine genes that are differentially regulated and their target pathways. Our current study extracted DEGs from 17 SLE patients and 9 healthy controls from the GEO database (GSE30153) (Garaud et al., 2011). The top 250 DEGs were identified, including 4 upregulated and 13 downregulated genes from the groups through bioinformatics strategies (**Table 2** and **Supplementary Table S1**). These identified DEGs were subjected to ClueGO and GeneGo Metacore^{TM} analysis for GO and pathway annotation, and constructed the interacting networks of PPI and used for cluster analysis. In the network, the nodes were considered proteins, and the edges were their interactions. Using network topology features, the PPI network can be analyzed to distinguish the core proteins that are involved in the pathways (Barabási and Oltvai, 2004; Ideker and Sharan, 2008; Keskin et al., 2016; Kumar et al., 2019). The identified DEGs from the present study were analyzed with STRING to exploit the complex interactions between the DEGs via text mining, evidence from experiments, and repositories (**Figure 2**). We performed module screening of the PPI networks using the MCODE plugin from Cytoscape. As a result, we obtained significant clusters that are densely interlinked regions in the PPI network (**Figures 3A,B**). Screening of these clusters from the network might help to identify the essential genes that are involved in the pathogenesis and progression of SLE. The obtained clusters mostly contained protein complexes or proteins present in the pathways in the PPI network, and cluster visualization is essential for comprehending the properties of the network functionally and systematically (Krogan et al., 2006; Rahman et al., 2013).

Furthermore, to identify the functional enrichment of these subnetworks from MCODE, we implemented the ClueGO plugin for analysis. This revealed that the DEGs were enriched in most essential pathways, which are highly associated with the immune system. The GO and KEGG enrichment analyses of the DEGs from cluster 1 showed that they were mostly enriched in T-cell costimulation, lymphocyte costimulation, negative regulation of vascular permeability, the metaphase/anaphase transition of the mitotic cell cycle, regulation of the transcription involved in the G1/S transition of mitotic cell cycle, the hematopoietic cell lineage, the B-cell receptor signaling pathway, the ErbB signaling pathway, the AGE-RAGE signaling pathway in diabetic complications, and pancreatic cancer. Interestingly, the costimulation of T-cell and lymphocyte receptors is recognized to be important in SLE pathogenesis by enabling communicating with B-cells for the production of autoantibodies (Shlomchik et al., 2001; Mak and Kow, 2014). In SLE, negative regulation of vascular permeability may be induced by different mechanisms; the dysregulated genes from the cluster 1 subnetwork might lead to endothelial cell damage and vasculopathy (Favero et al., 2014; Lee et al., 2019). The differential cell signaling results in the recruitment of various proteins and inappropriate activation of B-cells (Zhou et al., 2009; Comte et al., 2015). Oxidative stress is common in inflammatory disorders and results in the increased production of reactive carbonyl groups that are partially converted to AGEs, and the DEGs in the AGE-RAGE signaling pathway might also be involved in the accumulation of AGEs in SLE patients and lead to diabetic complications (de Leeuw et al., 2007; Li et al., 2007; Kurien and Scofield, 2008; Nienhuis et al., 2008). Interestingly, our enrichment analysis found that the identified differential expression of the genes (*AKT1*, *VEGFA*, *CDK6*, and *MAPK9*) that were involved in the risk of developing pancreatic cancer in SLE patients was due to chronic inflammation, suggesting that these genes might be involved in the pathogenesis of SLE. Our findings are therefore consistent with the roles of genes that are differentially expressed in SLE-causing pathways (**Figure 4A**). The enrichment analysis of the cluster 2 subnetwork showed that the DEGs were mostly enriched in the regulation of transcription involved in the G1/S transition of the mitotic cell cycle, the negative regulation of the G0 and G1 transitions, and the p53 signaling pathway. It has been reported that the proliferation of T-cells is followed by lowered levels of cyclin-dependent kinase (CDK) inhibitors, and alterations in the expression of CDKs in the G0/G1 phase were seen in the lymphocytes of SLE patients (Yamauchi and Bloom, 1997; Tang et al., 2009). The DEGs involved in the cluster 2 subnetwork might negatively regulate these pathways. Alterations in cyclin-CDK complex behavior and cyclin-dependent kinase inhibitors (CDKIs) have been reported to alter the proliferation of T-cells, oxidative stress, and immune responses (Santiago-Raber et al., 2001; Tang et al., 2009). p53 signaling is essential for various cellular mechanisms, and defects in this signaling pathway are associated with SLE development. Considerably elevated levels of p53 protein are found in SLE patients with active inflammatory disorders (Miret et al., 2003; Veeranki and Choubey, 2010). Apoptosis dysregulation appears to be another cause of SLE pathogenesis because the possible sources of autoantigens are cell debris from apoptosis in SLE, and excessive cellular senescence of the immune cells, especially T-cells, was reported in SLE patients with peripheral blood mononuclear cells (PBMCs) and skin lesions (Colonna et al., 2014; Sáenz-Corral et al., 2015). Thus, our identified DEGs (*RRM2*, *APC*, *CHEK1*, *E2F6*, *TYMS*, *E2F7*, and *CDK6*) from the

cluster 2 subnetwork are highly related to and consistent with the members of the signaling pathways associated with the immune system, apoptosis, the cell cycle, and vasculopathy.

To clearly define the interactions between the proteins and signaling pathways examined from the interpretation of STRING, Cytoscape, MCODE, and ClueGO analyses, we implemented the GeneGo Metacore software, which incorporates extensive data on metabolic signaling pathways and their regulatory mechanisms and contains accurately complied networks of biological systems. By utilizing the GeneGo Metacore software, we obtained a detailed description of the DEGs that participate in SLE pathogenesis based on the determined *p*-values. Among the top 10 metabolic networks, the phosphatidylinositol-3,4,5-triphosphate pathway, O-hexadecanoyl-(L)-carnitine pathway, 1,2-didocosapentaenoyl-sn-glycerol 3-phosphate pathway, and 1-linoleoyl-glycerol-3-phosphate pathway were profoundly enriched and significant in the SLE DEGs (**Figure 5A**). The increased activity of phosphatidylinositol-3,4,5-triphosphate stimulates essential cell signaling pathways such as the pathways involved in cell division, survival, and the rapid increase in T-lymphocytes in SLE (Comte et al., 2015). PI3K (phosphatidylinositol 3-kinase) is a protein kinase that phosphorylates phosphatidylinositol 4,5-phosphate to regulate the signaling of T-lymphocytes; an increased amount of PI3K was also observed in an animal model of lupus (Liu et al., 1998; Grolleau et al., 2000; Niculescu et al., 2003; Joseph et al., 2014). Modification of the carnitine signaling pathway results in various organ failures by producing effective responses to pathogens (Famularo and De Simone, 1995; Famularo et al., 2004). Thus, the DEGs involved in the O-hexadecanoyl-(L)-carnitine pathway might lead to increased immune responses. In addition, the top three pathways associated with the DEGs of sorted B-cells from SLE patients were mostly enriched in oxidative stress- and ROS-induced cellular signaling (**Figure 6A**), chemotaxis and lysophosphatidic acid signaling via GPCRs (**Figure 6B**), and androgen receptor activation and downstream signaling in prostate cancer (**Figure 6C**). Recent findings have shown that oxidative stress and ROS induce molecular alterations that have adverse effects in SLE patients (Choi et al., 2016; Tsokos et al., 2016; Lightfoot et al., 2017). Elevated oxidative stress in SLE patients leads to the accumulation of higher amounts of oxidative lipoproteins, which are harmful in zebrafish models and cause additional oxidative damage to the system (Chung et al., 2007; Park et al., 2016; Lightfoot et al., 2017). Interestingly, our study identified the *EGR1* gene as downregulated in the SLE patients in comparison to controls, and it also plays a role in ROS signaling. This clearly indicates that *EGR1* might be required to maintain the oxidative stress and ROS signaling pathways.

Moreover, the DEGs involved in the oxidative stress signaling pathway might contribute to peripheral neuropathy, damage to blood vessels, and cardiovascular events, which are the prominent clinical conditions found in SLE patients. Chemotaxis and lysophosphatidic acid (LPA) signaling are essential pathways in autoimmune inflammatory disorders, and GPCRs are responsible for regulating immune cells via LPA receptors (Yang et al., 2005; Skoura and Hla, 2009). *G2A* gene knockout resulted in the hyperresponsiveness of T-cells to T-cell receptor stimulation, manifesting as an increased proliferation of T-cells, which may promote inflammatory phenotypes in G2A-deficient mice (Le et al., 2001). Various studies have suggested that LPA plays a vital role in atherosclerosis progression and development by promoting neutrophil and monocyte adherence and enhancing inflammatory events (Siess et al., 1999; Smyth et al., 2008; Skoura and Hla, 2009). The androgen receptor (AR) is a transcription factor that is activated by a ligand and is essential for cells targeted by the androgen response (Robeva et al., 2013; Gubbels Bupp and Jorgensen, 2018). AR also regulates immune function in SLE via transcriptional regulation of various genes. Our study identified the transcription factor *AR,* which positively regulates the *c-Myc, SCAP, prosaposin,* and *KLF5* genes, which are responsible for inflammatory responses, and promotes tumor growth factors and cytokine signaling when activated (**Figure 6C**). The enriched terms from ClueGO modules and the GeneGo-identified terms correlated well in this study and validate the significance of the findings from the pathway maps. The combined results from these two enrichment analyses suggest that B-cells from SLE patients and B-cells from healthy controls undergo differential gene expression associated with positive regulation of kidney development, the hematopoietic cell lineage, positive regulation of vasoconstriction, T-cell costimulation, and regulation of the transcription involved in the G1/S transition of the mitotic cell cycle.

Furthermore, the interaction results from the GeneGo analysis provided the essential genes (*EGR1*, *CD38*, *CAV1*, and *AKT1*) from the pathway maps constructed from the DEGs. Among them, *EGR1* (early growth response 1) is a transcription factor shown to interact with the *IGF-2* (insulin-like growth factor 2), *APEX* (apurinic/apyrimidinic endodeoxyribonuclease 1), *SRD5A1* (steroid 5 alpha-reductase 1), *CD44* (cell surface glycoprotein CD44), and *EGFR* (epidermal growth factor receptor) genes and transcriptionally regulate them by activating or promoting their expression in sorted B-cells from patients with SLE (Liu et al., 1995; Recio and Merlino, 2003; Lee et al., 2005; Pines et al., 2005; Blanchard et al., 2007; Rui et al., 2008; Cullen et al., 2010; Sauer et al., 2010). The cyclic ADP ribose hydrolase (CD38) is also known as cluster of differentiation 38 protein, can be found on several immune cells, and activates SEMA4D (semaphorin-4D or cluster of differentiation 100), CD19 (B-lymphocyte antigen CD19 or cluster of differentiation 19), and c-Cbl (Casitas B-lineage lymphoma proto-oncogene) (Deaglio et al., 2005, 2007; Shen and Yen, 2008; Vences-Catalán et al., 2012). These interactions with *CD38* result in the activation of B-lymphocytes and increase immune responses in SLE patients. The protein caveolin 1 (CAV1) has been shown to interact with the ErbB2 (Erb-B2 receptor tyrosine kinase 2), MDR1 (multidrug resistance protein 1), HTR2A (5-hydroxytryptamine receptor 2A), and AR (androgen receptor) genes. Several studies have suggested that caveolin 1 physically interacts with HER2, p-gp, HTR2A, and AR and activates/inhibits them by binding to their specific caveolin-binding motif (Couet et al., 1997; Lu et al., 2001; Razani and Lisanti, 2001; Bhatnagar et al., 2004; Bennett et al., 2010, 2014; Yu et al., 2012). AKT1 is a protein kinase that interacts with *FKHR* (Forkhead box protein O1), mTOR (mechanistic target of rapamycin), Bcl-10 (B-cell lymphoma/leukemia 10), FOXO3A (Forkhead box O3),

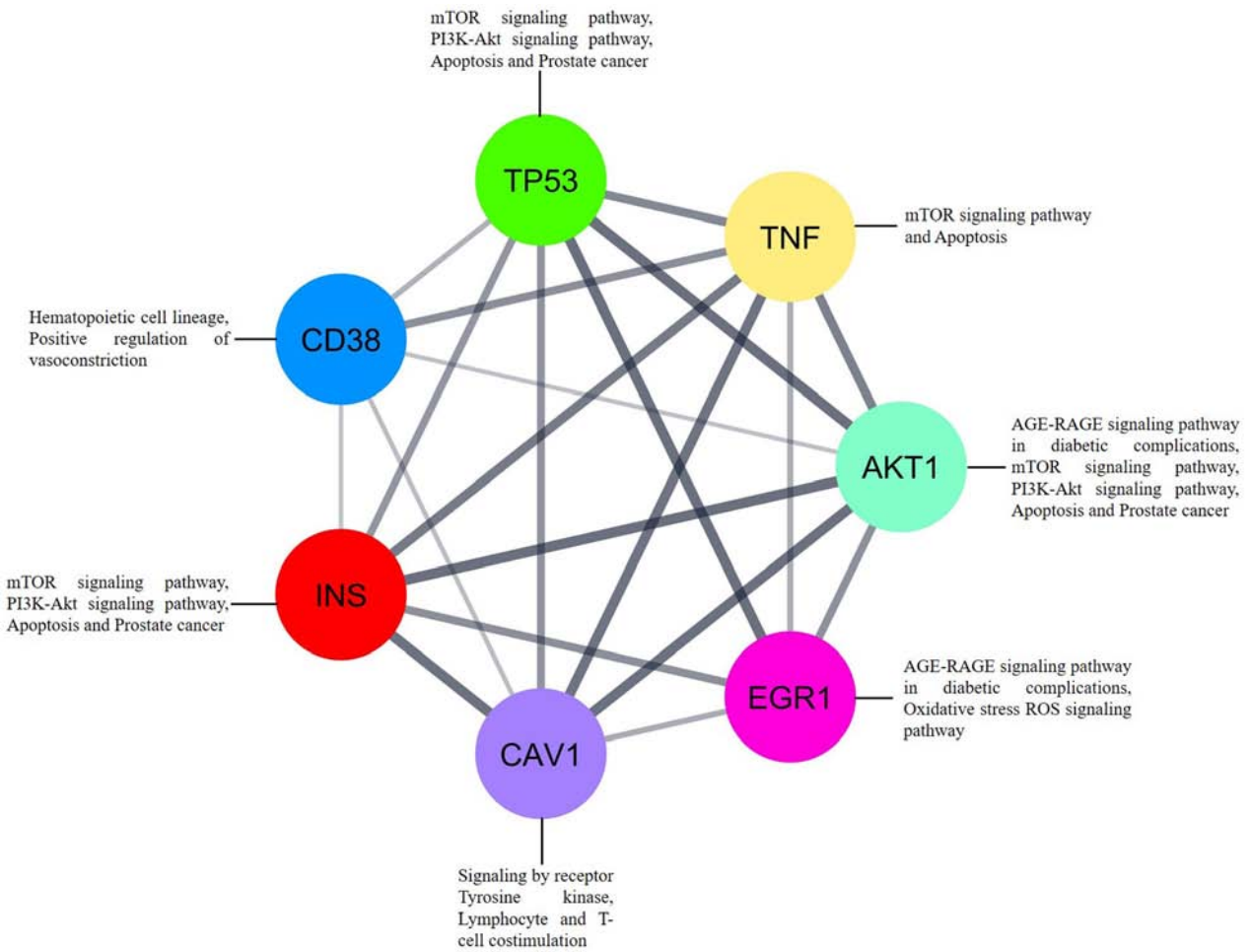


FIGURE 7 | The interrelation analysis of genes *EGR1*, *CD38*, *CAV1*, and *AKT1* that strongly associated to SLE. Each gene involved in different pathways via interacting to each other. Inbuilt color code was provided to all the genes based on the STRING tool from Cytoscape.

HNF3-beta (hepatocyte nuclear factor 3-beta), and GSK3 beta (glycogen synthase kinase three beta). The protein kinase AKT1 inhibits FKHR via phosphorylation and decreases its activity (Biggs et al., 1999; Rena et al., 1999; Tang et al., 1999; Hay, 2011), whereas it increases its activity via phosphorylation of mTOR (Navé et al., 1999; Sekuliæ et al., 2000; Ikenoue et al., 2008; Thirumal Kumar et al., 2019). Additionally, AKT1 phosphorylates Bcl-10 at the specific residues Ser231 and Ser218, increasing its activity (Yeh et al., 2006), while it inhibits the action of FOXO3A via phosphorylation and decreases its activity, increasing the survival of cells (Brunet et al., 1999; Linding et al., 2007; Calnan and Brunet, 2008; Li et al., 2008; Tzivion et al., 2011). AKT1 decreases HNF3-beta activity by phosphorylating it at Thr156 (Wolfrum et al., 2003), whereas phosphorylation of GSK3-beta by AKT1 occurs at Ser9 to inhibit its activity (Brazil and Hemmings, 2001; Salas et al., 2004; Kuemmerle, 2005; Shin et al., 2006; Markou et al., 2008). This suggests the vital genes we identified from the DEGs of patients with SLE play essential roles in the development and progression of SLE via different signaling pathways to increase autoimmune responses.

In addition to the interaction analysis, we carried out interrelation analysis for the essential genes to determine the relationships between the genes, which implicitly or explicitly interacted with each other. Interestingly, the identified genes indirectly communicated with each other via molecular signaling pathways related to mTOR signaling, apoptosis, PI3K-Akt signaling, the hematopoietic cell lineage, positive regulation of vasoconstriction, signaling by receptor tyrosine kinases, AGE-RAGE signaling, and lymphocyte and T-cell costimulation (**Figure 7**). *EGR1* and *AKT1* are directly involved in oxidative stress via ROS and AGE-RAGE signaling, whereas *CAV1* is directly involved in tyrosine kinase receptor signaling and lymphocyte and T-cell costimulation. *CD38* is directly associated with the hematopoietic cell lineage and positive regulation of vasoconstriction. Overall, the dysregulation of the indicated pathways in SLE patients is a result of differential gene expression. The essential genes are differentially expressed between cells from patients with SLE and cells from healthy controls and are present in important signaling cascades, which could be a crucial factor for SLE development.

CONCLUSION

Taken together, the results of our comprehensive bioinformatics analysis showed that the DEGs identified between sorted B-cells from patients with SLE sorted B-cells from controls could play a significant role in the growth, progression, and development of SLE. This study identified 4 upregulated and 13 downregulated genes, including essential genes (*EGR1*, *CD38*, *CAV1*, and *AKT1*), from the pathway enrichment analysis. Indeed, the identified pathways from the enrichment analysis were strongly related to the immune system, vasculopathy, cardiovascular functions, and inflammatory responses, which are processes that can lead to the development of SLE. The broad understanding of SLE pathophysiology from this study will allow us to identify and develop therapies targeting SLE and contribute to personalized treatment strategies. Collectively, the study findings could aid in enhancing our understanding of the fundamental molecular processes of SLE and provide possible strategies for early diagnosis in SLE; in addition, combinatorial therapeutic strategies using oxidative stress and ROS cellular signaling and lysophosphatidic acid signaling via GPCRs might have symbiotic effects on the molecular events in SLE.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE30153.

AUTHOR CONTRIBUTIONS

SU, DT, HZ, and CG were involved in the design of the study and the acquisition, analysis, and interpretation of the data. SU, DT, CG, SY, NY, MS, and HZ were involved in the interpretation of the data and drafting the manuscript. CG, RS, and HZ supervised the entire study and were involved in study design, the acquisition, analysis, and interpretation of the data, and drafting the manuscript. The manuscript was reviewed and approved by all the authors.

FUNDING

This work was supported by Qatar University Grant# QUST-1-CHS-2020-2.

ACKNOWLEDGMENTS

We would like to take this opportunity to thank the management of VIT for providing the necessary facilities and encouragement to carry out this work.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fbioe.2020.00276/full#supplementary-material

REFERENCES

Alibés, A., Yankilevich, P., Cañada, A., and Díaz-Uriarte, R. (2007). IDconverter and IDClight: conversion and annotation of gene and protein IDs. *BMC Bioinformatics* 8:9. doi: 10.1186/1471-2105-8-9

Aubert, J., Bar-Hen, A., Daudin, J. J., and Robin, S. (2004). Determination of the differentially expressed genes in microarray experiments using local FDR. *BMC Bioinformatics* 5:125. doi: 10.1186/1471-2105-5-125

Babu, M. M. (2004). *An Introduction to Microarray Data Analysis*. Cambridge: Horizon Bioscience, 225–249.

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

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, D991–D995. doi: 10.1093/nar/gks1193

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

Bennett, N. C., Gardiner, R. A., Hooper, J. D., Johnson, D. W., and Gobe, G. C. (2010). Molecular cell biology of androgen receptor signalling. *Int. J. Biochem. Cell Biol.* 42, 813–827. doi: 10.1016/j.biocel.2009.11.013

Bennett, N. C., Hooper, J. D., Johnson, D. W., and Gobe, G. C. (2014). Expression profiles and functional associations of endogenous androgen receptor and caveolin-1 in prostate cancer cell lines. *Prostate* 74, 478–487. doi: 10.1002/pros.22767

Bentham, J., Morris, D. L., Cunninghame Graham, D. S., Pinder, C. L., Tombleson, P., Behrens, T. W., et al. (2015). Genetic association analyses implicate aberrant regulation of innate and adaptive immunity genes in the pathogenesis of systemic lupus erythematosus. *Nat. Genet.* 47, 1457–1464. doi: 10.1038/ng.3434

Bhatnagar, A., Sheffler, D. J., Kroeze, W. K., Compton-Toth, B., and Roth, B. L. (2004). Caveolin-1 interacts with 5-HT2A serotonin receptors and profoundly modulates the signaling of selected Gαq-coupled protein receptors. *J. Biol. Chem.* 279, 34614–34623. doi: 10.1074/jbc.M404673200

Biggs, W. H., Meisenhelder, J., Hunter, T., Cavenee, W. K., and Arden, K. C. (1999). Protein kinase B/Akt-mediated phosphorylation promotes nuclear exclusion of the winged helix transcription factor FKHR1. *Proc. Natl. Acad. Sci. U.S.A.* 96, 7421–7426. doi: 10.1073/pnas.96.13.7421

Bindea, G., Mlecnik, B., Hackl, H., Charoentong, P., Tosolini, M., Kirilovsky, A., et al. (2009). ClueGO: a cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. *Bioinformatics* 25, 1091–1093. doi: 10.1093/bioinformatics/btp101

Blanchard, Y., Seenundun, S., and Robaire, B. (2007). The promoter of the rat 5α-reductase type 1 gene is bidirectional and Sp1-dependent. *Mol. Cell. Endocrinol.* 264, 171–183. doi: 10.1016/j.mce.2006.11.007

Borrebaeck, C. A. K., Sturfelt, G., and Wingren, C. (2014). "Recombinant antibody microarray for profiling the serum proteome of SLE," in *Systemic Lupus Erythematosus: Methods and Protocols* Methods in Molecular Biology, eds P. Eggleton and F. J. Ward (New York, NY: Springer), 67–78. doi: 10.1007/978-1-4939-0326-9_6

Brazil, D. P., and Hemmings, B. A. (2001). Ten years of protein kinase B signalling: a hard Akt to follow. *Trends Biochem. Sci.* 26, 657–664. doi: 10.1016/S0968-0004(01)01958-2

Brunet, A., Bonni, A., Zigmond, M. J., Lin, M. Z., Juo, P., Hu, L. S., et al. (1999). Akt promotes cell survival by phosphorylating and inhibiting a forkhead transcription factor. *Cell* 96, 857–868. doi: 10.1016/S0092-8674(00)80595-4

Calnan, D. R., and Brunet, A. (2008). The FoxO code. *Oncogene* 27, 2276–2288. doi: 10.1038/onc.2008.21

Choi, S.-C., Titov, A. A., Sivakumar, R., Li, W., and Morel, L. (2016). Immune cell metabolism in systemic lupus erythematosus. *Curr. Rheumatol. Rep.* 18:66. doi: 10.1007/s11926-016-0615-7

Chung, C. P., Avalos, I., Oeser, A., Gebretsadik, T., Shintani, A., Raggi, P., et al. (2007). High prevalence of the metabolic syndrome in patients with systemic lupus erythematosus: association with disease characteristics and cardiovascular risk factors. *Ann. Rheum. Dis.* 66, 208–214. doi: 10.1136/ard.2006.054973

Colonna, L., Lood, C., and Elkon, K. (2014). Beyond apoptosis in lupus. *Curr Opin Rheumatol.* 26, 459–466. doi: 10.1097/BOR.0000000000000083

Comte, D., Karampetsou, M. P., and Tsokos, G. C. (2015). T cells as a therapeutic target in SLE. *Lupus* 24, 351–363. doi: 10.1177/0961203314556139

Costa-Reis, P., and Sullivan, K. E. (2013). Genetics and epigenetics of systemic lupus erythematosus. *Curr. Rheumatol. Rep.* 15:369. doi: 10.1007/s11926-013-0369-4

Couet, J., Sargiacomo, M., and Lisanti, M. P. (1997). Interaction of a receptor tyrosine kinase, EGF-R, with caveolins. caveolin binding negatively regulates tyrosine and serine/threonine kinase activities. *J. Biol. Chem.* 272, 30429–30438. doi: 10.1074/jbc.272.48.30429

Cui, Y., Sheng, Y., and Zhang, X. (2013). Genetic susceptibility to SLE: Recent progress from GWAS. *J. Autoimm.* 41, 25–33. doi: 10.1016/j.jaut.2013.01.008

Cullen, E. M., Brazil, J. C., and O'Connor, C. M. (2010). Mature human neutrophils constitutively express the transcription factor EGR-1. *Mol. Immunol.* 47, 1701–1709. doi: 10.1016/j.molimm.2010.03.003

Dang, J., Li, J., Xin, Q., Shan, S., Bian, X., Yuan, Q., et al. (2016). Gene–gene interaction of ATG5, ATG7, BLK and BANK1 in systemic lupus erythematosus. *Int. J. Rheum. Dis.* 19, 1284–1293. doi: 10.1111/1756-185X.12768

de Leeuw, K., Graaff, R., de Vries, R., Dullaart, R. P., Smit, A. J., Kallenberg, C. G., et al. (2007). Accumulation of advanced glycation endproducts in patients with systemic lupus erythematosus. *Rheumatology (Oxford)* 46, 1551–1556. doi: 10.1093/rheumatology/kem215

Deaglio, S., Vaisitti, T., Bergui, L., Bonello, L., Horenstein, A. L., Tamagnone, L., et al. (2005). CD38 and CD100 lead a network of surface receptors relaying positive signals for B-CLL growth and survival. *Blood* 105, 3042–3050. doi: 10.1182/blood-2004-10-3873

Deaglio, S., Vaisitti, T., Billington, R., Bergui, L., Omede', P., Genazzani, A. A., et al. (2007). CD38/CD19: a lipid raft–dependent signaling complex in human B cells. *Blood* 109, 5390–5398. doi: 10.1182/blood-2006-12-061812

Deng, Y., and Tsao, B. P. (2010). Genetic susceptibility to systemic lupus erythematosus in the genomic era. *Nat. Rev. Rheumatol.* 6, 683–692. doi: 10.1038/nrrheum.2010.176

Ducreux, J., Houssiau, F. A., Vandepapelière, P., Jorgensen, C., Lazaro, E., Spertini, F., et al. (2016). Interferon α kinoid induces neutralizing anti-interferon α antibodies that decrease the expression of interferon-induced and B cell activation associated transcripts: analysis of extended follow-up data from the interferon α kinoid phase I/II study. *Rheumatology (Oxford)* 55, 1901–1905. doi: 10.1093/rheumatology/kew262

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

Famularo, G., and De Simone, C. (1995). A new era for carnitine? *Immunol. Today* 16, 211–213. doi: 10.1016/0167-5699(95)80159-6

Famularo, G., Simone, C., de, Trinchieri, V., and Mosca, L. (2004). Carnitines and its congeners: a metabolic pathway to the regulation of immune response and inflammation. *Ann. N. Y. Acad. Sci.* 1033, 132–138. doi: 10.1196/annals.1320.012

Favero, G., Paganelli, C., Buffoli, B., Rodella, L. F., and Rezzani, R. (2014). Endothelium and its alterations in cardiovascular diseases: life style intervention. *Biomed. Res. Int.* 2014:801896. doi: 10.1155/2014/801896

Garaud, J.-C., Schickel, J.-N., Blaison, G., Knapp, A.-M., Dembele, D., Ruer-Laventie, J., et al. (2011). B cell signature during inactive systemic lupus is heterogeneous: toward a biological dissection of lupus. *PLoS One* 6:e23900. doi: 10.1371/journal.pone.0023900

Grolleau, A., Kaplan, M. J., Hanash, S. M., Beretta, L., and Richardson, B. (2000). Impaired translational response and increased protein kinase PKR expression in T cells from lupus patients. *J. Clin. Invest.* 106, 1561–1568. doi: 10.1172/JCI9352

Gubbels Bupp, M. R., and Jorgensen, T. N. (2018). Androgen-Induced Immunosuppression. *Front. Immunol.* 9:794. doi: 10.3389/fimmu.2018.00794

Gurevitz, S., Snyder, J., Wessel, E., Frey, J., and Williamson, B. (2013). Systemic lupus erythematosus: a review of the disease and treatment options. *Consult. Pharm.* 28, 110–121. doi: 10.4140/TCP.n.2013.110

Harley, I. T. W., Kaufman, K. M., Langefeld, C. D., Harley, J. B., and Kelly, J. A. (2009). Genetic susceptibility to SLE: new insights from fine mapping and genome-wide association studies. *Nat. Rev. Genet.* 10, 285–290. doi: 10.1038/nrg2571

Hay, N. (2011). Interplay between FOXO, TOR, and Akt. *Biochim. Biophys. Acta (BBA) Mol. Cell Res.* 1813, 1965–1970. doi: 10.1016/j.bbamcr.2011.03.013

Ideker, T., and Sharan, R. (2008). Protein networks in disease. *Genome Res.* 18, 644–652. doi: 10.1101/gr.071852.107

Ikenoue, T., Inoki, K., Yang, Q., Zhou, X., and Guan, K.-L. (2008). Essential function of TORC2 in PKC and Akt turn motif phosphorylation, maturation and signalling. *EMBO J.* 27, 1919–1931. doi: 10.1038/emboj.2008.119

Irizarry, R. A., Bolstad, B. M., Collin, F., Cope, L. M., Hobbs, B., and Speed, T. P. (2003). Summaries of affymetrix genechip probe level data. *Nucleic Acids Res.* 31:e15. doi: 10.1093/nar/gng015

Joseph, N., Reicher, B., and Barda-Saad, M. (2014). The calcium feedback loop and T cell activation: How cytoskeleton networks control intracellular calcium flux. *Biochim. Biophys. Acta (BBA) Biomemb.* 1838, 557–568. doi: 10.1016/j.bbamem.2013.07.009

Keskin, O., Tuncbag, N., and Gursoy, A. (2016). Predicting protein–protein interactions from the molecular to the proteome level. *Chem. Rev.* 116, 4884–4909. doi: 10.1021/acs.chemrev.5b00683

Krogan, N. J., Cagney, G., Yu, H., Zhong, G., Guo, X., Ignatchenko, A., et al. (2006). Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. *Nature* 440, 637–643. doi: 10.1038/nature04670

Kuemmerle, J. F. (2005). Endogenous IGF-I protects human intestinal smooth muscle cells from apoptosis by regulation of GSK-3β activity. *Am. J. Physiol. Gastroin. Liver Physiol.* 288, G101–G110. doi: 10.1152/ajpgi.00032.2004

Kumar, S. U., Kumar, D. T., Siva, R., Doss, C. G. P., and Zayed, H. (2019). Integrative bioinformatics approaches to map potential novel genes and pathways involved in ovarian cancer. *Front. Bioeng. Biotechnol.* 7:391. doi: 10.3389/fbioe.2019.00391

Kurien, B. T., and Scofield, R. H. (2008). Autoimmunity and oxidatively modified autoantigens. *Autoimmun. Rev.* 7, 567–573. doi: 10.1016/j.autrev.2008.04.019

Le, L. Q., Kabarowski, J. H., Weng, Z., Satterthwaite, A. B., Harvill, E. T., Jensen, E. R., et al. (2001). Mice lacking the orphan G protein-coupled receptor G2A develop a late-onset autoimmune syndrome. *Immunity* 14, 561–571. doi: 10.1016/s1074-7613(01)00145-5

Lee, W.-F., Wu, C.-Y., Yang, H.-Y., Lee, W.-I., Chen, L.-C., Ou, L.-S., et al. (2019). Biomarkers associating endothelial Dysregulation in pediatric-onset systemic lupus erythematous. *Pediat. Rheumatol.* 17:69. doi: 10.1186/s12969-019-0369-7

Lee, Y.-S., Jang, H.-S., Kim, J.-M., Lee, J.-S., Lee, J.-Y., Kim, K. L., et al. (2005). Adenoviral-mediated delivery of early growth response factor-1 gene increases tissue perfusion in a murine model of hindlimb ischemia. *Mol. Ther.* 12, 328–336. doi: 10.1016/j.ymthe.2005.03.027

Li, J.-T., Hou, F.-F., Guo, Z.-J., Shan, Y.-X., Zhang, X., and Liu, Z.-Q. (2007). Advanced glycation end products upregulate C-reactive protein synthesis by human hepatocytes through stimulation of monocyte IL-6 and IL-1β production. *Scand. J. Immunol.* 66, 555–562. doi: 10.1111/j.1365-3083.2007.02001.x

Li, Y., Wang, Z., Kong, D., Li, R., Sarkar, S. H., and Sarkar, F. H. (2008). Regulation of Akt/FOXO3a/GSK-3β/AR signaling network by isoflavone in prostate cancer cells. *J. Biol. Chem.* 283, 27707–27716. doi: 10.1074/jbc.M802759200

Lightfoot, Y. L., Blanco, L. P., and Kaplan, M. J. (2017). Metabolic abnormalities and oxidative stress in lupus. *Curr. Opin. Rheumatol.* 29, 442–449. doi: 10.1097/BOR.0000000000000413

Linding, R., Jensen, L. J., Ostheimer, G. J., Vugt, M. A. T. M., van Jørgensen, C., Miron, I. M., et al. (2007). Systematic discovery of in vivo phosphorylation networks. *Cell* 129, 1415–1426. doi: 10.1016/j.cell.2007.05.052

Liu, K.-Q., Bunnell, S. C., Gurniak, C. B., and Berg, L. J. (1998). T cell receptor–initiated calcium release is uncoupled from capacitative calcium entry in itk-deficient T cells. *J. Exp. Med.* 187, 1721–1727. doi: 10.1084/jem.187.10.1721

Liu, Z., Mittanck, D. W., Kim, S., and Rotwein, P. (1995). Control of insulin-like growth factor-II/mannose 6-phosphate receptor gene transcription by proximal

promoter elements. *Mol. Endocrinol.* 9, 1477–1487. doi: 10.1210/mend.9.11.8584025

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

Lu, M. L., Schneider, M. C., Zheng, Y., Zhang, X., and Richie, J. P. (2001). Caveolin-1 interacts with androgen receptor a positive modulator of androgen receptor mediated transactivation. *J. Biol. Chem.* 276, 13442–13451. doi: 10.1074/jbc.M006598200

Mak, A., and Kow, N. Y. (2014). The pathology of T cells in systemic lupus erythematosus. *J. Immunol. Res.* 2014:419029. doi: 10.1155/2014/419029

Markou, T., Cullingford, T. E., Giraldo, A., Weiss, S. C., Alsafi, A., Fuller, S. J., et al. (2008). Glycogen synthase kinases 3α and 3β in cardiac myocytes: regulation and consequences of their inhibition. *Cell. Signal.* 20, 206–218. doi: 10.1016/j.cellsig.2007.10.004

Miret, C., Molina, R., Filella, X., García-Carrasco, M., Claver, G., Ingelmo, M., et al. (2003). Relationship of p53 with other oncogenes, cytokines and systemic lupus erythematosus activity. *TBI* 24, 185–188. doi: 10.1159/000074428

Navé, B. T., Ouwens, M., Withers, D. J., Alessi, D. R., and Shepherd, P. R. (1999). Mammalian target of rapamycin is a direct target for protein kinase B: identification of a convergence point for opposing effects of insulin and amino-acid deficiency on protein translation. *Biochem. J.* 344, 427–431.

Niculescu, F., Nguyen, P., Niculescu, T., Rus, H., Rus, V., and Via, C. S. (2003). Pathogenic T cells in murine lupus exhibit spontaneous signaling activity through phosphatidylinositol 3-kinase and mitogen-activated protein kinase pathways. *Arthr. Rheum.* 48, 1071–1079. doi: 10.1002/art.10900

Nienhuis, H. L., de Leeuw, K., Bijzet, J., Smit, A., Schalkwijk, C. G., Graaff, R., et al. (2008). Skin autofluorescence is increased in systemic lupus erythematosus but is not reflected by elevated plasma levels of advanced glycation endproducts. *Rheumatology (Oxford)* 47, 1554–1558. doi: 10.1093/rheumatology/ken302

Park, J. K., Kim, J.-Y., Moon, J. Y., Ahn, E. Y., Lee, E. Y., Lee, E. B., et al. (2016). Altered lipoproteins in patients with systemic lupus erythematosus are associated with augmented oxidative stress: a potential role in atherosclerosis. *Arthr. Res. Ther.* 18:306. doi: 10.1186/s13075-016-1204-x

Perez-Diez, A., Morgun, A., and Shulzhenko, N. (2013). *Microarrays for Cancer Diagnosis and Classification. Landes Bioscience*. Available online at: https://www.ncbi.nlm.nih.gov/books/NBK6624/ (accessed July 19, 2019).

Pines, A., Bivi, N., Romanello, M., Damante, G., Kelley, M. R., Adamson, E. D., et al. (2005). Cross-regulation between Egr-1 and APE/Ref-1 during early response to oxidative stress in the human osteoblastic HOBIT cell line: Evidence for an autoregulatory loop. *Free Radical Res.* 39, 269–281. doi: 10.1080/10715760400028423

Pons-Estel, G. J., Alarcón, G. S., Scofield, L., Reinlib, L., and Cooper, G. S. (2010). Understanding the epidemiology and progression of systemic lupus erythematosus. *Sem. Arthr. Rheum.* 39, 257–268. doi: 10.1016/j.semarthrit.2008.10.007

Prokunina, L., and Alarcon-Riquelme, M. (2004). The genetic basis of systemic lupus erythematosus—knowledge of today and thoughts for tomorrow. *Hum. Mol. Genet.* 13, R143–R148. doi: 10.1093/hmg/ddh076

Rahman, K. M. T., Islam, M. F., Banik, R. S., Honi, U., Diba, F. S., Sumi, S. S., et al. (2013). *Changes in Protein Interaction Networks Between Normal and Cancer Conditions: Total Chaos or Ordered Disorder? Network Biology*. Available online at: http://agris.fao.org/agris-search/search.do?recordID=CN2013200002 (accessed November 19, 2019).

Razani, B., and Lisanti, M. P. (2001). Caveolins and caveolae: molecular and functional relationships. *Exp. Cell Res.* 271, 36–44. doi: 10.1006/excr.2001.5372

Recio, J. A., and Merlino, G. (2003). Hepatocyte growth factor/scatter factor induces feedback up-regulation of CD44v6 in melanoma cells through Egr-1. *Cancer Res.* 63, 1576–1582.

Rena, G., Guo, S., Cichy, S. C., Unterman, T. G., and Cohen, P. (1999). Phosphorylation of the transcription factor forkhead family member FKHR by protein kinase B. *J. Biol. Chem.* 274, 17179–17183. doi: 10.1074/jbc.274.24.17179

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

Robeva, R., Tanev, D., Andonova, S., Kirilov, G., Savov, A., Stoycheva, M., et al. (2013). Androgen receptor (CAG)n polymorphism and androgen levels in women with systemic lupus erythematosus and healthy controls. *Rheumatol. Int.* 33, 2031–2038. doi: 10.1007/s00296-013-2687-2

Rui, C., Li, C., Xu, W., Zhan, Y., Li, Y., and Yang, X. (2008). Involvement of Egr-1 in HGF-induced elevation of the human 5α-R1 gene in human hepatocellular carcinoma cells. *Biochem. J.* 411, 379–386. doi: 10.1042/BJ20071343

Russo, G., Zegar, C., and Giordano, A. (2003). Advantages and limitations of microarray technology in human cancer. *Oncogene* 22:6497. doi: 10.1038/sj.onc.1206865

Sáenz-Corral, I., Vega-Memíje, M. E., Martínez-Luna, E., Cuevas-González, J. C., Rodríguez-Carreón, A. A., de la Rosa, J. J. B.-G., et al. (2015). Apoptosis in chronic cutaneous lupus erythematosus, discoid lupus, and lupus profundus. *Int. J. Clin. Exp. Pathol.* 8, 7260–7265.

Salas, T. R., Kim, J., Vakar-Lopez, F., Sabichi, A. L., Troncoso, P., Jenster, G., et al. (2004). Glycogen synthase kinase-3β is involved in the phosphorylation and suppression of androgen receptor activity. *J. Biol. Chem.* 279, 19191–19200. doi: 10.1074/jbc.M309560200

Santiago-Raber, M.-L., Lawson, B. R., Dummer, W., Barnhouse, M., Koundouris, S., Wilson, C. B., et al. (2001). Role of cyclin kinase inhibitor p21 in systemic autoimmunity. *J. Immunol.* 167, 4067–4074. doi: 10.4049/jimmunol.167.7.4067

Sauer, L., Gitenay, D., Vo, C., and Baron, V. T. (2010). Mutant p53 initiates a feedback loop that involves Egr-1/EGF receptor/ERK in prostate cancer cells. *Oncogene* 29, 2628–2637. doi: 10.1038/onc.2010.24

Sekuliæ, A., Hudson, C. C., Homme, J. L., Yin, P., Otterness, D. M., Karnitz, L. M., et al. (2000). A Direct linkage between the phosphoinositide 3-kinase-AKT signaling pathway and the mammalian target of rapamycin in mitogen-stimulated and transformed cells. *Cancer Res.* 60, 3504–3513.

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

Shen, M., and Yen, A. (2008). c-Cbl Interacts with CD38 and promotes retinoic acid–induced differentiation and G0 arrest of human myeloblastic leukemia cells. *Cancer Res.* 68, 8761–8769. doi: 10.1158/0008-5472.CAN-08-1058

Shin, S. Y., Chin, B. R., Lee, Y. H., and Kim, J.-H. (2006). Involvement of glycogen synthase kinase-3β in hydrogen peroxide-induced suppression of Tcf/Lef-dependent transcriptional activity. *Cell. Signal.* 18, 601–607. doi: 10.1016/j.cellsig.2005.06.001

Shlomchik, M. J., Craft, J. E., and Mamula, M. J. (2001). From T to B and back again: positive feedback in systemic autoimmune disease. *Nat. Rev. Immunol.* 1, 147–153. doi: 10.1038/35100573

Siess, W., Zangl, K. J., Essler, M., Bauer, M., Brandl, R., Corrinth, C., et al. (1999). Lysophosphatidic acid mediates the rapid activation of platelets and endothelial cells by mildly oxidized low density lipoprotein and accumulates in human atherosclerotic lesions. *Proc. Natl. Acad. Sci. U.S.A.* 96, 6931–6936.

Skoura, A., and Hla, T. (2009). Lysophospholipid receptors in vertebrate development, physiology, and pathology. *J. Lipid Res.* 50, S293–S298. doi: 10.1194/jlr.R800047-JLR200

Smith, S., Fernando, T., Wu, P. W., Seo, J., Ní Gabhann, J., Piskareva, O., et al. (2017). MicroRNA-302d targets IRF9 to regulate the IFN-induced gene expression in SLE. *J. Autoimmu.* 79, 105–111. doi: 10.1016/j.jaut.2017.03.003

Smyth, G. K. (2005). "limma: linear models for microarray data," in *Bioinformatics and Computational Biology Solutions Using R and Bioconductor Statistics for Biology and Health*, eds R. Gentleman, V. J. Carey, W. Huber, R. A. Irizarry, and S. Dudoit (New York, NY: Springer), 397–420. doi: 10.1007/0-387-29362-0_23

Smyth, S. S., Cheng, H.-Y., Miriyala, S., Panchatcharam, M., and Morris, A. J. (2008). Roles of lysophosphatidic acid in cardiovascular physiology and disease. *Biochim. Biophys. Acta (BBA) Mol. Cell Biol. Lipids* 1781, 563–570. doi: 10.1016/j.bbalip.2008.05.008

Szklarczyk, D., Gable, A. L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., et al. (2019). STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* 47, D607–D613. doi: 10.1093/nar/gky1131

Szklarczyk, D., Morris, J. H., Cook, H., Kuhn, M., Wyder, S., Simonovic, M., et al. (2017). The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible. *Nucleic Acids Res.* 45, D362–D368. doi: 10.1093/nar/gkw937

Tan, E. M., Cohen, A. S., Fries, J. F., Masi, A. T., Mcshane, D. J., Rothfield, N. F., et al. (1982). The 1982 revised criteria for the classification of systemic

lupus erythematosus. *Arthr. Rheum.* 25, 1271–1277. doi: 10.1002/art.17802 51101

Tang, E. D., Nuñez, G., Barr, F. G., and Guan, K.-L. (1999). Negative regulation of the forkhead transcription factor FKHR by Akt. *J. Biol. Chem.* 274, 16741–16746. doi: 10.1074/jbc.274.24.16741

Tang, H., Tan, G., Guo, Q., Pang, R., and Zeng, F. (2009). Abnormal activation of the Akt-GSK3β signaling pathway in peripheral blood T cells from patients with systemic lupus erythematosus. *Cell Cycle* 8, 2789–2793. doi: 10.4161/cc.8.17.9446

Thirumal Kumar, D., Jain, N., Evangeline, J., Kamaraj, B., Siva, R., Zayed, H., et al. (2019). A computational approach for investigating the mutational landscape of RAC-alpha serine/threonine-protein kinase (AKT1) and screening inhibitors against the oncogenic E17K mutation causing breast cancer. *Comput. Biol. Med.* 115:103513. doi: 10.1016/j.compbiomed.2019.103513

Tsokos, G. C., Lo, M. S., Reis, P. C., and Sullivan, K. E. (2016). New insights into the immunopathogenesis of systemic lupus erythematosus. *Nat. Rev. Rheumatol.* 12, 716–730. doi: 10.1038/nrrheum.2016.186

Tzivion, G., Dobson, M., and Ramakrishnan, G. (2011). FoxO transcription factors; Regulation by AKT and 14-3-3 proteins. *Biochim. Biophy. Acta (BBA) Mol. Cell Res.* 1813, 1938–1945. doi: 10.1016/j.bbamcr.2011.06.002

Veeranki, S., and Choubey, D. (2010). Systemic lupus erythematosus and increased risk to develop B cell malignancies: role of the p200-family proteins. *Immunol. Lett.* 133, 1–5. doi: 10.1016/j.imlet.2010.06.008

Vences-Catalán, F., Rajapaksa, R., Levy, S., and Santos—Argumedo, L. (2012). The CD19/CD81 complex physically interacts with CD38 but is not required to induce proliferation in mouse B lymphocytes. *Immunology* 137, 48–55. doi: 10.1111/j.1365-2567.2012.03602.x

Wang, J., Liu, Y., Zhao, J., Xu, J., Li, S., and Qin, X. (2017). P-glycoprotein gene MDR1 polymorphisms and susceptibility to systemic lupus erythematosus in Guangxi population: a case-control study. *Rheumatol. Int.* 37, 537–545. doi: 10.1007/s00296-017-3652-2

Wolfrum, C., Besser, D., Luca, E., and Stoffel, M. (2003). Insulin regulates the activity of forkhead transcription factor Hnf-3β/Foxa-2 by Akt-mediated phosphorylation and nuclear/cytosolic localization. *Proc. Natl. Acad. Sci. U.S.A.* 100, 11624–11629. doi: 10.1073/pnas.1931483100

Yamauchi, A., and Bloom, E. T. (1997). Control of cell cycle progression in human natural killer cells through redox regulation of expression and phosphorylation of retinoblastoma gene product protein. *Blood* 89, 4092–4099. doi: 10.1182/blood.V89.11.4092

Yang, L. V., Radu, C. G., Wang, L., Riedinger, M., and Witte, O. N. (2005). Gi-independent macrophage chemotaxis to lysophosphatidylcholine via the immunoregulatory GPCR G2A. *Blood* 105, 1127–1134. doi: 10.1182/blood-2004-05-1916

Yang, W., and Lau, Y. L. (2015). Solving the genetic puzzle of systemic lupus erythematosus. *Pediatr. Nephrol.* 30, 1735–1748. doi: 10.1007/s00467-014-2947-8

Yeh, P. Y., Kuo, S.-H., Yeh, K.-H., Chuang, S.-E., Hsu, C.-H., Chang, W. C., et al. (2006). A pathway for tumor necrosis factor-α-induced Bcl10 nuclear translocation Bcl10 is up-regulated by nf-κb and phosphorylated by Akt1 and then complexes with bcl3 to enter the nucleus. *J. Biol. Chem.* 281, 167–175. doi: 10.1074/jbc.M511014200

Yu, J., Akishita, M., Eto, M., Koizumi, H., Hashimoto, R., Ogawa, S., et al. (2012). Src kinase-mediates androgen receptor-dependent non-genomic activation of signaling cascade leading to endothelial nitric oxide synthase. *Biochem. Biophys. Res. Commun.* 424, 538–543. doi: 10.1016/j.bbrc.2012.06.151

Zhou, Y., Yuan, J., Pan, Y., Fei, Y., Qiu, X., Hu, N., et al. (2009). T cell CD40LG gene expression and the production of IgG by autologous B cells in systemic lupus erythematosus. *Clin. Immunol.* 132, 362–370. doi: 10.1016/j.clim.2009.05.011

Zhu, H., Luo, H., Yan, M., Zuo, X., and Li, Q.-Z. (2015). Autoantigen microarray for high-throughput autoantibody profiling in systemic lupus erythematosus. *Genom. Proteom. Bioinformatics* 13, 210–218. doi: 10.1016/j.gpb.2015.09.001

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 Bioengineering and Biotechnology

ORIGINAL RESEARCH
published: 14 May 2020
doi: 10.3389/fbioe.2020.00428

Let-7i-5p Regulation of Cell Morphology and Migration Through Distinct Signaling Pathways in Normal and Pathogenic Urethral Fibroblasts

Kaile Zhang[1,2†], Ranxin Yang[1,2†], Jun Chen[1,2†], Er Qi[3], Shukui Zhou[1], Ying Wang[1], Qiang Fu[1,2], Rong Chen[1,2] and Xiaolan Fang[4*‡]*

[1] The Department of Urology, Affiliated Sixth People's Hospital, Shanghai Jiao Tong University, Shanghai, China, [2] Shanghai Eastern Institute of Urologic Reconstruction, Shanghai, China, [3] Shanghai Xuhui District Xietu Street Community Health Service Center, Shanghai, China

OPEN ACCESS

Edited by:
Jyoti Sharma,
Institute of Bioinformatics (IOB), India

Reviewed by:
Jing Zhang,
Virginia State University, United States
Richard L. Carpenter,
Indiana University, United States

***Correspondence:**
Xiaolan Fang
fangxiaolan@gmail.com
Rong Chen
chenrongshanghai@126.com

[†] These authors have contributed equally to this work

[‡] **Present address:**
Xiaolan Fang,
Greenwood Genetic Center,
Greenwood, SC, United States

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Bioengineering and Biotechnology

Received: 16 November 2019
Accepted: 14 April 2020
Published: 14 May 2020

Citation:
Zhang K, Yang R, Chen J, Qi E, Zhou S, Wang Y, Fu Q, Chen R and Fang X (2020) Let-7i-5p Regulation of Cell Morphology and Migration Through Distinct Signaling Pathways in Normal and Pathogenic Urethral Fibroblasts. Front. Bioeng. Biotechnol. 8:428. doi: 10.3389/fbioe.2020.00428

microRNAs regulate subcellular functions through distinct molecular mechanisms. In this study, we used normal and pathogenic fibroblasts in pelvic fracture urethral distraction defects (PFUDD) patients. PFUDD is a common disease that could severely affect patients' life quality, yet little is known about the molecular mechanism associated with pathogenic fibrosis in PFUDD. Our data showed that let-7i-5p performs a multi-functional role in distinct signaling transduction pathways involved in cell morphology and cell migration in both normal and pathogenic fibroblasts. By analyzing the molecular mechanism associated with its functions, we found that let-7i-5p regulates through its direct target genes involved in collagen metabolism, cell proliferation and differentiation, TGF-beta signaling, DNA repair and ubiquitination, gene silencing and oxygen homeostasis. We conclude that let-7i-5p plays an essential role in regulating cell shape and tissue elasticity, cell migration, cell morphology and cytoskeleton, and could serve as a potential target for clinical treatment of urethral stricture patients.

Keywords: let-7i-5p, microRNA, cell migration, cell morphology, fibroblast

INTRODUCTION

Pelvic fracture urethral distraction defects (PFUDD) is a common disease that could severely affects patients' life quality, largely due to excessive fibrosis and associated urethral stricture (Zhang et al., 2018). The current incidence of PFUDD is noted to be variable, usually between 5 and 25% of pelvic fractures, with a frequency of 0.32–5/100,000 for men and 0.46–7.25/100,000 for women (Alwaal et al., 2015; Barratt et al., 2018; Dixon et al., 2018). Pelvic fractures resulting in PFUDD has mortality rates between 5 and 33% (Barratt et al., 2018). Fibrosis is a key factor responsible for pathologic changes related to urethral stricture (in both primary or recurrent diseases; Zhao et al., 2018). Over the last few decades, microRNAs and their regulation of fibrosis have been studied in many specific organs, such as liver, heart, skin, kidney, and lung (Jiang et al., 2010, 2017; Vettori et al., 2012; Zhu et al., 2013; Rajasekaran et al., 2015; Li et al., 2016; O'Reilly, 2016; Bagnato et al., 2017). The major interests are in miR-29 and TGF-beta signaling pathway, focusing on their role of molecular

regulation of fibrosis and/or associated excessive extracellular matrix deposition (Rajasekaran et al., 2015; O'Reilly, 2016). miRNAs in specific diseases, such as idiopathic pulmonary fibrosis (IPF), together with their functions in epithelial-mesenchymal transition (EMT) and trans-differentiation, have also been studied (Li et al., 2016). Thus, it would be extremely helpful to further understand molecular mechanisms and related miRNA signaling involved in PFUDD-associated fibrosis, in order to discover novel targets to prevent PFUDD by suppressing urethral stricture.

Our group recently performed molecular profiling of microRNAs in PFUDD patients and summarized a few candidate genes that may serve regulatory functions in fibrosis (Zhang et al., 2018). We found that miR-29 expression is moderate in normal and pathogenic scar tissues in PFUDD patients, and that expression of hsa-miR-29b-3p and hsa-miR-29c-3p were both slightly downregulated in scar tissues (0.64 vs. 0.69, scar vs. normal) from PFUDD patients (Zhang et al., 2018). Interestingly, let-7i-5p expression appeared to be one of the highest among all the microRNAs, and its expression showed an increase in scar tissue comparing to normal tissue (Zhang et al., 2018). Based on raw counts, the expression of global miRNA in normal tissue is 1,522 ± 488 (mean ± standard error), and in scar tissue it is 1,512 ± 483 (mean ± standard error). For let-7i-5p, the expression is 57,325 in normal tissues and 76,083 in scar tissues. Given its impressive abundance in normal and pathogenic fibroblasts, we hypothesize that let-7i-5p may serve as an important regulator in cellular events. Thus, we extended our work of microRNA analysis in PFUDD and discovered let-7i-5p as a novel regulator in multiple cellular events in normal and pathogenic urethral tissues. By up- or down-regulating let-7i-5p in normal human fibroblasts and pathogenic tissues, we evaluated the expression of possible molecular targets involved in those cellular functions (COL1A1, COL3A1, ELN, MMP1, VIM, FN1, ACTIN, TGFBR1, and TIMP1). Our data confirmed that let-7i-5p regulates those cellular events in distinct signaling pathways and the multi-functional regulations are through corresponding downstream target genes. In conclusion, let-7i-5p plays an essential role in regulating cell shape and tissue elasticity, cell migration, cell morphology and cytoskeleton, and it could serve as a potential biomarker and therapeutic target for clinical treatment in PFUDD patients.

MATERIALS AND METHODS

Urethral Scar Samples

The study was approved by the Ethics Committee of Shanghai Sixth People's Hospital. Consents were obtained from all of the patients to use their samples in scientific research. Scar tissues in urethra (Human scar fibroblasts, or HSF) were harvested from PFUDD patients undergoing urethroplasty ($n = 5$). All five subjects are males (gender ratio is 100% male) with ages ranging from 16 to 59. Patients' baseline information was summarized in **Table 1**. The etiology of patients with urethral stricture was PFUDD. All the participating patients underwent primary surgery. The mean length of stricture is 1.5 cm and the locations were all at the membranous segment of urethra. Samples were harvested after surgery, sectioned and stored at −80°C until the process of RNA extraction.

TABLE 1 | Patient baseline characteristics.

Number of subject	Age	Gender	Health status
1	59	Male	PFUDD
2	50	Male	PFUDD
3	16	Male	PFUDD
4	43	Male	PFUDD
5	44	Male	PFUDD

Human Cell Line

Normal human foreskin fibroblasts (HFF, Catalog# SCSP-106) were provided by Stem Cell Bank, Chinese Academy of Sciences.

Cell Transfection

HSF and HFF cells were growing in 10 cm dishes in Dulbecco's MEM (DMEM, Gibco, Cat.# 12100-046, Carlsbad, CA, United States) supplemented with Fetal Bovine Serum (FBS) (Gibco, Cat.#10099-141, Carlsbad, CA, United States) to 10% by volume and Penicillin-Streptomycin (100 μ/mL) (Gibco, Cat.# 15140122, Carlsbad, CA, United States). 80–90% confluency cells were then detached using Trypsin (Gibco, Cat.# 25200056) and plated at 1×10^5 cells/mL in 6-well dishes (2 mL/well), and incubate at 37°C overnight. The cells were transfected with lentiviral constructs (empty control construct, customized lenti-KD miRNA and lenti-OE miRNA from GENECHEM, Shanghai, China) to overexpress or knock down of let-7i-5p according to the manufacturer's protocol. Transfection mixture was replaced by 2 mL DMEM (10% FBS) medium after 12 h. Transfected cells were grown for 72 h before imaging.

Imaging

To confirm stable expression of each transfected construct, GFP expression of transfected cells was observed and evaluated by an Olympus IX70 microscope under fluorescent channel. For regular bright field imaging (for Transwell assay), samples were imaged with the Olympus IX70 microscope under bright light channel.

Cell Migration Assay

Inserts with 8 μM pore size (Corning-Costar, Lowell MA) were used with matching 24-well transwell chambers. Cells were suspended in serum-free DMEM medium and adjusted to 2×10^5 cells/mL. 100–150 μL cell suspension were placed in the upper chambers. The lower chambers were filled with DMEM medium with 10% FBS (600–800 μL/well). Cells were incubated at 37C for 24 h, the inserts were removed and inner side was wiped with cotton swaps. The inserts were then fixed in methanol for 30 min at room temperature, and stained with crystal violet solution (Cat.#A100528-0025, Sangon Biotech Shanghai, China) for 15–30 min and were peeled off after washing and mounted on the slides. The migrated cells were imaged with an OLYMPUS IX70 microscope using bright light channel. Six cell

[HFF (NC (negative control)/SI ((for siRNA-led knockdown)/OE (overexpression), or FNC/FSI/FOE and HSF (NC (negative control)/SI ((for siRNA-led knockdown)/OE (overexpression), or SNC/SSI/SOE] were analyzed and triplicate experiments were done for all the cell types.

Reverse Transcription (RT)-qPCR for miRNAs and Targeted Genes

Total RNA was extracted following standard protocol by Servicebio, Inc. (Wuhan Servicebio Technology Co., Ltd., Wuhan, Hubei, China). The primers used for PCR were designed with Primer Premier software (version 5.0; Premier Biosoft International, Palo Alto, CA, United States; primer sequence details are summarized in **Table 2**). cDNA synthesis was performed on a GeneAmp PCR System 9700 (Applied Biosystems; Thermo Fisher Scientific, Inc., Waltham, MA, United States) following the manufacturer's instructions (RevertAid First Strand cDNA Synthesis kit, Cat.# K1622, Thermo Fisher Scientific, Inc., Waltham, MA, United States). qPCR was performed on a ViiA 7 Real-time PCR System (Applied Biosystems; Thermo Fisher Scientific, Inc.) using a PowerUp SYBR Green Master Mix (Cat.# A25778, Thermo Fisher Scientific, Inc.). The PCR thermal procedure is (1) 95°C, 10min, 2) 95°C, 15s– > 60°C, 60 s, 40 cycles. The fold change for each miRNA was calculated using the $2^{-\Delta\Delta Cq}$ method (Livak and Schmittgen, 2001). U6 expression level was used to normalize the mRNA expression data. Expression in six cell types (HFF (NC/SI/OE) and HSF (NC/SI/OE) were analyzed and triplicate experiments were done for all the cell types.

TABLE 2 | RT-PCR primers.

Gene RefSeq	Primer name	Primer Sequence(5′- > 3′)
NM_001101	H-ACTIN-S	CACCCAGCACAATGAAGATCAAGAT
	H-ACTIN-A	CCAGTTTTTAAATCCTGAGTCAAGC
	U6-S	CTCGCTTCGGCAGCACA
	U6-A	AACGCTTCACGAATTTGCGT
	General control primer-A	TGGTGTCGTGGAGTCG
NM_000090.3	H-COL3A1-S	TTCCTTCGACTTCTCTCCAGCC
	H-COL3A1-A	CCCAGTGTGTTTCGTGCAACC
NM_000501.3	H-ELASTIN-S	GGCATTCCTACTTACGGGGTT
	H-ELASTIN-A	GCTTCGGGGGAAATGCCAAC
NM_212482.2	H-FN1-S	ACACAGAACTATGATGCCGACCA
	H-FN1-A	TGTCCATTCCCCACGACCAT
NM_003380.3	H-VIMENTIN-S	GAAGCCGAAAACACCCTGCAATC
	H-VIMENTIN-A	TGCAGCTCCTGGATTTCCTCT
NM_004612.3	H-TGFBR1-S	GGACCCTTCATTAGATCGCCCTT
	H-TGFBR1-A	CAACTTCTTCTCCCCGCCACT
NM_001145938.1	H-MMP1-S	TACGATTCGGGGAGAAGTGAT
	H-MMP1-A	AAGCCCATTTGGCAGTTGTG
NM_003254.2	H-TIMP1-S	TCCTGTTGTTGCTGTGGCTGAT
	H-TIMP1-A	AAACTCCTCGCTGCGGTTGT
NM_000088.3	H-COL1A1-S	CCAAGACGAAGACATCCCACCA
	H-COL1A1-A	CCGTTGTCGCAGACGCAGAT
MIMAT0000415	hsa-let-7i-5p-RT	CTCAACTGGTGTCGTGGAGTCGG CAATTCAGTTGAGAACAGCAC
	hsa-let-7i-5p-S	ACACTCCAGCTGGGTGAGGTAGT AGTTTGT

Western Blotting

Whole protein lysate was extracted using RIPA Lysis buffer (Cat.#20101ES60, Yeasen, Shanghai, China) following instructions by the manufacturer. Equal concentration of protein was loaded on 5–10% SDS-PAGE gels and transferred onto a PVDF membrane (Cat.# IPVH000010, MilliporeSigma, Burlington, MA, United States). Primary antibodies for COPS6, COPS8, Ago1, Elf1,Tlr4, insulin-like growth factor 1 (somatomedin C) (IGF1), Collagen Type VIII (Collagen8), IL13, Bmp4 and tubulin were listed in **Table 3**. Bands were visualized using horse-radish peroxidase (HRP) conjugated secondary antibodies (**Table 3**) in conjunction with Immobilon ECL Ultra Western HRP Substrate (Cat.#WBKLS0100, MilliporeSigma, Burlington, MA, United States) via ImageQuant LAS 4000mini [HFF (NC/SI/OE) and HSF (NC/SI/OE)] were analyzed and triplicate experiments were done for all the cell types. AlphaEaseFC software (Genetic Technologies Inc., Miami, FL, United States) was used to analyze the density of electrophoretic Western blot bands by Servicebio, Inc. (Wuhan Servicebio Technology Co., Ltd., Wuhan, Hubei, China). GAPDH expression level was used to normalize the protein expression data. Intensity analysis was done for one of the experiments.

Enzyme Linked Immunoabsorbent Assay (ELISA)

Supernatant of cell lysates was collected for each of six cell types [HFF (NC/SI/OE) and HSF (NC/SI/OE)], and the levels of MMP2, TGFβ1, and TIMP1 were quantified using ELISA kits as per manufacturers' instructions (Human Matrix Metalloproteinase 2/Gelatinase A (MMP-2) ELISA Kit, Cat.#CSB-E04675h, CUSABio, Wuhan, Hubei, China; Human TGF-beta1 ELISA Kit, Cat.#EK1811, MultiSciences, Hangzhou, Zhejiang, China; Human TIMP1 ELISA Kit, Cat.#EK11382, MultiSciences).

Statistical Analysis

Student *t*-test were performed for let-7i-5p expression comparison (unpaired, two tails, heteroscedastic) in six cell types (FNC, FSI, FOE, SNC, SSI, and SOE). ANOVA One Way analysis were also performed for validation (**Supplementary File S1**).

Construction of the Let-7i-5p-Target Gene Regulatory Network and Functional Enrichment Analysis

miTarBase database (7.0) was used to predict the target genes of let-7i-5p[1]. The STRING database (www.string-db.org) was used to establish the protein-protein interaction (PPI) network. GO and KEGG pathway enrichment analyses were performed to determine the biological significance of associated proteins. Cytoscape version 3.7.2 was used to visualize the results.

[1]http://mirtarbase.mbc.nctu.edu.tw/php/index.php

TABLE 3 | Commercial antibodies.

Antibody	Target protein	Provider	Catalog number	Dilution
Peroxidase-Conjugated Goat anti-rabbit IgG (H + L)	Rabbit IgG (H + L)	Yeasen	33101ES60	1:5000
Peroxidase-Conjugated Goat anti-mouse IgG (H + L)	Mouse IgG	Yeasen	33201ES60	1:5000
Rabbit Anti-Goat IgG (H + L) HRP	Goat IgG (H + L)	Multisciences	70-RAG007	1:5000
Rabbit-COPS6 Polyclonal Antibody	COPS6	ABclonal	A7072	1:1000
Rabbit-anti-COPS8/COP9 (polyclonal)	COPS8/COP9	Proteintech	10089-2-AP	1:1000
Rabbit-anti-NEDD8 (polyclonal)	NEDD8	Proteintech	16777-1-AP	1:1000
Rabbit-anti-CUL1 (polyclonal)	Cullin-1	Proteintech	12895-1-AP	1:1000
Argonaute 1 (D84G10) XP Rabbit mAb #5053	Ago1	CST	5053T	1:1000
Rabbit-anti-ELF1 (polyclonal)	Elf1	Proteintech	22565-1-AP	1:1000
Mouse-anti-TLR4 (monoclonal)	Tlr4	Proteintech	66350-1-Ig	1:1000
IGF1B-specific polyclonal antibody	Insulin-like growth factor 1	Proteintech	20215-1-AP	1:1000
Rabbit-anti-Collagen Type VIII (polyclonal)	Collagen Type VIII	Proteintech	17251-1-AP	1:1000
Rabbit-anti-IL13 (polyclonal)	IL13	SAB	38354	1:1000
Rabbit-anti-BMP4 (polyclonal)	BMP4	Proteintech	12492-1-AP	1:1000
Mouse-anti -beta Tubulin Mouse mAb	Tubulin	Servibebio	GB13017-2	1:1000

RESULTS

Let-7i-5p Regulates Cell Morphology and Motility in Normal and Pathogenic Fibroblasts

Let-7i-5p is a member of Lethal-7 (let-7) microRNA family, which is widely observed and highly conservative across species, from reptiles to mammals (**Figure 1A**). Let-7 family members were among the first discovered microRNAs and were shown to be an essential regulator of development in *C. elegans* (Reinhart et al., 2000), and let-7 microRNA family has been reported to regulate allergic inflammation through T cells (O'Connell et al., 2012). In human tissues, hsa-let-7i-5p showed extremely high expression in thyroid, and relatively high expression in spinal cord, brain, muscle and vein, with tissue specific index score at 0.905 (indicating a high tissue expression specificity in thyroid; **Figure 1B**; Ludwig et al., 2016), suggesting a potential role of hsa-let-7i-5p in metabolism, fibroblast proliferation and differentiation and tissue development. However, little is known about the role of let-7i-5p or the related molecular mechanism involved in normal fibroblast growth or fibrosis-related scar formation.

Based on a recent miRNA profiling using PFUDD patients' tissues, we found that let-7i-5p expression was really high in both normal and pathogenic fibroblasts based on miRNA sequencing. To manipulate the knock-down or overexpression of let-7i-5p, the miRNA level was up- and down-regulated in normal (HFF) and pathogenic (HSF) fibroblasts using Lenti-viral transfection and significant expression changes were observed (**Figure 1C**). Dysregulation of let-7i-5p in normal fibroblasts caused cell morphology changes yet had little influence on that of pathogenic fibroblasts (**Figure 2**). Surprisingly, either overexpression or knockdown of let-7i-5p resulted in rounder but more spiky cells. Similar phenotypes were reported to be caused by null-functional Dematin (an actin binding/bundling protein), and was associated with null effect in mutant fibroblasts and impaired wound healing process (Mohseni and Chishti, 2008).

To see whether let-7i-5p could regulate cell motility, we performed cell migration assay for normal and pathogenic fibroblasts. Inhibition of let-7i-5p led to a clear promotion of cell motility, while overexpression of let-7i-5p displayed a severely suppression (**Figure 3**). The regulation patterns are similar in both normal and pathogenic fibroblasts (**Figure 3**).

Let-7i-5p Regulates Cellular Processes Through Three Distinct Signaling Pathways

We then performed real time quantitative PCR to evaluate the mRNA expression of potential molecular regulators in those cellular processes. Interestingly, we found that up- or down-regulation of let-7i-5p results in three different regulatory patterns of those genes (**Figure 4** and **Table 4**).

In the first group, let-7i-5p knockdown resulted in decreased mRNA expression of COL1A1, COL3A1, and ELN in normal fibroblasts, while overexpression of let-7i-5p resulted in significantly increased expression of those genes (**Figure 4**, Group 1). This suggests a positive correlation between let-7i-5p and those three genes. Pathogenic status (whether the cell is normal or pathogenic fibroblasts) doesn't seem to affect this regulation, as the positive regulatory pattern is consistent in normal and fibrotic tissues for COL1A1 and ELN. The only exception of COL3A1 in let-7i-5p knockdown pathogenic fibroblasts, which had a slight increased expression instead of down regulation.

In the second group, we observed quite opposite regulatory effects on mRNA expression of MMP1 and VIM by let-7i-5p (**Figure 4**, Group 2). Knockdown of let-7i-5p significantly enhanced MMP1 and VIM expression, while overexpression of let-7i-5p caused a decrease in expression. Strikingly, this negative regulation is completely reversed in pathogenic fibroblasts, as downregulation of let-7i-5p decreased the expression of

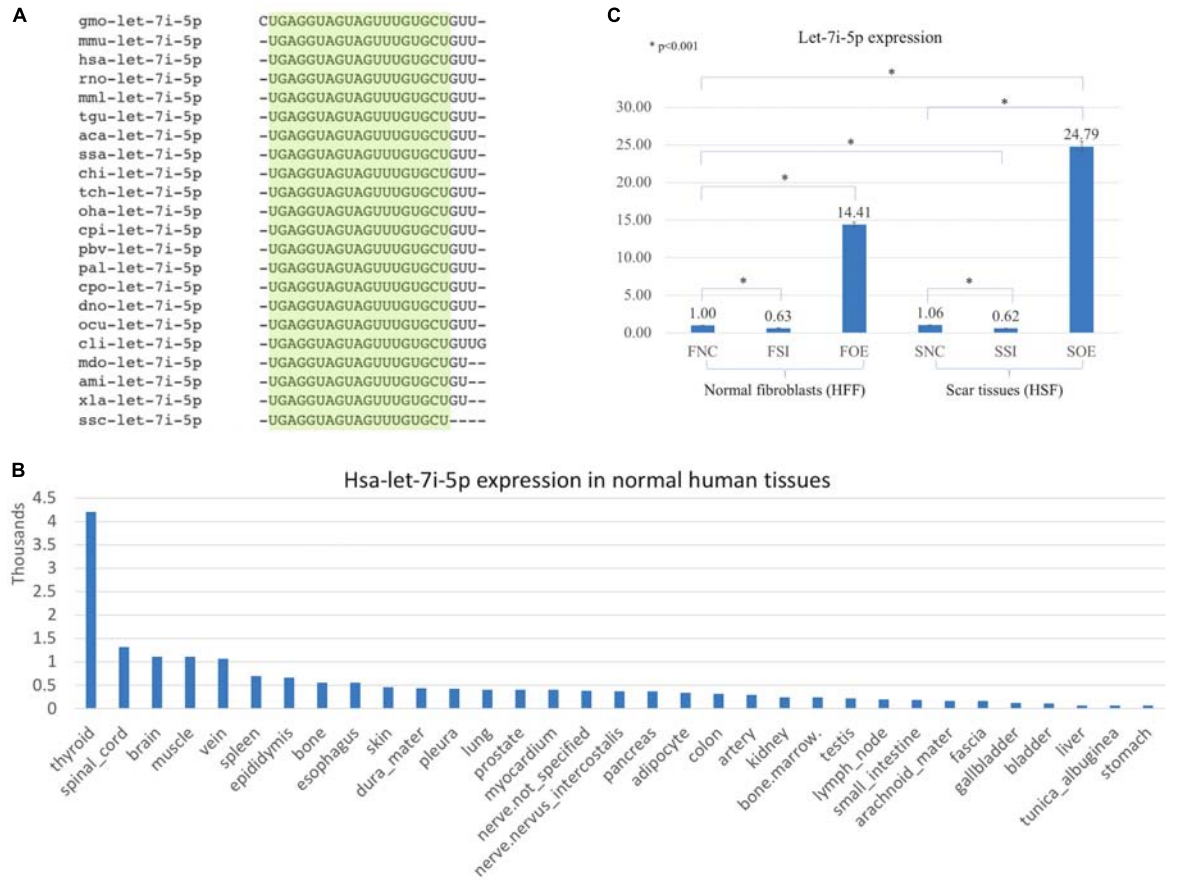


FIGURE 1 | Let-7i-5p is conservative among different species and hsa-let-7i-5p is expressed differentially in normal human tissues. By lenti-viral infection the let-7i-5p expression was manipulated either up or down in normal and scar tissues. **(A)** Let-7i-5p sequence comparison across different species. The conservative sequences are highlighted. aca, Anolis carolinensis; ami, Alligator mississippiensis; chi, Capra hircus; cli, Columba livia; cpi, Chrysemys picta; cpo, Cavia porcellus; dno, Dasypus novemcinctus; gmo, Gadus morhua; hsa, Homo sapiens; mdo, Monodelphis domestica; mmL, Macaca mulatta; mmu, Mus musculus; ocu, Oryctolagus cuniculus; oha, Ophiophagus hannah; pal, Pteropus alecto; pbv, Python bivittatus; rno, Rattus Norvegicus; ssa, Salmo Salar; ssc, Sus scrofa; tch, Tupaia chinensis; tgu, Taeniopygia guttata; xla, Xenopus laevis. **(B)** Hsa-let-7i-5p expression levels in normal human tissues. Data based on two individuals' microRNA sequencing results (Ludwig et al., 2016) and average of normalized value by quantile normalization were used. **(C)** let-7i-5p level was up- and down-regulated in normal and pathogenic fibroblasts by Lenti-viral transfection. *$p < 0.001$. F, normal fibroblasts (HFF). S, scar tissues. NC, non-transfected control. SI, transfected by lenti-KD miRNA to knock down hsa-let-7i-5p expression. OE, transfected by lenti-OE miRNA to overexpress hsa-let-7i-5p.

MMP1 and VIM, and upregulation of let-7i-5p increased their expression. MMP2 expression displayed a similar negative pattern at protein level, although the expression level of MMP2 was almost doubled in control pathogenic fibroblasts (SNC) comparing to normal control cells (FNC) (**Figure 5**). This strongly suggests that regulation of MMP1 and VIM by let-7i-5p is dependent on the pathogenic status of the fibroblasts.

The third group of regulated genes contains FN1, ACTIN, TGFBR1, and TIMP1, which had increased expression with either knockdown or overexpression of let-7i-5p in both normal and pathogenic fibroblasts (**Figure 4**, Group 3). This suggests dysregulated let-7i-5p could boost up the expression of those target genes, regardless of the actual expression change of let-7i-5p (whether it is up- or down-regulated). It is worth noting that the enhanced expression was impressively high for all four target genes in fibrotic cells when let-7i-5p is over-expressed (**Figure 4**), suggesting that pathogenic fibroblasts could further amplify the regulatory effect of those genes resulted from let-7i-5p overexpression, while normal cells still maintained a retraining ability to suppress the dysregulation of target genes caused by let-7i-5p level changes. We also evaluated TGF-beta1, the ligand of TgfbR1, in corresponding cell types, and observed an opposite pattern of regulation in normal fibroblasts (**Figure 5**), which indicates a negative feedback regulation of TGF-beta in response to the TGFbetaR1 protein level changes in normal cells, and this regulation was lost in the pathogenic cells.

Let-7i-5p Regulates Subcellular Functions in Normal Fibroblasts Through Direct Downstream Gene Targets

In normal human fibroblasts, manipulated let-7i-5p expression resulted in positive regulations of COL1A1, COL3A1 and ELN), negative regulations of MMP1 and VIM, and constitutively

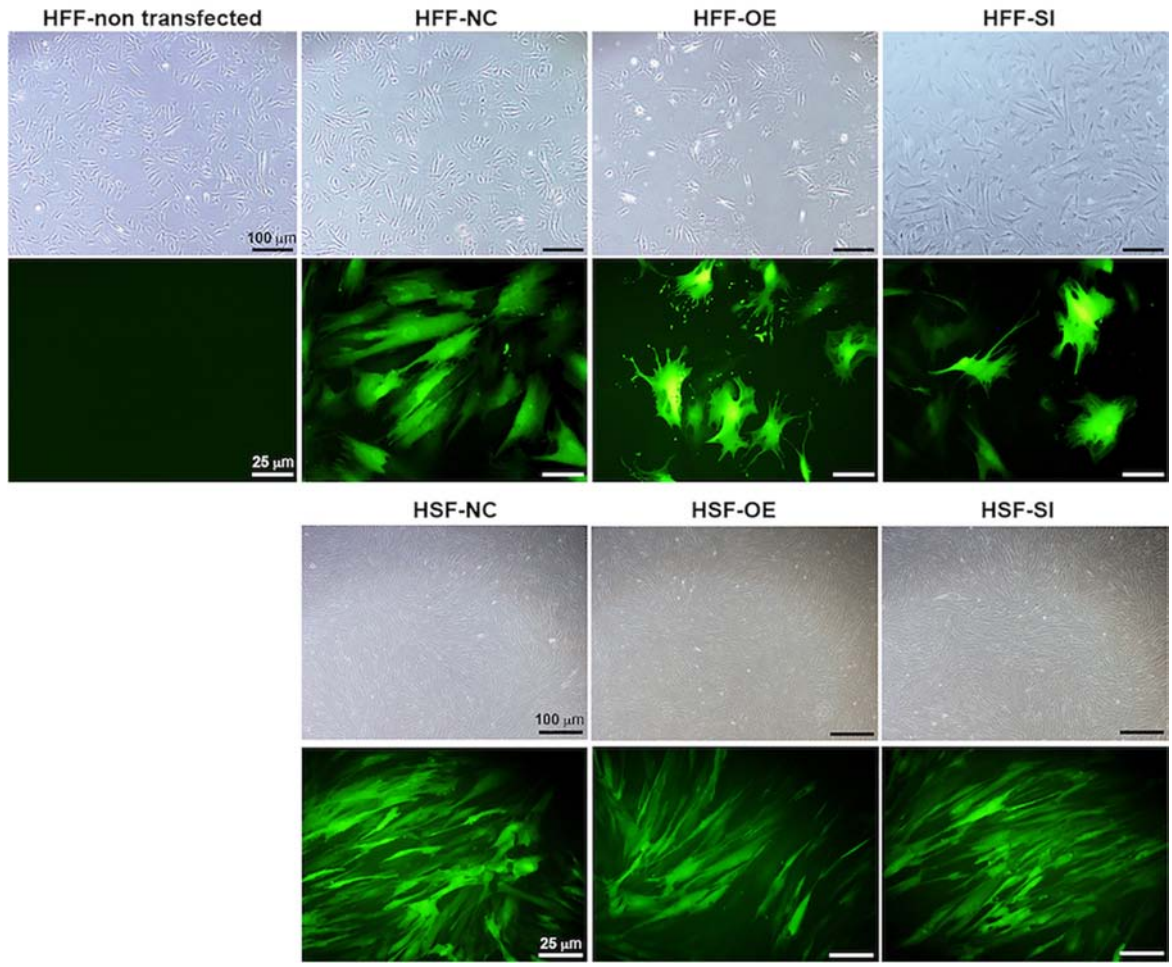


FIGURE 2 | Let-7i-5p level change results in cell morphology changes in normal fibroblasts, but not in pathogenic fibroblasts. Scale bar in bright field images, 100 μm. Scale bar in fluorescent images, 25 μm.

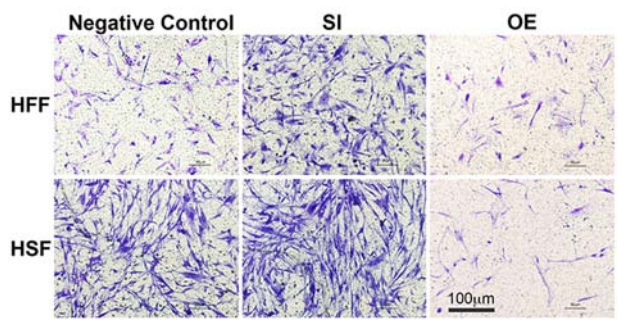


FIGURE 3 | Overexpression of let-7i-5p could suppress migration, while inhibition could promote cell migration in both normal and pathogenic fibroblasts.

increased expression of FN1, ACTIN, TGFBR1, and TIMP1 at mRNA level. Given the fact that those molecules are involved in different signaling transduction pathways and corresponding subcellular functions, we performed functional enrichment analysis for those genes and let-7i-5p. We generated let-7i-5p centered signaling network based on protein-protein interaction and direct binding targets for let-7i-5p (**Figure 6**). Our data strongly suggest a multi-functional role of let-7i-5p in normal fibroblasts, including protein deneddylation, posttranscriptional gene silencing, oxygen homeostatic process (HIF-1 signaling), regulation of fibroblast proliferation, collagen metabolic process, pathogenic *E. coli* infection, neural nucleus development, extracellular matrix disassembly, etc. (**Figure 6**). Among the twenty direct targets predicted *in silico*, we found nine genes (COL8A1, IL13, BMP4, LRIG3, COPS6, COPS8, AGO1, TLR4, and IGF-1) which are predicted to interact with the molecules in the three signaling transduction pathways and might be serving as the connectors. To confirm let-7i-5p regulates through those direct downstream targets, we checked their expression in normal fibroblasts with up- or down-regulated let-7i-5p expression. We confirmed that let-7i-5p could directly regulate collagen metabolic process through COL8A1, fibroblast proliferation and epithelial cell differentiation through IL13, TGFbeta

TABLE 4 | mRNA expressions of let-7i-5p regulated targets.

	Group 1			Group 2		Group 3			
mRNA	**COL1A1**	**COL3A1**	**ELN**	**MMP1**	**VIM**	**FN1**	**ACTIN**	**TGBR1**	**TIMP1**
1 FNC	Control	Control	Control	Control	Control	Control	Control	Control	Control
2 FSI	Down	Down	Down	Up	Up	Up	Up	Up	Up
3 FOE	Up	Up	Up	Down	Down	Up	Up	Up	Up
4 SNC	Control	Control	Control	Control	Control	Control	Control	Control	Control
5 SSI	Down	Up	Down	Down	Down	Up	Up	Up	Up
6 SOE	Up	Up	Up	Up	Up	Up	Up	Up	Up

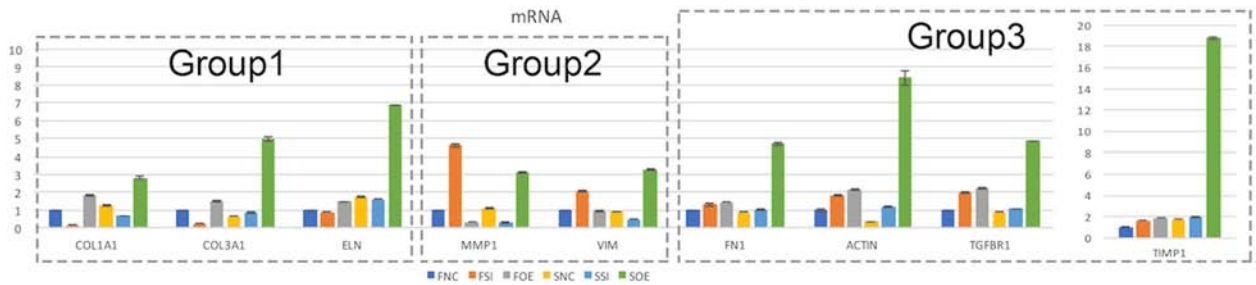


FIGURE 4 | Let-7i-5p regulates signaling molecules in normal and pathogenic fibroblasts. Quantitative analysis of mRNA expression levels of let-7i-5p's downstream targets were performed. Group 1, target genes positively regulated by hsa-let-7i-5p in normal fibroblast cells at mRNA level. Group 2, target genes negatively regulated by hsa-let-7i-5p in normal fibroblast cells at mRNA level. Group 3, target genes constitutively upregulated in normal and pathogenic fibroblasts at mRNA level with either suppressed or increased hsa-let-7i-5p expression. Error bar, standard error. mRNA level expression was evaluated by quantitative real-time PCR.

signaling through BMP4, DNA damage recognition, DNA repair and ubiquitination through LRIG3, COPS6 and COPS8, posttranscriptional gene silencing through AGO1 and ELF1 and oxygen homeostasis through TLR4 and IGF1 (**Figure 7**).

DISCUSSION

Let-7i-5p Regulates Collagen Metabolic Process and Tissue Elasticity Through COL8A1, COL1A1, and COL3A1

The positive regulation of COL1A1, COL3A1, and ELN mRNA expressions in normal fibroblasts was largely retained in pathogenic cells, except for COL3A1, which has a slightly increased expression with decreased let-7i-5p. This indicates that the regulation of those target genes by let-7i-5p was not interrupted in pathogenic fibroblasts, or the functions of let-7i-5p in this specific signaling pathway is independent of the fibrotic status of cells. Since those proteins function together to strengthen and support connective tissues in the body, our data suggest an independent role of let-7i-5p in regulating collagen metabolism and tissue elasticity. This is in concordance with several studies elucidating the association between tissue elasticity and miRNA regulation. For example, in primitive neuroectodermal tumor (PNET) stem cells, tissue elasticity was suggested to promote miRNA silencing and downregulation of target genes (Vu et al., 2015). In mouse models, miR-29-3p could suppress the mRNA expression of COL1A1 and COL1A3 either with or without the induction by TGF-β1 and prevent *S. japonicum*-induced liver fibrosis (Tao et al., 2018). Also, Col1a1 and Col3a1 were overexpressed during active inflammation and murine colitis induced by 2,4,6-trinitrobenzene sulfonic acid (TNBS) hapten (Wu and Chakravarti, 2007). Given that COL8A1 is a direct target of let-7i-5p and it is regulated in the same positive pattern as that for COL1A1, COL1A3 in normal cells, our data suggest a positive regulation pattern of collagen metabolic process by let-7i-5p through COL8A1, COL1A1, and COL3A1 (**Figures 4**, **7**).

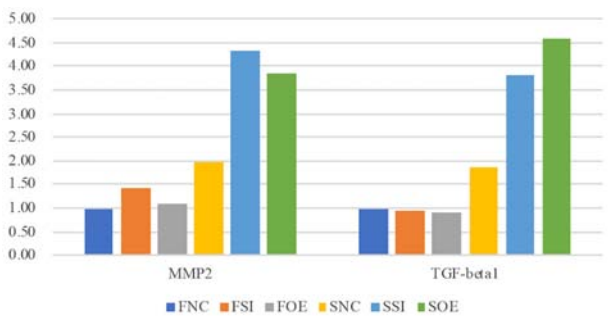


FIGURE 5 | Let-7i-5p regulates MMP2 and TGF-beta1 proteins in normal and pathogenic fibroblasts (ELISA). Protein level expression was evaluated by ELISA.

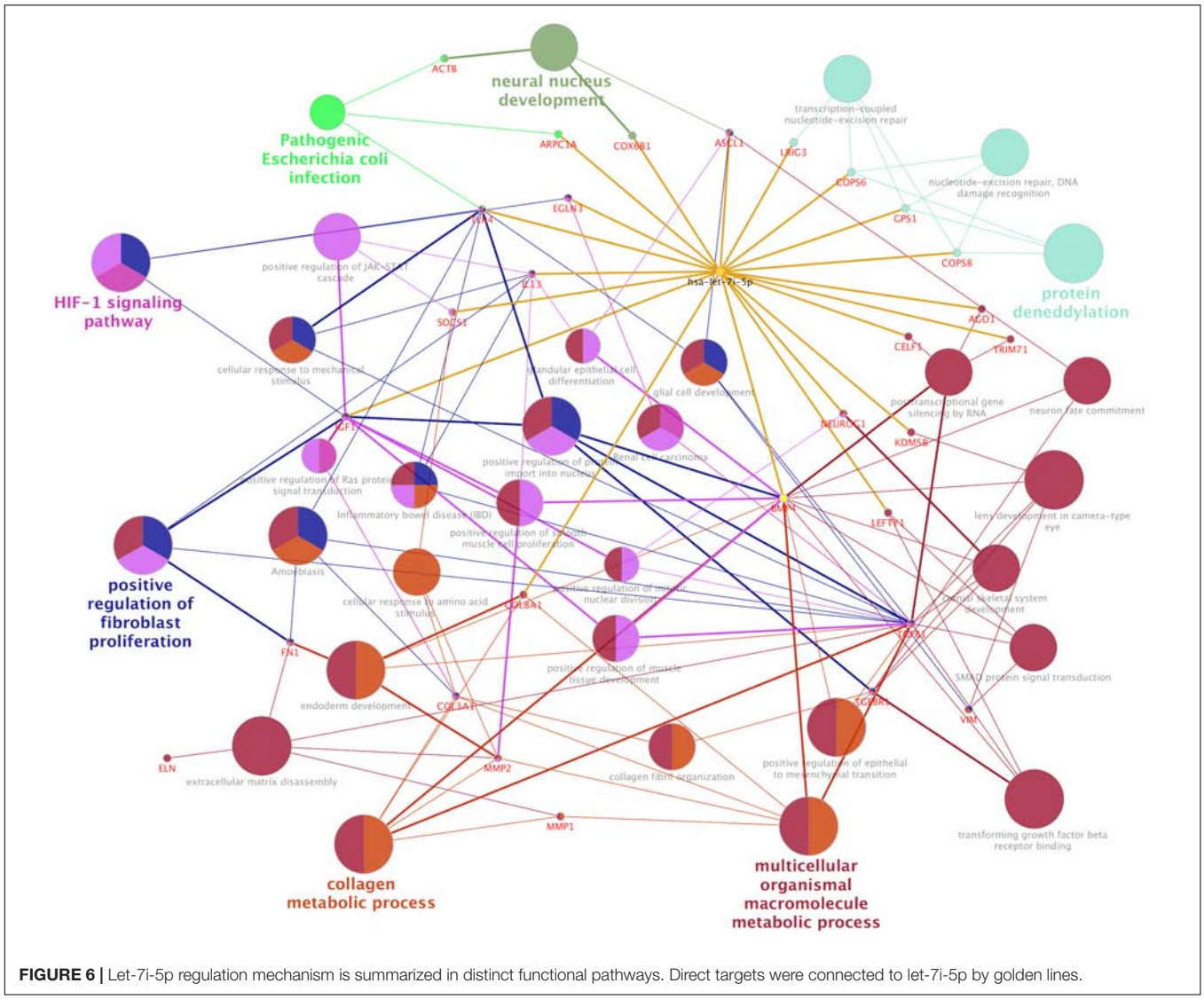


FIGURE 6 | Let-7i-5p regulation mechanism is summarized in distinct functional pathways. Direct targets were connected to let-7i-5p by golden lines.

Let-7i-5p Regulates Extracellular Matrix (ECM) and Cell Migration Through MMP1, MMP2, and Vimentin

Let-7i-5p negatively regulated VIM and MMP1 in normal cells, and this regulation was interrupted in pathogenic fibroblasts, as their mRNA expressions were completely reversed from negative (in normal cells) to positive (in pathogenic cells) regulatory pattern (**Figure 4**). This observation indicates that let-7i-5p functions as an upstream regulator of Vim and MMP1 and its regulation is dependent on pathogenic status of the cells.

TGF-β1 and MMP2 protein levels were up-regulated, together with increased cell migratory capability in pathogenic fibroblasts comparing to normal cells (**Figure 5**). This is consistent with what was reported in human hepatic stellate cells, in which stimulation with TGF-β1 resulted in an increase in migratory capacity and up-regulated MMP-2 activity (Yang et al., 2003). However, in pathogenic fibroblasts with overexpressed let-7i-5p, the cell motility was decreased comparing to control pathogenic fibroblasts with no let-7i-5p change (**Figure 3**, lower right panel and lower left panel), while TGF-β1 and MMP2 expression levels were actually higher in let-7i-5p overexpressed pathogenic cells than that in control pathogenic cells (**Figure 4**). It is worth noting that IL13, a direct downstream target gene of let-7i-5p, could regulate smooth muscle cell proliferation together MMP2 (**Figure 6**), and it is regulated by let-7i-5p in a constitutively negative pattern (**Figure 7**), which is the opposite of that for TGFbR1 (**Figures 4**, **7**).

Let-7i-5p Regulates TGF-Beta Signaling and Fibroblast Proliferation Through BMP4, Fibronectin, Actin, TGFbetaR1, TIMP1, and IL13

The third group of targets regulated by let-7i-5p contains FN1, ACTIN, TGFBR1 and TIMP1, which are overexpressed at mRNA level upon dysregulation of let-7i-5p, no matter let-7i-5p's expression is increased or decreased (**Figure 4A**). The consistent

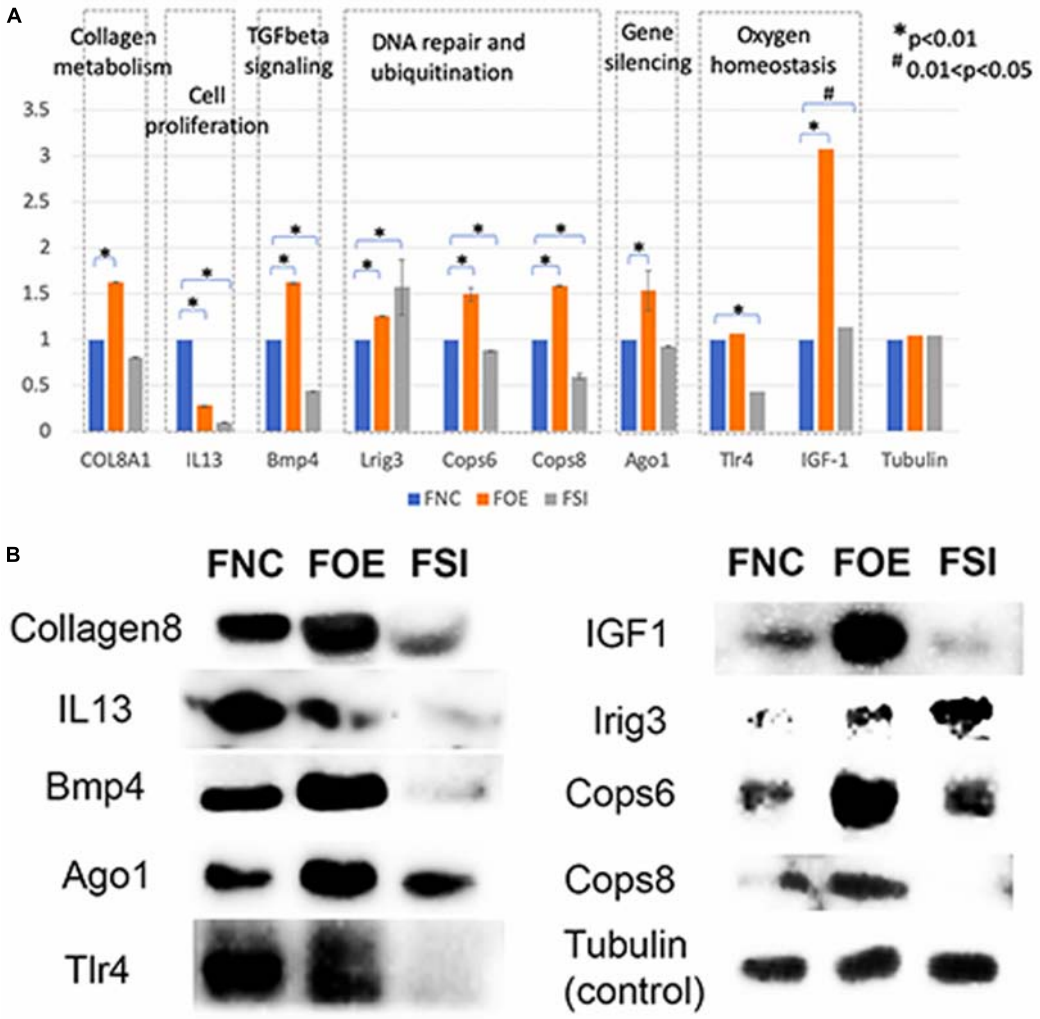


FIGURE 7 | Let-7i-5p regulates direct target genes involved in collagen metabolism, cell proliferation, TGFbeta signaling, DNA repair and ubiquitination, gene silencing and oxygen homeostasis. **(A)** Quantitative analysis of protein expression levels of target genes. Y axis represents the expression change normalized by FNC for each protein. **(B)** Representative western blot images.

pattern in normal and pathogenic fibroblasts implies that the fibrotic status of cells doesn't affect the regulation by let-7i-5p. Our data suggested that BMP4 is directly regulated by let-7i-5p, yet the regulation pattern is not similar to TGFBR1 or TGFbeta1 and let-7i-5p may regulate TGF-beta signaling in a parallel route that is independent from its regulation of BMP4.

Disruption of let-7i-5p regulation would result in morphological changes in normal fibroblasts (**Figure 2**), yet the mechanism remains unclear. In literature, let-7i-5p dysregulation phenotype mimics that of Dematin mutations (Mohseni and Chishti, 2008). Dematin is an actin binding/bundling protein that regulates FAK activation through RhoA and regulate cell morphology (Mohseni and Chishti, 2008) and is predicted to be a conserved target of miR181a-5p, miR181b-5p, miR181c-5p, miR181d-5p, and miR-4262 in human, yet little is known about its regulation by those miRNAs. In addition, it was reported that MMP1 overexpression could suppress Thioacetamide (TAA)-induced liver fibrosis in rat model (Iimuro et al., 2003). TIMP1's function in fibrosis has been in argument as the evidences are divergent from different studies, although many of the results suggests that its expression has no effect on fibrosis (Giannandrea and Parks, 2014).

Interestingly, we found that dysregulated let-7i-5p could result in suppression of IL13, which is a positive regulator of fibroblast proliferation and epithelial cell differentiation (**Figure 7**). It may also play a role in JAK-STAT cascade and inflammatory response (**Figure 6**).

Let-7i-5p and Its Potential Functions in DNA Repair and Ubiquitination, Gene Silencing and Oxygen Homeostasis

Hsa-let-7i-5p is predicted to target on several post-translational pathways, such as DNA repair and ubiquitination (through

TABLE 5 | miRNA associated clinical trials in urological diseases.

NCT Number	Status	Study title	Conditions	Type
NCT02470507	Active, not recruiting	Immune Function in Acute Kidney Injury	Acute Kidney Failure	General miRNA profile, observational study
NCT02289040	Completed	Acute Kidney Injury Following Paediatric Cardiac Surgery	Acute Kidney Injury	General miRNA profile, in microvesicles
NCT02315183	Completed	An Observational Case Control Study to Identify the Role of MV and MV Derived Micro-RNA in Post CArdiac Surgery AKI	Acute Kidney Injury	General miRNA profile, observational study
NCT03373786	Completed	A Study of RG-012 in Subjects With Alport Syndrome	Alport Syndrome	miR-21, renal
NCT00743054	Completed	microRNA Expression in Renal Cell Carcinoma	Carcinoma, Renal Cell	General miRNA profile, observational study
NCT03227055	Unknown	Cardiovascular Comorbidity in Children With Chronic Kidney Disease	Childhood Chronic Kidney Disease	urine exosome miRNA
NCT01114594	Completed	Pilot Study of RNA as a Biomarker for Autosomal Dominant Polycystic Kidney Disease	Chronic Kidney Disease Polycystic Kidney, Autosomal Dominant	General miRNA profile, urine, observational study
NCT02147782	Recruiting	Clinical Observation on Bone Metabolism Induced by Chronic Renal Insufficiency	Chronic Renal Insufficiency Renal Osteodystrophy	General miRNA profile, observational study
NCT02410876	Recruiting	Changes of microRNA Expression in Obstructive and Neurogenic Bladder Dysfunction	Disorder of the Lower Urinary Tract	General miRNA profile, comparison between BLUTD (bladder outlet obstruction (BOO)-induced) and NLUTD (neurogenic)
NCT00806650	Completed	Anti-IMP3 Autoantibody and MicroRNA Signature Blood Tests in Finding Metastasis in Patients With Localized or Metastatic Kidney Cancer	Kidney Cancer	General, miRNA profile, serum, observational study
NCT03089242	Unknown	MicroRNAs in Acute Kidney Injury	Kidney Injury in Cardiac Surgery - Expression of microRNAs	General miRNA profile
NCT01731158	Unknown	Sequential Therapy With Bevacizumab, RAd001 (Everolimus) and Tyrosinekinase Inhibitors (TKI) in Metastatic Renal Cell Carinoma (mRCC)	Metastatic Renal Cell Carcinoma	General miRNA profile
NCT03235128	Unknown	Clinical Significance of Assesment of Serum miRNA-30a in Childhood Nephrotic Syndrome	Nephrotic Syndrome Steroid-Resistant	miRNA-30a, serum, observational study
NCT00565903	Active, not recruiting	Elucidating the Genetic Basis of the Pleuropulmonary Blastoma (PPB) Familial Cancer Syndrome	Cystic Nephroma Pleuropulmonary Blastoma Sertoli-Leydig Cell Tumor of Ovary Medulloepithelioma Embryonal Rhabdomyosarcoma of Cervix Goiter Sarcoma Pineoblastoma Pituitary Tumors Wilms Tumor	General miRNA profile, observational study
NCT01482676	Completed	The Role of microRNAs in Organ Remodeling in Lower Urinary Tract Dysfunction	Urinary Bladder Neck Obstruction Cystitis, Interstitial Prostatic Hyperplasia	General miRNA profile, observational study
NCT02316522	Active, not recruiting	Epigenetic Contribution to the Pathogenesis of Diabetic Nephropathy in Qatari Population	Type 2 Diabetes	General miRNA profile, observational study
NCT01973088	Unknown	Screening and Identification of Human Urate Transporter hURAT1 MicroRNA	Urinary Calculi	miRNAs regulated by hURAT1
NCT03511924	Completed	Intradialytic Resistance Training in Haemodialysis Patients	Chronic Kidney Disease Requiring Chronic Dialysis	Renal specific miRNA profile
NCT03591367	Completed	The Potential Role Of MicroRNA-155 And Telomerase Reverse Transcriptase In Diagnosis Of Non-Muscle Invasive Bladder Cancer And Their Pathological Correlation	Bladder Cancer; Bladder Disease; Bladder Neoplasm; Micro-RNA	MicroRNAs-155
NCT04176276	Recruiting	Determining Serum and Urinary Levels of miRNA 192 and miRNA 25 in Patients With and Without Type 2 Diabetes.	Diabetic Kidney Disease; Type2 Diabetes	miR-192 and miR-25

(Continued)

TABLE 5 | Continued

NCT Number	Status	Study title	Conditions	Type
NCT03924089	Recruiting	Oral Nutritional Supplement on Nutritional and Functional Status, and Biomarkers in Malnourished Hemodialysis Patients.	Malnutrition; End Stage Renal Disease	Circulating miRNAs
NCT01829971	Terminated	A Multicenter Phase I Study of MRX34, MicroRNA miR-RX34 Liposomal Injection	Primary Liver Cancer; SCLC; Lymphoma; Melanoma; Multiple Myeloma; Renal Cell Carcinoma; NSCLC	liposomal miR-34a mimic
NCT03942744	Recruiting	The Effect of High-flux Hemodialysis and On-line Hemodiafiltration on Endothelial Function.	Chronic Kidney Disease Requiring Chronic Dialysis	General miRNA profile
NCT04300387	Recruiting	Chronic Kidney Disease at Northeast Taiwan: Biomarker and Multidisciplinary Care	Chronic Kidney Disease	General miRNA profile
NCT02593526	Recruiting	Diuretic/Cool Dialysate Trial	Chronic Kidney Insufficiency	General miRNA profile
NCT03202212	Completed	Effect of Mixed On-line Hemodiafiltration on Circulating Markers of Inflammation and Vascular Dysfunction	Chronic Kidney Failure; Dialysis Related Complication	General miRNA profile in plasmatic exosomes or microvesicles
NCT03780101	Recruiting	Pathology and Imaging in Kidney Allografts	Renal Transplant Rejection; Chronic Kidney Diseases; Fibrosis	miR-214, miR-21 and miR-29
NCT03476460	Completed	Sodium Chloride and Contrast Nephropathy	Kidney Failure, Chronic; Kidney Failure, Acute; Heart Failure; Diabetes	General miRNA profile
NCT03844412	Suspended	Vestibulodynia: Understanding Pathophysiology and Determining Appropriate Treatments	Vestibulodynia; Temporomandibular Disorder; Fibromyalgia Syndrome; Irritable Bowel Syndrome; Migraines; Tension Headache; Endometriosis; Interstitial Cystitis; Back Pain; Chronic Fatigue Syndrome	General miRNA profile
NCT03651388	Completed	Research Into the Molecular Bases of a New Phenotype Combining Premature White Hair, Polycystic Kidney Disease, Aortic Dilation/Dissection and Lymphopenia	New Phenotype (Combining Premature White Hair, Polycystic Kidney Disease, Aortic Dilation/Dissection and Lymphopenia)	Bcl-2-regulating miRNAs
NCT03246191	Unknown status	Screening and Assessing the Risk Factors and Complications of Chronic Kidney Disease	Chronic Kidney Disease	General miRNA profile, circulating microRNA
NCT02691546	Unknown status	Screening for Chronic Kidney Disease (CKD) Among Older People Across Europe (SCOPE)	Chronic Kidney Diseases	General miRNA profile, circulating microRNA

predicted target genes LRIG3, COPS6, COPS8, etc.) and gene silencing by RNA/miRNA (through target genes such as AGO1). It is also involved in homeostatic process [through predicted target genes TLR4 and IGF1, members of HIF-1 signaling pathway (Prabhakar and Semenza, 2015)] (**Figure 6**). Our data confirmed that let-7i-5p serves as a positive regulator of COPS6, COPS8, Ago1, and IGF-1 (**Figure 7**). LRIG3 is regulated in a quite different pattern comparing to COPS6 and COPS8, suggesting that let-7i-5p could regulate transcription-coupled nucleotide-excision repair (through LRIG3) separately from DNA damage recognition and protein deneddylation (through COPS6 and COPS8).

Potential Clinical Applications of Hsa-let-7i-5p and MicroRNAs in PFUDD and Urological Diseases

miRNAs have been discussed as potential therapeutic targets and clinical biomarkers in various diseases (Mlcochova et al., 2015; Christopher et al., 2016; Ji et al., 2017). microRNAs are under investigation in a number of recent clinical trials for various urologic complications (e.g., urinary bladder neck obstruction, urolithiasis, urinary tract disorders, renal carcinoma, kidney injury, etc. (**Table 5**), mainly by microRNA profiling in patients, with a few extended into studies targeting on a specific microRNA (such as miR-21). A recent clinical trial (NCT02639923) evaluates the correlation between serum let-7i expression and intracranial traumatic lesions, which is based on evidence from animal models (Balakathiresan et al., 2012). The unique regulatory functions of let-7i-5p in fibroblast proliferation, ECM regulation and homeostasis makes it an interesting drug target for complications involved with fibrosis, tissue reconstruction and cellular stress (**Figure 6**). We hope the data from this study could broaden our understanding of the function of hsa-let-7i-5p in normal and pathogenic fibroblasts and urethral tissues, so as to facilitate clinical diagnosis, treatment as well as tissue engineering as follow-up options for patients with PFUDD or other urological diseases.

CONCLUSION

In this study, we analyzed let-7i-5p and its potential downstream targets (COL1A1, COL3A1, ELN, MMP1, VIM, FN1, ACTIN,

TGFBR1, TIMP1, MMP2) in both normal and pathogenic fibroblasts. We found that let-7i-5p could regulate various signaling pathways and serve distinct functions in different cellular events, including tissue plasticity, cell motility and cell morphology. By functional enrichment analysis, we evaluated the potential direct targets of let-7i-5p that might be responsible for each signaling cascade. We conclude that let-7i-5p is a multi-functional regulator, and it could be affected by fibrosis and pathogenic status.

DATA AVAILABILITY STATEMENT

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

ETHICS STATEMENT

The experiments involving human participants were reviewed and approved by the Ethics Committee of Shanghai Sixth People's Hospital. Consents were obtained from all of the patients to participate in this study.

AUTHOR CONTRIBUTIONS

KZ, JC, SZ, RY, and YW: clinical sample collection and experiments. KZ, XF, QF, and RC: experimental design. XF and EQ: data analysis. KZ, QF, and RC: reagents, materials, and analysis tools contribution. XF, KZ, QF, and RC: manuscript writing.

FUNDING

This study was supported by National Natural Science Fund of China (Grant No. 81700590), the Science and Technology Commission of Shanghai (Grant No. 17410742800), the Shanghai Jiao Tong University Biomedical Engineering Cross Research Foundation (Grant No. YG2017QN15), Shanghai health committee (20184Y0053), Shanghai Rising stars of medical talent Youth development program and Shanghai Jiao Tong University K. C. Wong Medical Fellowship Fund. A previous version of this manuscript has been released as a Pre-Print at bioRxiv (https://doi.org/10.1101/330332).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fbioe.2020.00428/full#supplementary-material

FILE S1 | Raw data used for **Figures 1, 4,** and **7** and associated student *t*-test and ANOVA One Way analysis.

REFERENCES

Alwaal, A., Zaid, U. B., Blaschko, S. D., Harris, C. R., Gaither, T. W., McAninch, J. W., et al. (2015). The incidence, causes, mechanism, risk factors, classification, and diagnosis of pelvic fracture urethral injury. *Arab J. Urol.* 13, 2–6. doi: 10.1016/j.aju.2014.08.006

Bagnato, G., Roberts, W. N., Roman, J., and Gangemi, S. (2017). A systematic review of overlapping microRNA patterns in systemic sclerosis and idiopathic pulmonary fibrosis. *Eur. Respir. Rev.* 26:160125. doi: 10.1183/16000617.0125-2016

Balakathiresan, N., Bhomia, M., Chandran, R., Chavko, M., McCarron, R. M., and Maheshwari, R. K. (2012). MicroRNA Let-7i is a promising serum biomarker for blast-induced traumatic brain injury. *J. Neurotrauma* 29, 1379–1387. doi: 10.1089/neu.2011.2146

Barratt, R. C., Bernard, J., Mundy, A. R., and Greenwell, T. J. (2018). Pelvic fracture urethral injury in males—mechanisms of injury, management options and outcomes. *Transl. Androl. Urol.* 7, S29–S62. doi: 10.21037/tau.2017.12.35

Christopher, A. F., Kaur, R. P., Kaur, G., Kaur, A., Gupta, V., and Bansal, P. (2016). MicroRNA therapeutics: discovering novel targets and developing specific therapy. *Perspect. Clin. Res.* 7, 68–74. doi: 10.4103/2229-3485.179431

Dixon, A. N., Webb, J. C., Wenzel, J. L., Wolf, J. S. J., and Osterberg, E. C. (2018). Current management of pelvic fracture urethral injuries: to realign or not? *Transl. Androl. Urol.* 7, 593–602. doi: 10.21037/tau.2018.01.14

Giannandrea, M., and Parks, W. C. (2014). Diverse functions of matrix metalloproteinases during fibrosis. *Dis. Models Mech.* 7, 193–203. doi: 10.1242/dmm.012062

Iimuro, Y., Nishio, T., Morimoto, T., Nitta, T., Stefanovic, B., Choi, S. K., et al. (2003). Delivery of matrix metalloproteinase-1 attenuates established liver fibrosis in the rat. *Gastroenterology* 124, 445–458. doi: 10.1053/gast.2003.50063

Ji, W., Sun, B., and Su, C. (2017). Targeting MicroRNAs in cancer gene therapy. *Genes* 8:21. doi: 10.3390/genes8010021

Jiang, X., Tsitsiou, E., Herrick, S. E., and Lindsay, M. A. (2010). microRNAs and the regulation of fibrosis. *FEBS J.* 277, 2015–2021. doi: 10.1111/j.1742-4658.2010.07632.x

Jiang, X.-P., Ai, W.-B., Wan, L.-Y., Zhang, Y.-Q., and Wu, J.-F. (2017). The roles of microRNA families in hepatic fibrosis. *Cell Biosci.* 7:34. doi: 10.1186/s13578-017-0161-7

Li, H., Zhao, X., Shan, H., and Liang, H. (2016). MicroRNAs in idiopathic pulmonary fibrosis: involvement in pathogenesis and potential use in diagnosis and therapeutics. *Acta Pharm. Sin. B* 6, 531–539. doi: 10.1016/j.apsb.2016.06.010

Livak, K. J., and Schmittgen, T. D. (2001). Analysis of relative gene expression data using real-time quantitative PCR and the 2-ΔΔCT method. *Methods* 25, 402–408.

Ludwig, N., Leidinger, P., Becker, K., Backes, C., Fehlmann, T., Pallasch, C., et al. (2016). Distribution of miRNA expression across human tissues. *Nucleic Acids Res.* 44, 3865–3877. doi: 10.1093/nar/gkw116

Mlcochova, H., Hezova, R., Meli, A. C., and Slaby, O. (2015). Urinary MicroRNAs as a new class of noninvasive biomarkers in oncology, nephrology, and cardiology. *Methods Mol Biol.* 1218, 439–463. doi: 10.1007/978-1-4939-1538-5_26439-463

Mohseni, M., and Chishti, A. H. (2008). The headpiece domain of dematin regulates cell shape, motility, and wound healing by modulating RhoA activation. *Mol. Cell. Biol.* 28, 4712–4718. doi: 10.1128/MCB.00237-08

O'Connell, R. M., Rao, D. S., and Baltimore, D. (2012). microRNA regulation of inflammatory responses. *Annu. Rev. Immunol.* 30, 295–312. doi: 10.1146/annurev-immunol-020711-075013

O'Reilly, S. (2016). MicroRNAs in fibrosis: opportunities and challenges. *Arthritis Res. Ther.* 18:11. doi: 10.1186/s13075-016-0929-x

Prabhakar, N. R., and Semenza, G. L. (2015). Oxygen sensing and homeostasis. *Physiology* 30, 340–348. doi: 10.1152/physiol.00022.2015

Rajasekaran, S., Rajaguru, P., and Sudhakar Gandhi, P. S. (2015). MicroRNAs as potential targets for progressive pulmonary fibrosis. *Front. Pharmacol.* 6:254. doi: 10.3389/fphar.2015.00254

Reinhart, B. J., Slack, F. J., Basson, M., Pasquinelli, A. E., Bettinger, J. C., Rougvie, A. E., et al. (2000). The 21-nucleotide let-7 RNA regulates developmental timing in *Caenorhabditis elegans*. *Nature* 403, 901. doi: 10.1038/35002607

Tao, R., Fan, X.-X., Yu, H.-J., Ai, G., Zhang, H.-Y., Kong, H.-Y., et al. (2018). MicroRNA-29b-3p prevents *Schistosoma japonicum*-induced liver fibrosis by targeting COL1A1 and COL3A1. *J. Cell. Biochem.* 119, 3199–3209. doi: 10.1002/jcb.26475

Vettori, S., Gay, S., and Distler, O. (2012). Role of MicroRNAs in fibrosis. *Open Rheumatol. J.* 6, 130–139. doi: 10.2174/1874312901206010130

Vu, L. T., Keschrumrus, V., Zhang, X., Zhong, J. F., Su, Q., Kabeer, M. H., et al. (2015). Tissue elasticity regulated tumor gene expression: implication for diagnostic biomarkers of primitive neuroectodermal tumor. *PLoS One* 10:e0120336. doi: 10.1371/journal.pone.0120336

Wu, F., and Chakravarti, S. (2007). Differential expression of inflammatory and fibrogenic genes and their regulation by NF-κB inhibition in a mouse model of chronic colitis. *J. Immunol.* 179:6988. doi: 10.4049/jimmunol.179.10.6988

Yang, C., Zeisberg, M., Mosterman, B., Sudhakar, A., Yerramalla, U., Holthaus, K., et al. (2003). Liver fibrosis: insights into migration of hepatic stellate cells in response to extracellular matrix and growth factors. *Gastroenterology* 124, 147–159. doi: 10.1053/gast.2003.50012

Zhang, K., Chen, J., Zhang, D., Wang, L., Zhao, W., Lin, D. Y., et al. (2018). microRNA expression profiles of scar and normal tissue from patients with posterior urethral stricture caused by pelvic fracture urethral distraction defects. *Int. J. Mol. Med.* 41, 2733–2743. doi: 10.3892/ijmm.2018.3487

Zhao, J., Ren, L., Liu, M., Xi, T., Zhang, B., and Yang, K. (2018). Anti-fibrotic function of Cu-bearing stainless steel for reducing recurrence of urethral stricture after stent implantation. *J. Biomed. Mater. Res. Part B Appl. Biomater.* 106, 2019–2028. doi: 10.1002/jbm.b.34005n/a-n/a

Zhu, H., Luo, H., and Zuo, X. (2013). MicroRNAs: their involvement in fibrosis pathogenesis and use as diagnostic biomarkers in scleroderma. *Exp. Amp Mol. Med.* 45:e41. doi: 10.1183/16000617.0125-2016

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 Bioengineering and Biotechnology

ORIGINAL RESEARCH
published: 29 May 2020
doi: 10.3389/fbioe.2020.00522

Truncation of C-Terminal Intrinsically Disordered Region of Mycobacterial Rv1915 Facilitates Production of "Difficult-to-Purify" Recombinant Drug Target

Monika Antil[1], *Sébastien G. Gouin*[2] *and Vibha Gupta*[1]*

[1] Department of Biotechnology, Jaypee Institute of Information Technology, Noida, India, [2] CEISAM, Chimie Et Interdisciplinarité, Synthèse, Analyse, Modélisation, UMR CNRS 6230, UFR des Sciences et des Techniques, Université de Nantes, Nantes, France

OPEN ACCESS

Edited by:
Manoj Kumar Kashyap,
Amity University Gurgaon, India

Reviewed by:
Vikram Saini,
All India Institute of Medical Sciences, India
Rahul Singh,
University of Pennsylvania, United States
Vineel P. Reddy,
University of Alabama at Birmingham, United States

***Correspondence:**
Vibha Gupta
vibha.gupta@jiit.ac.in

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Bioengineering and Biotechnology

Received: 25 January 2020
Accepted: 01 May 2020
Published: 29 May 2020

Citation:
Antil M, Gouin SG and Gupta V (2020) Truncation of C-Terminal Intrinsically Disordered Region of Mycobacterial Rv1915 Facilitates Production of "Difficult-to-Purify" Recombinant Drug Target. Front. Bioeng. Biotechnol. 8:522. doi: 10.3389/fbioe.2020.00522

Availability of purified drug target is a prerequisite for its structural and functional characterization. However, aggregation of recombinant protein as inclusion bodies (IBs) is a common problem during the large scale production of overexpressed protein in heterologous host. Such proteins can be recovered from IB pool using some mild solubilizing agents such as low concentration of denaturants or detergents, alcohols and osmolytes. This study reports optimization of solubilization buffer for recovery of soluble and biologically active recombinant mycobacterial Rv1915/ICL2a from IBs. Even though the target protein could be solubilized successfully with mild agents (sarcosine and βME) without using denaturants, it failed to bind on Ni-NTA resin. The usual factors such as loss of His6-tag due to proteolysis, masking of the tag due to its location or protein aggregation were investigated, but the actual explanation, provided through bioinformatics analysis, turned out to be presence of intrinsically disordered protein regions (IDPRs) at the C-terminus. These regions due to their inability to fold into ordered structure may lead to non-specific protein aggregation and hence reduced binding to Ni-NTA affinity matrix. With this rationale, 90 residues from the C-terminal of Rv1915/ICL2 were truncated, the variant successfully purified and characterized for ICL and MICL activities, supporting the disordered nature of Rv1915/ICL2a C-terminal. When a region that has definite structure associated in some mycobaterial strains such as CDC 1551 and disorder in others for instance *Mycobacterium tuberculosis* H37Rv, it stands to reason that larger interface in the later may have implication in binding to other cellular partner.

Keywords: *Mycobacterium tuberculosis* H37Rv, Rv1915, isocitrate lyase 2, inclusion bodies, solubilization and IDPRs

INTRODUCTION

Soluble expression of potential drug targets in heterologous host is the limiting factor for their production in amounts required for their structure function characterization, screening of potential inhibitors and for unraveling the mechanism of inhibition. Although *Escherichia coli* is the most popular choice of host for production of recombinant proteins, however, low or no protein expression, incorrect folding or inclusion body formation (IBs), protein inactivity are some common problems during expression in this workhorse organism. Some of the factors responsible for these difficulties are high rates of transcription and translation processes, codon bias, absence of posttranscriptional modification in *E. coli*, unfavorable environment of bacterial cytoplasm for the formation of disulphide-bonds resulting in misfolding/unfolding of the proteins and ultimately leading to protein instability, aggregation and accumulation of the recombinant protein as IBs (Vincentelli et al., 2003; Zhang et al., 2004; Choi et al., 2006; Rosano and Ceccarelli, 2014).

IBs are the pool of partially folded or misfolded proteins which are biologically inactive. IBs were well-characterized in terms of their secondary structure and morphology, indicating that they possess a native-like secondary structure which may have biological activity (Bowden et al., 1991; Oberg et al., 1994; Przybycien et al., 1994). IBs accounts for the 25% of total cellular protein and are enriched in the recombinant protein as opposed to other protein of *E. coli*. In fact, if functionally active protein can be recovered from IBs, then their formation is advantageous as it provides a method for isolation of highly purified protein by (i) isolating IBs from the bacterial cytoplasm, (ii) solubilizing them by using denaturing agents such as urea and guanidine hydrochloride, followed by (iii) refolding via removal of denaturing agents to recover bioactive protein (Rudolph and Lilie, 1996; Vallejo and Rinas, 2004). Of these, solubility and refolding are the two critical steps which affect the time and cost of protein recovery, and thus determine the overall yield of active protein (Rudolph and Lilie, 1996; Burgess, 2009). Generally, use of high concentration of denaturing agents such as urea and guanidine hydrochloride in presence of reducing agent is the most commonly process for solubilization of IBs. However, the high concentration of detergents disrupts the complete secondary structures of the protein which may lead to the aggregation of the protein during refolding process. This problem is overcome by using mild solubilizing agents such as lower concentration of detergents, alcohols, DMSO, high pH, reducing agents (Khan et al., 1998; Process for solubilization of recombinant proteins expressed as inclusion body, 2003; Mohan Singh and Kumar Panda, 2005).

Mycobacterial infections, are the major concern for public health due to the emergence of drug resistant strains of the pathogen. During its persistence phase *Mtb* resides inside the granulomas which are rich in even and odd chain fatty acids. Activation of glyoxylate pathway allows the pathogen to utilize acetate or propionate (degradation product of fatty acids) as carbon sources for its growth (Bloom, 1994; McKinney et al., 2000). The two important enzymes of this pathway are Isocitrate lyase (ICL) and Malate synthase (MS), the former encoded by 2 genes (smaller *icl1* and larger *icl2*/*aceA*) and the later by *aceB,* respectively. Here it may be helpful to point out for readers that the term "*ace*" came up because the genes that encode for MS, ICL and isocitrate dehydrogenase kinase/phosphatase form an operon aceBAK in *E. coli* (Chung et al., 1988) that functions in acetate utilization. However, operonic arrangement of these genes is not true in all organisms and therefore annotating such genes as "*ace*" is a misnomer and confusing. Specially, in case of *Mtb* H37Rv, the two *icls* that together play an important role in pathogenesis and persistence of the bacterium, are annotated as *icl1* and *aceA* in literature (Cole et al., 1998; Höner Zu Bentrup et al., 1999; Muñoz-Elías and McKinney, 2005). Due to presence of a stop codon in between, the larger *aceA* (766 residues in *Mtb* strain CDC 1551) is split into *aceAa/rv1915* (367 residues) and *aceAb/rv1916* (398 residues) in *Mtb* H37Rv strain (**Figure 1**). The authors suggest that these split genes be termed as *icl2a* (*rv1915*) and *icl2b* (*rv1916*) for clarity and consistency and the same has been followed in the current study. In case of H37Rv, as evident from sequence mapping of full length and spilt versions of ICL2, ∼90 residues involved in the formation of domain II are present in Rv1915/ICL2a, whereas the rest of the 59 residues of domain II are present in Rv1916/ICL2b (**Figure 1**). Literature documents *Mtb* ICLs to be novel antitubercular drug targets (Wang et al., 2011). The crystal structure of ICL1 (Rv0467), determined in 2,000 by Sharma et al. (2000), was a momentous discovery for structure-based drug designing against *Mtb*. However, although some inhibitors have been reported against *Mtb* Rv0467/ICL1, but no drug is available till date to treat persistent *Mtb*. The possible reasons for this failure are the undiscovered roles of split ICL2 (Rv1915 and Rv1916) that may assist the pathogen to survive in granulomas. In a recent study we reported structure function insights into Rv1916/ICL2b (Antil et al., 2019), but Rv1915 is yet to be characterized. This study reports the difficulties encountered in obtaining soluble expression of the target protein in heterologous host *E. coli* BL21 (DE3) and mainly focuses on strategies adopted for recovery of active Rv1915/ICL2a.

MATERIALS AND METHODS

Chemicals

All chemicals used in this study are commercially available except 2-Methylisocitrate which was synthesized in the laboratory of Dr. Sébastien Gouin, University of Nantes France. The genomic DNA of *Mycobacterium tuberculosis* H37Rv, Luria-Bertani (LB) medium for bacterial growth and Isopropyl β-D-1-thiogalactopyranoside (IPTG) were purchased from Hi-Media Laboratories, India. Primers used for gene amplification were synthesized through Eurofins Genomics India Pvt. Ltd. Restriction enzymes (NheI and HindIII), Alkaline Phosphtase and T4 DNA Ligase were procured from Fermentas, US. DL-Isocitric acid and Phenylhydrazine were obtained from Sigma (India).

Cloning of *rv1915*

The gene coding for *Mtb* H37Rv Rv1915/ICL2a was amplified by Polymerase chain reaction using a pair of gene specific primers

listed in **Table 1** from genomic DNA of *Mtb* H37Rv DNA. For directional cloning of the insert DNA, recognition sites for NheI and HindIII restriction enzymes were designed into the 5' end of the forward and reverse primers, respectively (underlined in the **Table 1**). In addition, to facilitate the target purification, His_6-tag was incorporated either in the forward primer for N-terminus tag or in the reverse primer for C-terminus tag (highlighted in bold letters in the **Table 1**). The PCR reaction mix comprised of 1x Taq buffer, 10 ng/μl of genomic DNA, 20 pmoles of each forward and reverse primers, 200 μM dNTPs mix and 4:1 ratio of Taq (Geno biosciences):Pfu polymerase (Fermentas). The standardized PCR cycle for all the three constructs was: initial denaturation at 95°C for 5 min, denaturation at 95°C for 1 min, annealing at 65°C for 1 min and extension at 72°C for 1 min. These conditions were repeated for 30 cycles before a final extension at 72°C for 10 min. The amplified insert and pET-21c expression vector (Novagen) were digested with NheI and HindIII restriction enzymes at 37°C for 1–2 h. The vector and insert were then ligated using T4 DNA ligase followed by the transformation of *E. coli* DH5α with the ligated product. The cells were plated on LB agar plate containing 100 μg/ml ampicillin. Random colonies were picked from the plate and inoculated in 3 ml LB broth supplemented with ampicillin (100 μg/ml). Plasmid was isolated from each of the colonies by alkaline lysis method (Sambrook and Russell, 2006) followed by double digestion with NheI and HindIII restriction enzymes. All the digested products were analyzed by running on 1% agarose gel.

Expression and Cellular Localization of Rv1915/ICL2a

For expression studies the *E. coli* BL21 (DE3) competent cells were transformed with recombinant plasmid. The transformed cells were plated onto LB agar plate containing 100 μg/ml ampicillin and incubated at 37°C overnight. A single colony was inoculated in 5 ml of LB broth supplemented with 100 μg/ml ampicillin and incubated overnight at 37°C with continuous shaking at 200 rpm. 50 μl of this primary culture was then transferred into 5 ml of LB broth containing 100 μg/ml ampicillin and incubated with shaking at 37°C till the OD of culture at 600 nm was 0.5–0.6. The expression of recombinant protein

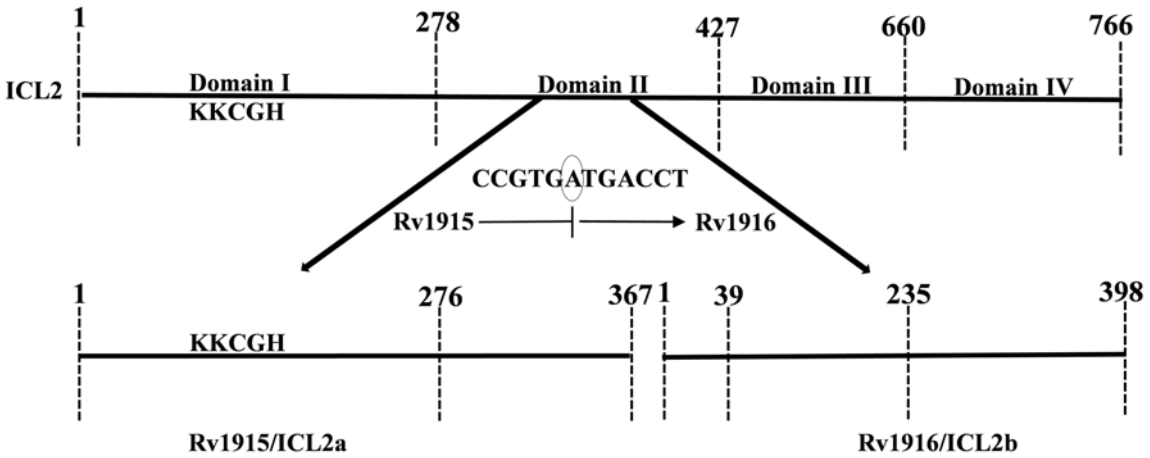


FIGURE 1 | Mapping of full length (*Mtb* strain CDC 1551) and split (*Mtb* strain H37Rv) ICL2. The unique eukaryotic domain II (278–427) in ICL2 of Erdman and CDC1551 strains is divided into two overlapping ORFs namely, Rv1915 (278–367) and Rv1916 (1–59 residues) in H37Rv strain (shown by bold arrows), a consequence of translational coupling phenomenon. Presence of circled nucleotide "A" in the reading frame, results into translation of TGA into stop codon and termination of upstream *rv1915*. The ribosome that has just finished is in a position to initiate translation with ATG start codon for the downstream *rv1916.* The ICL signature motif (KKCGH) present in domain I of both are also shown in the figure.

TABLE 1 | List of primers used for preparing different constructs of Rv1915/ICL2a.

S. No.	Name (details) of the construct	Primer	Primer sequence (5′ to 3′)
1	His_6-Rv1915 (FL Rv1915 with His_6 tag at N-terminal)	FP	5′-gatttagctagc**catcaccatcaccatcac**gccatcgccgaaacggacaccg-3′
		RP	5′-gatttaaagctttcaggcccgcgtcgtcctc-3′
2	Rv1915-His_6 (FL Rvk with His_6 tag at C-terminal)	FP	5′-gatttagctagcgccatcgccgaaacggacaccg-3′
		RP	5′-gatttaaagctttcaggcccgcgtcgtcctccgcgccgagaaggaacggctg-3′
3	Rv1915Δ35CT- His_6 (Rv1915 with 35 residues deleted from C-terminal)	FP	5′-gatttagctagcgccatcgccgaaacggacaccg-3′
		RP	5′-gatttaaagctttcaagccgaaaatgccttggcgctccgcgccgagaaggaacgg-3′
4	Rv1915Δ90CT- His_6 (Rv1915 with 90 residues deleted from C-terminal)	FP	5′-gatttagctagccatcaccatcaccatcacgccatcgccgaaacggacaccg- 3′
		RP	5′-gatttaaagctttcagtgatggtgatggtgatgcgcgccgagaaggaacggctg-3′

FL, Full length; FP, Forward Primer; RP, Reverse Primer.

was induced with IPTG concentration, temperature and time as indicated at relevant positions. The culture was harvested by centrifugation at 5,000 rpm for 10 min at 4°C. For analysis of protein induction, cell pellets were directly resuspended in 50 μl of 5X SDS loading dye containing 50 mM Tris-HCl pH 8, 0.25% β-mercaptoethanol (βME), 1% SDS, 10% glycerol and 0.04% Bromophenol blue. On the other hand, for determining localization of the expressed protein, cell pellet of 1 ml culture was lysed by mixing the pellet with 200 μl of 50 mM Sodium Phosphate buffer pH 8, comprising 300 mM NaCl and 20 mg/ml lysozyme, followed by incubation on ice for 15 min and sonication (pulse: 5 s ON and 5 s OFF at 40% amplitude) using ultrasonic water bath (Citizen). After lysis, centrifugation at 12,000 rpm for 15 min segregated the soluble (supernatant) and insoluble fractions (pellet) of the total cell lysate. All samples were boiled at 100°C for 10 min in 1X SDS loading dye before subjecting to 10% SDS-PAGE for expression analysis. For visualization of proteins, gels were stained in 0.25% Coomassie Brilliant Blue R-250 and then destained in 30% (v/v) methanol in water with 10% (v/v) acetic acid solution.

IBs Isolation and Solubilization of Rv1915/ICL2a

The induced *E. coli* cell culture (1L) was harvested by centrifugation at 5,000 rpm for 10 min at 4°C. The cell pellet was resuspended in 40 ml of 50 mM Tris buffer pH 8.5 containing 5 mM EDTA and 1 mM PMSF. The cells were lysed by sonication on ice for a total time of 20 min (1-min burst and 1-min OFF time) and centrifuged at 15,000 rpm for 30 min at 4°C. The pellet thus obtained was washed with wash buffers A (50 mM Tris pH 8.5, 5 mM EDTA, 1 mM PMSF and 2.5% Triton X-100) and B (50 mM Tris pH 8.5) for three repeated cycles of sonication and centrifugation in each buffer. Finally, the pellet of IBs was dissolved in 2 ml of Milli-Q water and processed for solubilization. In order to obtain bioactive Rv1915, six different solubilization buffers (**Table 2**) containing mild solubilizing agents were used. For solubilization of IBs, the 2 ml of purified IBs were equally divided (330 μl) in microcentrifuge tube and diluted to a final volume of 1 ml by adding 670 μl of solubilization buffers. The suspensions were vortexed and incubated at room temperature on an end-to-end rotator for an hour. The solubilized samples were separated from insoluble fraction by centrifugation at 15,000 rpm for 30 min at 4°C and the samples were analyzed on 10% SDS-PAGE.

TABLE 2 | List of Solubilization Buffers.

Buffer code	Buffer composition	Concentration of Rv1915 (mg/ml)
SB1	50 mM Tris, 5 mM EDTA, 1 mM PMSF, 20 mM βME, 0.25 M Urea, 0.5% Sarcosine pH 8	2.2
SB2	50 mM Tris, 1 mM PMSF, 20 mM βME, 0.5% Sarcosine pH 8	2
SB3	50 mM Tris, 1 mM EDTA, 1 mM PMSF, 10 mM βME, 0.5 M Urea, 0.25% Sarcosine pH 8	1.85
SB4	50 mM Tris, 1 mM EDTA, 1 mM PMSF, 20 mM βME, 0.25 M Urea, 0.5% Sarcosine pH 8	1.99
SB5	50 mM Tris, 5 mM EDTA, 1 mM PMSF, 20 mM βME, 0.5 M Urea, 0.5% Sarcosine pH 8	2.5
SB6	50 mM Tris, 1 mM EDTA, 1 mM PMSF, 5 mM βME, 0.5 M Urea, 0.5% Sarcosine pH 8	1.85

SB, Solubilization buffer; βME, β-mercaptoethanol.

Protein Quantification Using ImageJ Software

ImageJ is a freely available software (https://imagej.nih.gov/ij/download.html), used to determine the protein concentration from SDS-PAGE gels. This software measures the relative density of each protein band from a selected lane of the gel and plot a graph according to their densities. In order to determine the protein concentration, peak area of the band of interest was calculated and compare protein band with the known concentration. To estimate the protein concentration, known amount of BSA (2–10 μg/μl) was run on 10% SDS-PAGE and the relative density of each band was calculated using ImageJ software. BSA standard curve was then prepared using calculated peak area from the software and plotted against the known concentration of BSA. The standard curve thus prepared was used for determining the concentration of solubilized protein from each buffer.

Ni-NTA Purification of Rv1915/ICL2a

Standard protocol of Ni-NTA affinity chromatography was used for recombinant protein purification. Induced cell pellet of 100 ml was dissolved in 20 ml of lysis buffer containing 50 mM Sodium Phosphate buffer pH 8, 10 mM Imidazole, 300 mM NaCl, 0.5% Sarcosine, 2 mM βME, 1 mM PMSF and 20 mg/ml lysozyme. The buffer optimized from IBs solubilization experiments was further modified according to the standard buffer composition for Ni-NTA affinity chromatography. For lysis, the cell suspension was subjected to sonication for a total time of 10 min which consisted of 10 s ON and 10 s OFF cycles. After sonication, the cell lysate was centrifuged at 12,000 rpm for 30 min to remove the cell debris and the clear lysate was loaded on to 0.5 ml of Ni-NTA column pre-equilibrated with equilibration buffer (50 mM Sodium Phosphate buffer pH 8, 300 mM NaCl and 10 mM Imidazole). After the binding period of an hour, the column was washed with washing buffer A (50 mM Sodium Phosphate buffer pH 8, 300 mM NaCl, 20 mM Imidazole, 2 mM βME and 1 mM PMSF) and B (50 mM Sodium Phosphate buffer pH 8, 300 mM NaCl, 40 mM Imidazole, 2 mM βME and 1 mM PMSF). The bound protein was eluted from the resin with 250 mM Imidazole buffer, quantified by Bradford protein assay, dialyzed against the storage buffer (20 mM Tris pH 8, 100 mM NaCl, 2 mM βME, 1 mM PMSF and 5% glycerol) and aliquots stored at −80°C.

ICL and MICL Activity Assays

ICL activity of Rv1915/ICL2a was determined by a coupled assay that monitored the formation of glyoxylate-phenylhydrazone complex at 324 nm, generated because the glyoxylate produced

in the reaction was further made to react with phenylhydrazine-HCl. In brief, 1 ml of reaction mixture included 50 mM MOPS buffer (pH 7), 1 mM DL-isocitrate trisodium salt, 6 mM $MgCl_2$, and 4 mM phenylhydrazine-HCl and 10 μg protein either from solubilized IBs or total cell lysate. Reaction mixture containing solubilization buffers/lysis buffer without enzyme was used as blank. As the protein sample was not pure, total cell lysate of uninduced sample was also assayed for ICL activity as a negative control. For accuracy in comparative analysis, kinetic experiments were carried out under similar experimental conditions as described for Rv1916/ICL2b (Antil et al., 2019). *Mtb* ICLs, also reported to catalyze 2-methylisocitrate and convert it into pyruvate and succinate (Gould et al., 2006). Therefore, 10 and 15 μg of purified enzyme was used for assaying isocitrate and methylisocitrate activity, respectively.

Bioinformatics Analysis

Sequences of ICL2s from different strains of *Mtb* were aligned using EMBL-EBI tool- Clustal Omega (https://www.ebi.ac.uk/Tools/msa/clustalo/). All the sequences were retrieved from KEGG Database (https://www.genome.jp/kegg/). Expasy tools (https://www.expasy.org/proteomics/protein_structure) were used to predict the secondary structures (alpha, beta random coils and turns) of Rv1915/ICL2a. The disordered regions of Rv1915/ICL2a were further verified by different online servers namely- Prediction of Amyloid Structure Aggregation 2.0 (PASTA 2.0 -http://protein.bio.unipd.it/pasta2/) server, Predictor of Natural Disordered Regions (PONDER - http://www.pondr.com/) and Protein Disorder Prediction Server (PrDOS - http://prdos.hgc.jp/cgi-bin/top.cgi). These servers are freely available and predict the disordered regions of a given protein using its amino acid sequence. PASTA 2.0 predicts the formation of amyloids due to self-aggregation of the given protein using its energy function (Walsh et al., 2014). PONDER uses composition, complexity and hydropathy index of amino acid sequence of a protein to find the disordered regions (Peng et al., 2005). Similarly, PrDOS calculate probability of every amino acid in a protein of being unstructured/disordered (Ishida and Kinoshita, 2007). The quaternary model structure of Rv1915/ICL2a was generated by GalaxyWeb online server (http://galaxy.seoklab.org/cgi-bin/submit.cgi?type=HOMOMER) (Ko et al., 2012)

RESULTS

Expression and Localization of Rv1915/ICL2a as IBs

The successful cloning of *His_6-rv1915* in pET-21c was confirmed by double digestion with restriction enzymes NheI and HindIII. The fall out of 1.1 kb confirms the presence of insert *rv1915/icl2a* in pET-21c vector (**Figure 2A**). The expression of recombinant His_6-Rv1915, induced with 1 mM IPTG for 16–18 h at 18°C was analyzed on 10% SDS-PAGE. **Figure 2B**, confirms the expression of Rv1915/ICL2a at their expected size i.e., ∼40.5 kDa. Unfortunately, accumulation of induced protein in the insoluble pellet/IBs (**Figure 2B**, lane 4) leave negligible or no Rv1915 protein in the soluble fraction of the lysate (**Figure 2B**, lane 5). Despite extensive efforts involving variation in the induction temperature and IPTG concentration, media optimization, addition of osmolytes/chaotropes/additives in the culture media during cell growth etc., soluble expression of the induced protein could not be achieved (**Figures S1**–**S3**).

Recovery of Bioactive Rv1915/ICL2a From IBs

IBs of Rv1915/ICL2a were isolated as described in section IBs Isolation and Solubilization of Rv1915/ICL2a and six different solubilization buffers (SBs) varying in concentrations of urea, EDTA, βME and sarcosine were designed for solubilizing inclusion body protein Rv1915. **Table 2** represents the composition of SBs and concentration of solubilized His_6-Rv1915 in respective buffers. SDS-PAGE analysis of IBs solubilization using different buffers is depicted in **Figure 3A**. Almost all the buffers were able to solubilize the IBs of protein of interest to some extent, with highest concentration (2.5 mg/ml) of His_6-Rv1915 achieved in buffer SB5, composed of 50 mM Tris pH 8, 5 mM EDTA, 1 mM PMSF, 20 mM βME, 0.5 M Urea and 0.5% Sarcosine (**Table 2**). In order to select the appropriate buffer for the recovery of bioactive His_6-Rv1915, activity assay was performed with soluble fraction of His_6-Rv1915 from each buffer. As lowest activity of His_6-Rv1915 was observed in SB5 buffer, it was deemed unsuitable (**Figure 3B**). The highest activity was achieved in buffer SB2 where the solubilizing additive was only 0.5% sarcosine without urea or EDTA. EDTA appears to be more detrimental for the activity of Rv1915 than urea, as reducing the concentration of ETDA to 1 mM in SB4 (but keeping urea same) increases ICL activity almost comparable to the SB2. Furthermore, SB3 and SB6 shows substantial reduction in amount and activity of soluble protein due to the decrease in concentration of sarcosine and βME, respectively (**Figure 3**A). From all these observation it was concluded that sarcosine and βME plays an important role in solubility and activity of His_6-Rv1915, so these standardized conditions were employed in purification of His_6-Rv1915. Unfortunately, even after the solubilization of His_6-Rv1915 with 0.5% sarcosine, Ni-NTA affinity purification of the protein could not be achieved due to inefficient binding of the protein to the Ni-NTA beads (**Figure 3C**, lanes 2 & 3). The possible reasons could be inaccessibility/masking of the His_6-tag due to formation of soluble aggregates of the protein or loss of His-tag due to proteolysis. Alternatively, if proteins possess signal peptide and transmembrane domain at their N-terminus which when liberated will result in loss of tag and therefore reduced binding to Ni^{2+} matrix. The presence of these signal peptides at N-terminus was checked with the online servers - SignalP.4 (http://www.cbs.dtu.dk/services/SignalP-4.0/) (Petersen et al., 2011) and TMHMM (http://www.cbs.dtu.dk/services/TMHMM/) (Sonnhammer et al., 1998; Krogh et al., 2001) and negated the possibility. To overcome the problem of the His_6-tag being masked or degraded at the N-terminus, recombinant Rv1915-His_6 was prepared where the His tag was placed at C-terminal of the protein (**Figure S4**), but the problem

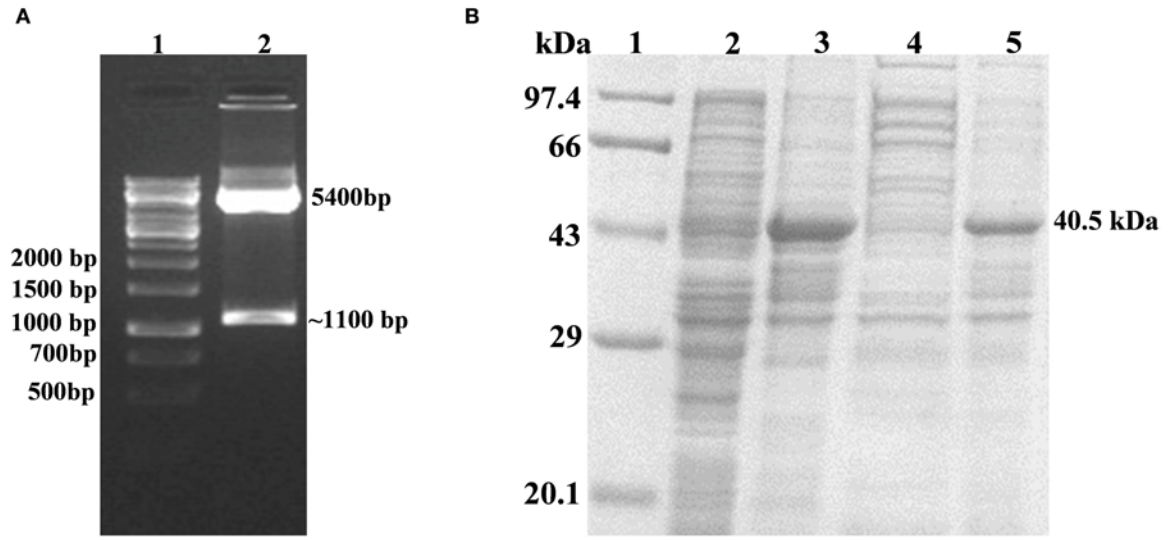


FIGURE 2 | Cloning, Expression and Localization of Rv1915/ICL2a with N-terminal His_6 tag: **(A)** Confirmation of cloning of Rv1915/ICL2a in pET-21c by double digestion with restriction enzymes NheI and HindIII: Lane 1- 1kb DNA ladder; Lane 2- positive clone of Rv1915 **(B)** Expression and localization of Rv1915: Lane 1- Medium range protein marker; Lane 2- Total cell lysate of uninduced sample; Lane 3- Total cell lysate of induced sample; Lane 4- Soluble fraction; Lane 5- Insoluble fraction.

still persisted (**Figure 3C**, lanes 8 & 9). Finally, bioinformatics analysis was performed that provided some clue for resolving the problem.

Sequence and Structure Analysis of Rv1915/ICL2a

Multiple sequence alignment of Rv1915/ICL2a with ICL2s from other *Mtb* strains reveal variability mostly in the C-terminal residues (**Figure 4A**). This difference was somewhat anticipated, as compared to larger ICL2s (∼766 amino acids), H37Rv ICL2 is split in two ORFs where Rv1915 forms the first part and Rv1916 the later. Secondary structure prediction based on the primary sequence of Rv1915 estimated ∼41.96% disorder (**Figure S5**), where out of 72 C-terminal residues 40 of them are random coil (highlighted in the black box). In the united version of ICL2, the equivalent region is comprised of helices, therefore, the splitting of this helical region is increasing disorder at the C-terminal of Rv1915. Further analysis with the PASTA 2.0 server corroborated that the 314–367 residues of Rv1915/ICL2a are disordered and have the propensity for self-aggregation and amyloids formation (**Table 3**). The software also predicts two additional segments (248–251 and 304–307) with tendency toward parallel aggregation. Two other servers PrDOS and PONDER endorsed the presence of disordered regions at the termini of Rv1915/ICL2a (**Figure S6**). Specifically, ∼35 residues from the C-terminus and ∼15 residues from the N-terminus of the queried protein was predicted to be unstructured by all the three servers. *In silico* deletion of either of these in PASTA 2.0 server did not reduce the number of amyloids, which could be achieved only after truncation of ∼90 residues (278–367) from the C-terminal end of Rv1915 (**Table 3**). This 90 residue long C-terminal region encompasses the second aggregation segment (residues 304–307), whose removal appears to reduce amyloid formation. As discernible from multiple sequence alignment (**Figure 4A**), this section has low similarity with the larger ICL2, reflected in the structural differences as well. The comparison of equivalent/similar structural region of *Mtb* ICL2 (**Figure 4B**) and the model structure of Rv1915/ICL2a also illustrates the disordered nature of C-terminal 90 residues (colored pink in **Figure 4C**). Therefore, in order to minimize the probability of the non-specific aggregation of the expressed target, two deletion variants of *rv1915/icl2a* were designed with 35 and 90 residues truncated from the C-terminus.

Effect of C-Terminal Truncation on the Solubility and Activity of Rv1915/ICL2a

Two C-terminal truncated variants of *rv1915*, namely, Rv1915Δ35CT-His_6 and Rv1915Δ90CT-His_6, were cloned in pET-21c vector using methodology detailed in section Cloning of *rv1915*. Recombinant clones were confirmed by double digestion with the same restriction enzymes (**Figure S7**). The deletion variants were expressed in *E. coli* BL21 (DE3), induced with 1 mM IPTG (**Figures 5A,C**) and purified with Ni-NTA affinity chromatography (**Figures 5B,D**). Out of the two variants, purification could be achieved only for Rv1915Δ90CT-His_6 due to inept binding of Rv1915Δ35CT-His_6 to Ni-NTA resin (**Figures 5A,B**), similar to the problem encountered in the case of full length Rv1915-His_6 (**Figure 3C**). **Figure 5D** shows the profile of Ni-NTA eluted fractions of Rv1915Δ90CT on 10% SDS-PAGE. The total yield corresponded to ∼20 mg/l. In order to compare the effect of C-terminal truncation on the functionality of Rv1915/ICL2a, and since purified full length Rv1915 could not be achieved, ICL activity was carried out

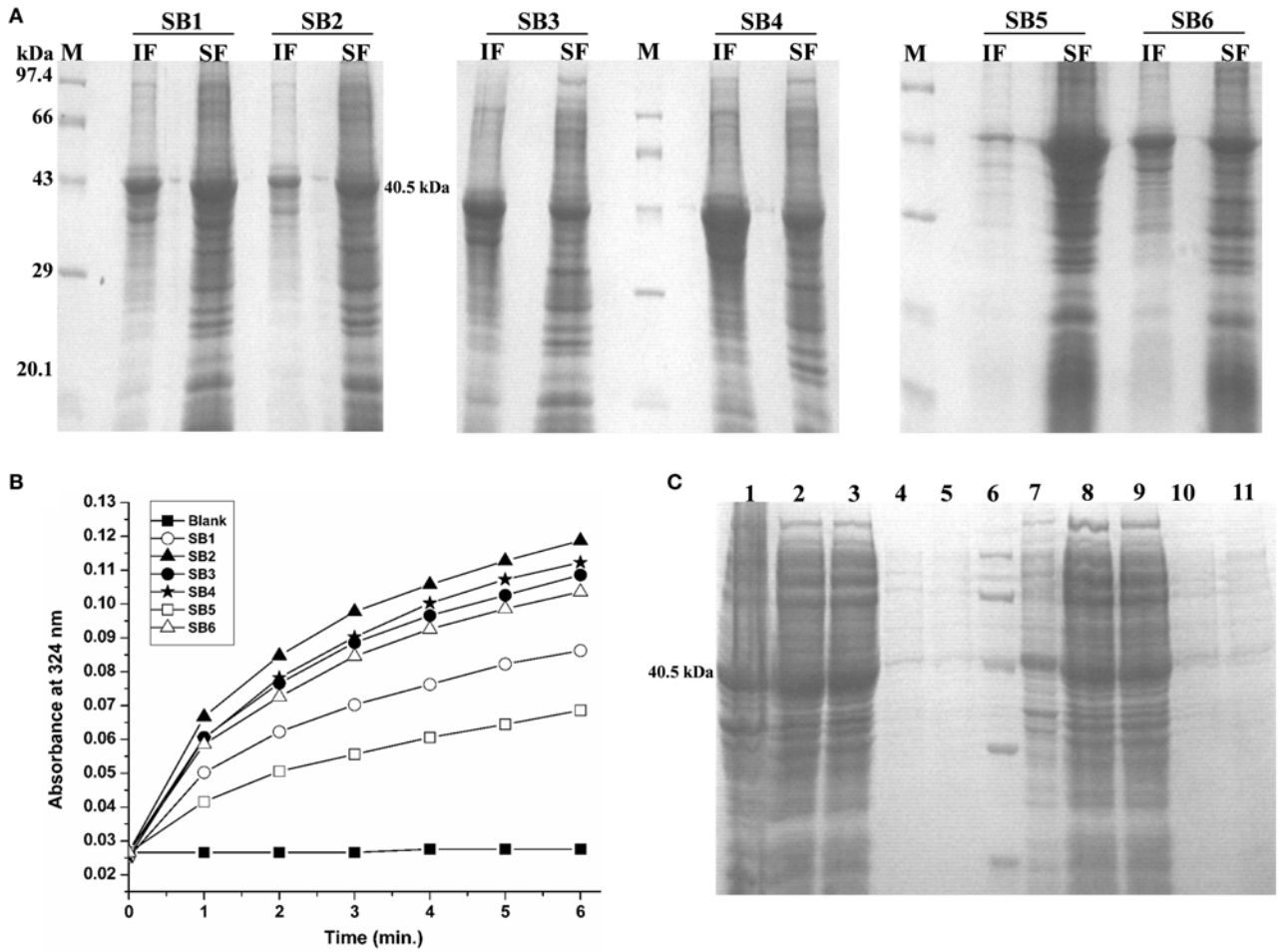


FIGURE 3 | Standardization of solubilization buffer and purification of Rv1915/ICL2a **(A)** SDS-PAGE analysis for solubilization of His_6-Rv1915 IBs in different SBs: Lane M: Medium range protein marker; Lane IF: Insoluble fraction; Lane SF: Soluble fraction. Composition of all SBs are listed in **Table 2**; **(B)** Activity assay of Rv1915/ICL2a solubilized from IBs: Increased absorbance at 324 nm with the addition of Rv1915 (extracted from different SBs) indicate formation of glyoxylate phenylhydrazone complex and hence ICL activity; **(C)** Purification of full length His_6-Rv1915 at N-terminal using SB2 buffer (Lane 1- Insoluble fraction; Lane 2- Soluble fraction; Lane 3- Flow through; Lane 4 and 5- Eluted fractions; Lane 6- Medium range protein marker) and Rv1915-His_6 at C-terminal (Lane 7- Insoluble fraction; Lane 8- Soluble fraction; Lane 9- Flow through; Lane 10 and 11- Eluted fractions).

with crude lysates (**Figure S8**). The observations show that the C-terminal truncation improves the activity of Rv1915/ICL2a. Kinetic parameters of Rv1915Δ90CT were determined using Lineweaver-Burk plot for both the substrates; isocitrate and 2-methylisocitrate (**Figure 6**). The kinetic parameters reveal that Rv1915Δ90CT has ∼50 fold higher affinity for isocitrate (5.2 μM) than 2-methylisocitrate (279 μM). Calculation of catalytic efficiency turns out to be 0.83 μM^{-1} min^{-1} for isocitrate and 0.0137 μM^{-1} min^{-1} for 2-methylisocitrate, reinforcing faster turnover of isocitrate into glyoxylate and succinate (as compared to conversion of 2-methylisocitrate to pyruvate and succinate). Similar trend was observed for Rv1916/ICL2b (**Table 4**). However, Rv0467/ICL1 unequivocally displays much higher activity for both the substrates in comparison to Rv1915/ICL2a and Rv1916/ICL2b (Antil et al., 2019).

DISCUSSION

This work was initiated with an aim to produce Rv1915/ICL2a, an important drug target of *Mtb* H37Rv, in ample amounts for structure function studies. Therefore, Rv1915/ICL2a was cloned in pET-21c vector and expressed in *E. coli* BL21 (DE3) strain but the recombinant protein localized in the insoluble fraction of cell lysate. Protein misfolding/unfolding/formation of insoluble IBs is often a problem during the overexpression of recombinant proteins. Employing different strategies such as reducing IPTG concentration, expression temperature, choice of right expression vector and host strain, optimizing composition of the culture media, co-expression with molecular chaperones, purification from inclusion bodies, etc. may help in overcoming the problem of insoluble expression in *E. coli*. Therefore, different induction temperature, inducer

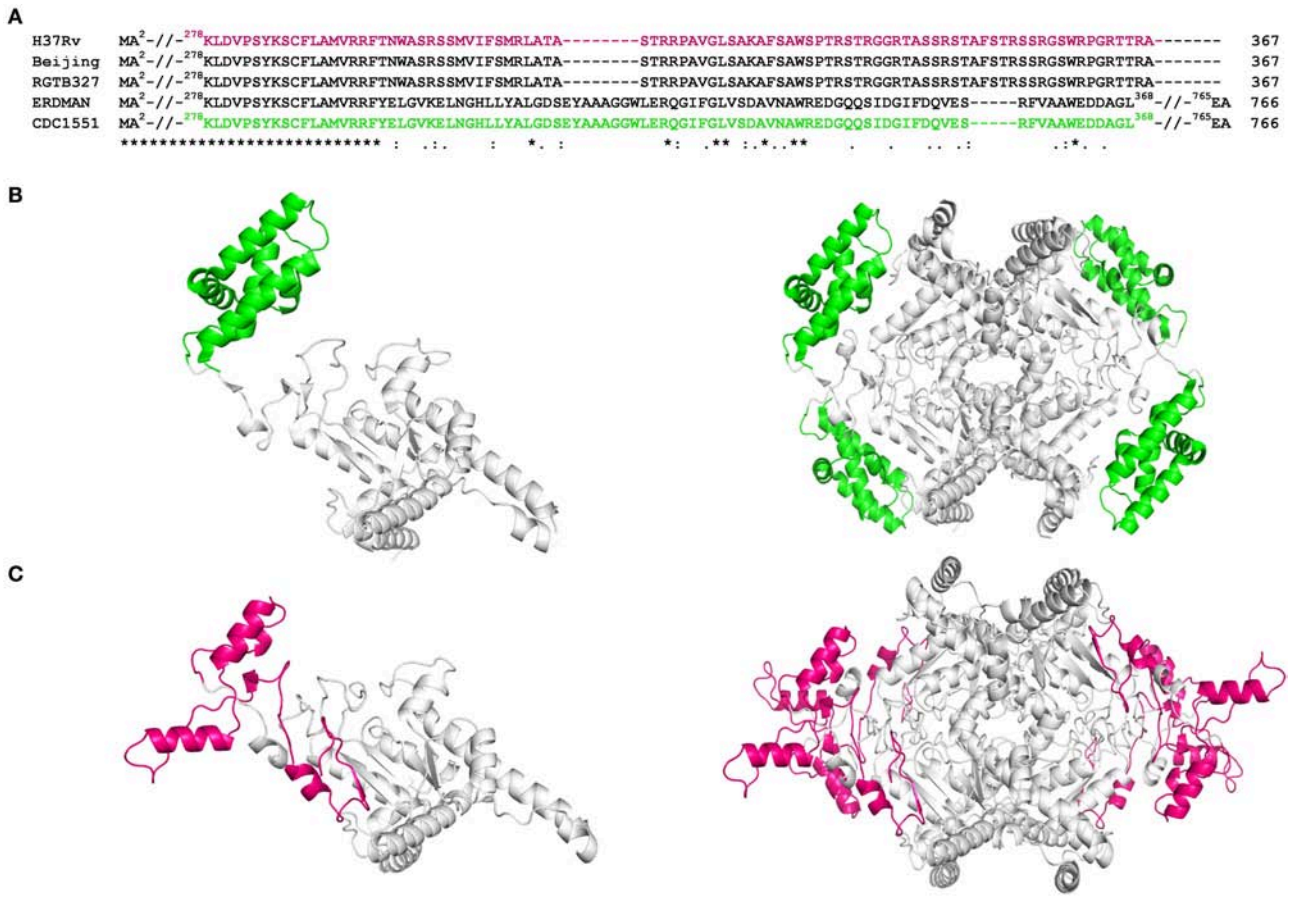


FIGURE 4 | Sequence and structure analysis of Rv1915/ICL2a: **(A)** Multiple sequence alignment of Rv1915/ICL2a and ICL2 from different strains of *Mtb* represents the variation at the C-terminal (residues 278–367) of spilt versions of ICL2s (H37Rv, RGTB327 and Beijing) vs. larger ICL2s (Erdman and CDC1551) (residues 278-368); **(B)** Crystal structure of larger ICL2 from *Mtb* CDC1551 (extracted for residues 1–368 from PDB code 6EDW) illustrates monomer (left) and tetramer (right) highlighting the variable residues (278–368) that form part of domain II in green color; **(C)** Model structure of monomeric (left) and tetrameric (right) Rv1915/ICL2a generated by GalaxyWeb online server shows the disordered nature and different orientation of equivalent C-terminal tail in magenta color. Complete sequence identity between the ICL2s in the region before the variable residues is reinforced by undistinguishable gray color cartoon structures in both the homologs.

TABLE 3 | Prediction of IDPRs of Rv1915/ICL2a using PASTA 2.0 server.

S. no.	Different variant Rv1915/ICL2a	Length of Rv1915/ICL2a	No. of amyloids	Best energy
1.	Full length Rv1915	367	2	−5.448
2.	Rv1915 with of 15-residues truncated from N-terminus	352	2	−5.448
3.	Rv1915 with of 35-residues truncated from C-terminus	332	2	−5.448
4.	Rv1915 with of 90-residues truncated from C-terminus	277	1	−5.448

concentration, growth media were explored but none of these increased the solubility of the recombinant protein (**Figures S1**–**S3**). Attempt to isolate Rv1915/ICL2a IBs followed by solubilization in mild buffers did yield active protein but could not be purified further as it did not bind to Ni-NTA (**Figure 3C**). The possible reasons could be masking or degradation of the His_6-tag. In any case all efforts to obtain soluble Rv1915 in amounts enough for further studies reached a dead end.

Multiple sequence alignment of *Mtb* ICL2a with 766 amino acid long ICL2s show high variability in the region of domain II where the later divided into two ORFs (**Figure 4A**). It appears that gene duplication in domain II may be responsible for structural divergence and evolution of this split version of ICL2. Fortunately the crystal structure of ICL2 from *Mtb* strain CDC 1551/Oshkosh (PDB code 6EDW), has recently become available (**Figure 4B**) (Bhusal et al., 2019) and helped in building homology model of Rv1915/ICL2a (**Figure 4C**). Both

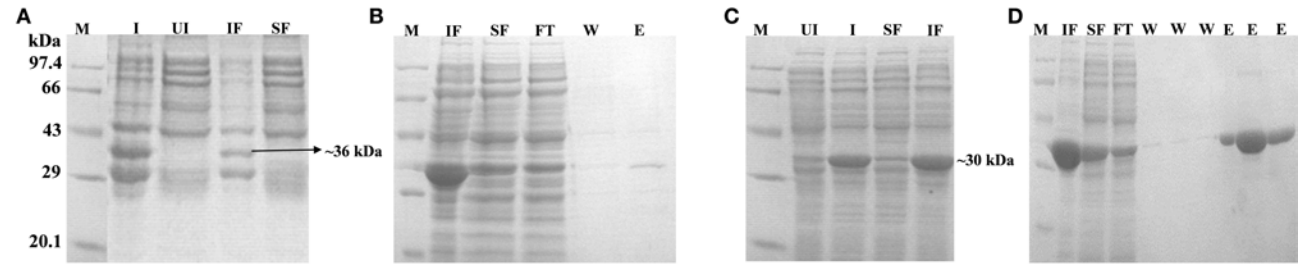


FIGURE 5 | Expression, localization and purification of Rv1915Δ35CT **(A,B)** and Rv1915Δ90CT **(C,D)**: Lane M, Medium range protein marker; Lane I, Total cell lysate of induced sample; Lane UI, Total cell lysate of uninduced sample; Lane IF, Insoluble fraction; Lane SF, Soluble Fraction; Lane FT, Flow through after binding to the Ni-NTA resin; Lane W, Washes with different concentrations of imidazole; and Lane E, Elutions with 250 mM imidazole.

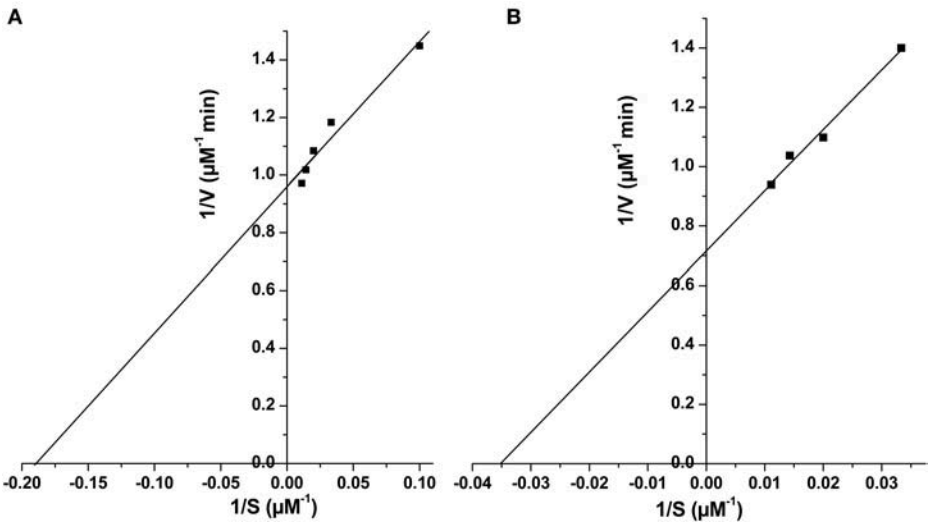


FIGURE 6 | Determination of kinetic parameters of Rv1915Δ90CT using Lineweaver Burk plot for both the substrate, isocitrate **(A)** and methylisocitrate **(B)**.

secondary structure and 3D model predicted disordered C-terminus Rv1915/ICL2a (residues 278–367) as opposed helical region of long ICL2. For reducing non-specific aggregation due to presence of floppy tails, recombinant DNA technology was employed for generating two variants of Rv1915 where 35 (Rv1915Δ35CT) and 90 (Rv1915Δ90CT) residues from the C-terminal were deleted. Only the later could be purified successfully that exhibited dual ICL and MICL activities, observed first time for *Mtb* H37Rv strain. Dual activity has been reported for the complete *Mtb* strain CDC 1551/Oshkosh ICL2 (Bhusal et al., 2019), but no activity data exists for Rv1915/ICL2a till date. Functional characterization of Rv1915/ICL2a follows our previous study on Rv1916/ICL2b (Antil et al., 2019) and although both Rv1915/ICL2a and Rv1916/ICL2b display dual activities, ICL and MICL activities of both the proteins are much lower than that exhibited by Rv0467/ICL1 (**Table 4**). Nevertheless, all three *Mtb* ICLs show preference for isocitrate over methylisocitrate.

It may be worth pondering on the biological significance of presence of IDPR in otherwise structured Rv1915. The propensity of these IDPRs to bind to multiple biological partners and their role in cellular activities such as gene replication, transcription, regulation and signal transduction is becoming evident (Uversky, 2013). Under physiological conditions, IDPRs do not have stable three dimensional structure, but they may attain a stable conformation after binding to their biological ligands or other cellular proteins (Dunker et al., 2001, 2002; Uversky, 2019). Disordered structure provides larger surface area for binding to its partner, perform various regulatory roles and have the ability to respond quickly to environmental cues. Specifically, IDPRs of bacterial pathogens can alter the host immune responses either by mimicking host cell signaling components or by forming complexes with proteins of the host cells and thereby disturbing its protein-protein interactions (Marín et al., 2013). It stands to reason that *Mtb* ICL2 (Rv1915 and Rv1916), known to be essential for chronic infection,

TABLE 4 | Kinetic parameters of recombinant ICLs of *Mtb* H37Rv.

Enzyme	Amount of enzyme (μg)	Kinetic parameters			References
		K_m (μM)	K_{cat} (min^{-1})	K_{cat}/K_m (μM^{-1} min^{-1})	
ICL Activity					
1915Δ90CT	10	5.22	4.33	0.83	This study
Rv1916	10	13	6.87	0.53	Antil et al., 2019
Rv0467	2	3.25	30.9	9.5	
MICL Activity					
1915Δ90CT	15	279	3.84	0.0137	This study
Rv1916	25	300	1.2	0.004	Antil et al., 2019
Rv0467	5	240.9	8.05	0.033	

may be playing similar regulatory role facilitated by IDPR. Recently, 3-dimensional structure of *Mtb* ICL2 (CDC1551 strain) along with Small-angle X-ray scattering analyses and Molecular Dynamic simulations have led to molecular level understanding of its allosteric activation at high lipid concentrations and of its function (Bhusal et al., 2019). Future structural studies of Rv1915 and Rv1916 will provide a better picture of their roles in *Mtb*'s virulence.

CONCLUSION

This study reports the cloning and accumulation of recombinant Rv1915/ICL2a as IBs. Although soluble protein could be recovered from these aggregates using βME and sarcosine, however, purification could not be achieved. Amino acid sequence and structure analysis predicted IDPRs in Rv1915/ICL2a, which were further confirmed by *in silico* deletion of the disordered regions and correlation with reduced number of amyloids in the query protein. C-terminal 90 residues deleted recombinant Rv1915Δ90CT could be purified to homogeneity, implementing IDPRs to be responsible for aggregation of Rv1915/ICL2a and deterrent in purification. Presence of IDPR suggests regulatory role for Rv1915/ICL2a by interaction with some cellular partners. Availability of this "difficult to purify" has led to its biochemical characterization and opens venue for structure function studies and inhibitor discovery.

DATA AVAILABILITY STATEMENT

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

AUTHOR CONTRIBUTIONS

MA designed, carried out all the experiments, and compiled the data. SG synthesized 2-methylisocitrate substrate for analyzing MICL activity of Rv1915. VG conceived the idea, procured the extramural funds, analyzed the results, and supervised the study. MA and VG wrote the manuscript.

FUNDING

This work was supported by financial aid from Indian Council of Medical Research (ICMR), Govt. of India through research grant BIC/12(16)/2012). MA is a recipient of INSPIRE fellowship provided by Department of Science and Technology, Govt. of India.

ACKNOWLEDGMENTS

The authors are grateful to Dr. Amulya Kumar Panda (Scientist at National Institute of Immunology, Delhi) for sharing his research knowledge in the field of IBs solubilization and for his guidance in designing experiments. The fellowship and contingency grant of MA is funded through the INSPIRE scheme of Department of Science and Technology, Govt. of India.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fbioe.2020.00522/full#supplementary-material

REFERENCES

Antil, M., Sharma, J., Brissonnet, Y., Choudhary, M., Gouin, S., and Gupta, V. (2019). Structure function insights into elusive *Mycobacterium* tuberculosis protein Rv1916. *Int. J. Biol. Macromol.* 141, 927–936. doi: 10.1016/J.IJBIOMAC.2019.09.038

Bhusal, R. P., Jiao, W., Kwai, B. X. C., Reynisson, J., Collins, A. J., Sperry, J., et al. (2019). Acetyl-CoA-mediated activation of *Mycobacterium* tuberculosis isocitrate lyase 2. *Nat. Commun.* 10, 4639–4639. doi: 10.2210/PDB6EDW/PDB

Bloom, B. R. (1994). *Tuberculosis : Pathogenesis, Protection, and Control.* Washington, DC: ASM Press. Available online at: https://www.worldcat.org/title/tuberculosis-pathogenesis-protection-and-control/oclc/29797333 (accessed June 11, 2019).

Bowden, G. A., Paredes, A. M., and Georgiou, G. (1991). Structure and morphology of protein inclusion bodies in *Escherichia coli*. *Biotechnology* 9, 725–730.

Burgess, R. R. (2009). "Chapter 17 Refolding solubilized inclusion body proteins," in *Methods in Enzymology*, Vol. 463. 259–282. doi: 10.1016/S0076-6879(09)63017-2

Choi, J. H., Keum, K. C., and Lee, S. Y. (2006). Production of recombinant proteins by high cell density culture of *Escherichia coli*. *Chem. Eng. Sci.* 61, 876–885. doi: 10.1016/j.ces.2005.03.031

Chung, T., Klumpp, D. J., and LaPorte, D. C. (1988). Glyoxylate bypass operon of *Escherichia coli*: cloning and determination of the functional map. *J. Bacteriol.* 170, 386–392. doi: 10.1128/jb.170.1.386-392.1988

Cole, S. T., Brosch, R., Parkhill, J., Garnier, T., Churcher, C., Harris, D., et al. (1998). Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature* 393, 537–544. doi: 10.1038/31159

Dunker, A. K., Brown, C. J., Lawson, J. D., Iakoucheva, L. M., and Obradović, Z. (2002). Intrinsic disorder and protein function. *Biochemistry* 41, 6573–6582. doi: 10.1021/bi012159+

Dunker, A. K., Lawson, J. D., Brown, C. J., Williams, R. M., Romero, P., Oh, J. S., et al. (2001). Intrinsically disordered protein. *J. Mol. Graph. Model.* 19, 26–59. doi: 10.1016/S1093-3263(00)00138-8

Gould, T. A., van de Langemheen, H., Munoz-Elias, E. J., McKinney, J. D., and Sacchettini, J. C. (2006). Dual role of isocitrate lyase 1 in the glyoxylate and methylcitrate cycles in *Mycobacterium tuberculosis*. *Mol. Microbiol.* 61, 940–947. doi: 10.1111/j.1365-2958.2006.05297.x

Höner Zu Bentrup, K., Miczak, A., Swenson, D. L., and Russell, D. G. (1999). Characterization of activity and expression of isocitrate lyase in *Mycobacterium avium* and *Mycobacterium tuberculosis*. *J. Bacteriol.* 181, 7161–7167.

Ishida, T., and Kinoshita, K. (2007). PrDOS: prediction of disordered protein regions from amino acid sequence. *Nucleic Acids Res.* 35, W460–W464. doi: 10.1093/nar/gkm363

Khan, R. H., Appa Rao, K. B., Eshwari, A. N. S., Totey, S. M., and Panda, A. K. (1998). Solubilization of recombinant ovine growth hormone with retention of native-like secondary structure and its refolding from the inclusion bodies of *Escherichia coli*. *Biotechnol. Prog.* 14, 722–728. doi: 10.1021/bp980071q

Ko, J., Park, H., Heo, L., and Seok, C. (2012). GalaxyWEB server for protein structure prediction and refinement. *Nucleic Acids Res.* 40, W294–W297. doi: 10.1093/nar/gks493

Krogh, A., Larsson, B., von Heijne, G., and Sonnhammer, E. L. (2001). Predicting transmembrane protein topology with a hidden markov model: application to complete genomes11Edited by F. *Cohen. J. Mol. Biol.* 305, 567–580. doi: 10.1006/jmbi.2000.4315

Marín, M., Uversky, V. N., and Ott, T. (2013). Intrinsic disorder in pathogen effectors: protein flexibility as an evolutionary hallmark in a molecular arms race. *Plant Cell* 25, 3153–3157. doi: 10.1105/tpc.113.116319

McKinney, J. D., zu Bentrup, K. H., Muñoz-Elías, E. J., Miczak, A., Chen, B., Chan, W.-T., et al. (2000). Persistence of *Mycobacterium tuberculosis* in macrophages and mice requires the glyoxylate shunt enzyme isocitrate lyase. *Nature* 406, 735–738. doi: 10.1038/35021074

Mohan Singh, S., and Kumar Panda, A. (2005). Solubilization and refolding of bacterial inclusion body proteins. *J. Biosci. Bioeng.* 99, 303–310. doi: 10.1263/jbb.99.303

Muñoz-Elías, E. J., and McKinney, J. D. (2005). *Mycobacterium* tuberculosis isocitrate lyases 1 and 2 are jointly required for *in vivo* growth and virulence. *Nat. Med.* 11, 638–644. doi: 10.1038/nm1252

Oberg, K., Chrunyk, B. A., Wetzel, R., and Fink, A. L. (1994). Native-like Secondary Structure in Interleukin-1.beta. inclusion bodies by attenuated total reflectance FTIR. *Biochemistry* 33, 2628–2634. doi: 10.1021/bi00175a035

Peng, K., Vucetic, S., Radivojac, P., Brown, C. J., Dunker, A. K., and Obradovic, Z. (2005). Optimizing long intrinsic disorder predictors with protein evolutionary information. *J. Bioinform. Comput. Biol.* 3, 35–60. doi: 10.1142/s0219720005000886

Petersen, T. N., Brunak, S., von Heijne, G., and Nielsen, H. (2011). SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nat. Methods* 8, 785–786. doi: 10.1038/nmeth.1701

Process for solubilization of recombinant proteins expressed as inclusion body (2003). Available online at: https://patents.google.com/patent/US20040235089A1/en (accessed June 13, 2019).

Przybycien, T. M., Dunn, J. P., Valax, P., and Georgiou, G. (1994). Secondary structure characterization of beta-lactamase inclusion bodies. *Protein Eng.* 7, 131–6.

Rosano, G. L., and Ceccarelli, E. A. (2014). Recombinant protein expression in *Escherichia coli*: advances and challenges. *Front. Microbiol.* 5:172. doi: 10.3389/fmicb.2014.00172

Rudolph, R., and Lilie, H. (1996). *In vitro* folding of inclusion body proteins. *FASEB J.* 10, 49–56.

Sambrook, J., and Russell, D. W. (2006). Preparation of plasmid DNA by alkaline lysis with SDS: minipreparation. *Cold Spring Harb. Protoc.* 2006, 911–917. doi: 10.1101/pdb.prot4084

Sharma, V., Sharma, S., Hoener zu Bentrup, K., McKinney, J. D., Russell, D. G., Jacobs, W. R., et al. (2000). Structure of isocitrate lyase, a persistence factor of *Mycobacterium* tuberculosis. *Nat. Struct. Biol.* 7, 663–668. doi: 10.1038/77964

Sonnhammer, E. L., von Heijne, G., and Krogh, A. (1998). A hidden Markov model for predicting transmembrane helices in protein sequences. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* 6, 175–182.

Uversky, V. N. (2013). Intrinsic Disorder-based Protein Interactions and their Modulators. *Curr. Pharm. Des.* 19, 4191–4213. doi: 10.2174/1381612811319230005

Uversky, V. N. (2019). Intrinsically disordered proteins and their "mysterious" (meta) physics. *Front. Phys.* 7:10. doi: 10.3389/fphy.2019.00010

Vallejo, L. F., and Rinas, U. (2004). Strategies for the recovery of active proteins through refolding of bacterial inclusion body proteins. *Microb. Cell Fact.* 3:11. doi: 10.1186/1475-2859-3-11

Vincentelli, R., Bignon, C., Gruez, A., Canaan, S., Sulzenbacher, G., Tegoni, M., et al. (2003). Medium-scale structural genomics: strategies for protein expression and crystallization. *Acc. Chem. Res.* 36, 165–172. doi: 10.1021/ar010130s

Walsh, I., Seno, F., Tosatto, S. C. E., and Trovato, A. (2014). PASTA 2.0: an improved server for protein aggregation prediction. *Nucleic Acids Res.* 42, W301–W307. doi: 10.1093/nar/gku399

Wang, X., Wang, H., and Xie, J. (2011). Genes and regulatory networks involved in persistence of *Mycobacterium tuberculosis*. *Sci. China Life Sci.* 54, 300–310. doi: 10.1007/s11427-011-4134-5

Zhang, Y.-B., Howitt, J., McCorkle, S., Lawrence, P., Springer, K., and Freimuth, P. (2004). Protein aggregation during overexpression limited by peptide extensions with large net negative charge. *Protein Expr. Purif.* 36, 207–216. doi: 10.1016/j.pep.2004.04.020

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 Bioengineering and Biotechnology

ORIGINAL RESEARCH
published: 04 June 2020
doi: 10.3389/fbioe.2020.00569

An Ensemble Approach to Predict Schizophrenia Using Protein Data in the N-methyl-D-Aspartate Receptor (NMDAR) and Tryptophan Catabolic Pathways

Eugene Lin[1,2,3], Chieh-Hsin Lin[3,4,5], Chung-Chieh Hung[6] and Hsien-Yuan Lane[3,6,7,8]**

[1] Department of Biostatistics, University of Washington, Seattle, WA, United States, [2] Department of Electrical & Computer Engineering, University of Washington, Seattle, WA, United States, [3] Graduate Institute of Biomedical Sciences, China Medical University, Taichung, Taiwan, [4] Department of Psychiatry, Kaohsiung Chang Gung Memorial Hospital, Chang Gung University College of Medicine, Kaohsiung, Taiwan, [5] School of Medicine, Chang Gung University, Taoyuan, Taiwan, [6] Department of Psychiatry, China Medical University Hospital, Taichung, Taiwan, [7] Brain Disease Research Center, China Medical University Hospital, Taichung, Taiwan, [8] Department of Psychology, College of Medical and Health Sciences, Asia University, Taichung, Taiwan

OPEN ACCESS

Edited by:
Lavanya Balakrishnan, Mazumdar Shaw Medical Centre, India

Reviewed by:
Sitanshu Sekhar Sahu, Birla Institute of Technology, Mesra, India
Nagarajan Raju, Vanderbilt University Medical Center, United States

***Correspondence:**
Chieh-Hsin Lin
cyndi36@gmail.com
Hsien-Yuan Lane
hylane@gmail.com

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Bioengineering and Biotechnology

Received: *28 October 2019*
Accepted: *11 May 2020*
Published: *04 June 2020*

Citation:
Lin E, Lin C-H, Hung C-C and Lane H-Y (2020) An Ensemble Approach to Predict Schizophrenia Using Protein Data in the N-methyl-D-Aspartate Receptor (NMDAR) and Tryptophan Catabolic Pathways. Front. Bioeng. Biotechnol. 8:569. doi: 10.3389/fbioe.2020.00569

In the wake of recent advances in artificial intelligence research, precision psychiatry using machine learning techniques represents a new paradigm. The D-amino acid oxidase (DAO) protein and its interaction partner, the D-amino acid oxidase activator (DAOA, also known as G72) protein, have been implicated as two key proteins in the N-methyl-D-aspartate receptor (NMDAR) pathway for schizophrenia. Another potential biomarker in regard to the etiology of schizophrenia is melatonin in the tryptophan catabolic pathway. To develop an ensemble boosting framework with random undersampling for determining disease status of schizophrenia, we established a prediction approach resulting from the analysis of genomic and demographic variables such as DAO levels, G72 levels, melatonin levels, age, and gender of 355 schizophrenia patients and 86 unrelated healthy individuals in the Taiwanese population. We compared our ensemble boosting framework with other state-of-the-art algorithms such as support vector machine, multilayer feedforward neural networks, logistic regression, random forests, naive Bayes, and C4.5 decision tree. The analysis revealed that the ensemble boosting model with random undersampling [area under the receiver operating characteristic curve (AUC) = 0.9242 ± 0.0652; sensitivity = 0.8580 ± 0.0770; specificity = 0.8594 ± 0.0760] performed maximally among predictive models to infer the complicated relationship between schizophrenia disease status and biomarkers. In addition, we identified a causal link between DAO and G72 protein levels in influencing schizophrenia disease status. The study indicates that the ensemble boosting framework with random undersampling may provide a suitable method to establish a tool for distinguishing schizophrenia patients from healthy controls using molecules in the NMDAR and tryptophan catabolic pathways.

Keywords: ensemble boosting, multilayer feedforward neural networks, N-methyl-D-aspartate receptor, precision psychiatry, schizophrenia

INTRODUCTION

Precision psychiatry, an emerging interdisciplinary paradigm of psychiatry and precision medicine, is progressing into the cornerstone of public health practice (Katsanis et al., 2008; Snyderman, 2012). In terms of diagnostic and therapeutic decisions, precision psychiatry is tailored to the specific patient with psychiatric disorders (Katsanis et al., 2008; Snyderman, 2012). More generally, multiple data types such as genomics and protein data are integrated with state-of-the-art artificial intelligence and machine learning algorithms. Thereby, these integrated frameworks are able to correspondingly learn to provide proper clinical decisions during nearly every stage of patient care in an individual manner, such as diagnosis and treatment of psychiatric disorders (Lin and Chen, 2008a; Lane et al., 2012; Lin and Lane, 2015, 2017). For example, a recent study utilized machine learning models to optimize prediction of antidepressant treatment outcome in patients with major depressive disorder by using genetic and clinical datasets (Lin et al., 2018a).

The N-methyl-D-aspartate receptor (NMDAR) pathway has been a focus of attention in schizophrenia research. The D-amino acid oxidase (DAO) protein and its putative activator, the D-amino acid oxidase activator (DAOA, also known as G72) protein, are two proteins in the NMDAR pathway. *In vitro* studies reported that the G72 protein activates and binds to the DAO protein (Chumakov et al., 2002; Sacchi et al., 2008). Next, the DAO protein in turn oxidizes D-amino acids such as D-serine, an agonist of NMDAR (Chumakov et al., 2002; Sacchi et al., 2008). It has been hypothesized that patients who over-yield the G72 protein may reduce the NMDAR activities, thereby inclining them to schizophrenia (Hashimoto et al., 2003; Lin et al., 2014; Lin and Lane, 2019). Furthermore, it has been suggested that plasma G72 protein levels are notably higher in patients with schizophrenia than in healthy individuals (Lin et al., 2014). Moreover, it has been indicated that the agonist activities in the NMDAR pathway possess appropriate importance in developing novel drug targets for treatment of schizophrenia (Coyle et al., 2003; Goff, 2012; Javitt, 2012; Moghaddam and Javitt, 2012; Ermilov et al., 2013; Lane et al., 2013; Lin et al., 2017a, 2018; Chang et al., 2019). To distinguish healthy individuals from patients with schizophrenia, a previous study also utilized machine learning algorithms (such as logistic regression, naive Bayes, and C4.5 decision tree) to construct predictive models by using the G72 protein and genetic variants (Lin et al., 2018b).

Melatonin, which has an impact on the tryptophan catabolic pathway, is another probable factor with respect to the developmental etiology of schizophrenia (Anderson and Maes, 2012). It is proposed that melatonin plays a role as a biomarker of schizophrenia although the findings were controversial (Morera-Fumero and Abreu-Gonzalez, 2013). It has been reported that plasma melatonin levels were higher, lower, or similar in patients with schizophrenia as compared to healthy controls (Morera-Fumero and Abreu-Gonzalez, 2013). Schizophrenia is also linked with both circadian and metabolic disorders, which are modulated by melatonin (Wulff et al., 2012).

Here, in order to distinguish schizophrenia patients from healthy controls in the Taiwanese population, we employed an ensemble boosting algorithm to build predictive models of schizophrenia disease status by using DAO and G72 protein levels in the NMDAR pathway as well as by using melatonin levels in the tryptophan catabolic pathway. To deal with imbalanced data, we also utilized the random undersampling method at the data level (Galar et al., 2011). To the best of our knowledge, no previous studies have been performed to evaluate predictive models for schizophrenia disease status by using ensemble boosting techniques with random undersampling. We selected the ensemble boosting algorithms because these algorithms are regularly applied to solve complex problems in classification and predictive modeling owing to their superiority in reduction of overfitting, consistency, robust prediction, and better generalization (Yang et al., 2010; Galar et al., 2011; Zhang et al., 2019). This study directly compared the performance of the ensemble boosting models to widely used machine learning algorithms, including support vector machine (SVM), multi-layer feedforward neural networks (MFNNs), logistic regression, random forests, naive Bayes, and C4.5 decision tree. Our analysis demonstrated that our ensemble boosting approach with random undersampling led to better performance.

MATERIALS AND METHODS

Study Population

The study cohort consisted of 355 schizophrenia patients and 86 unrelated healthy controls, who were recruited from the China Medical University Hospital in Taiwan. In this study, both schizophrenia patients and healthy controls were aged 18–65 years, were healthy in the neurological and physical conditions, and had obtained normal laboratory assessments (such as blood routine and biochemical tests). Details of the diagnosis of schizophrenia were published previously (Lin et al., 2014). Briefly, the research psychiatrists evaluated both patients and healthy volunteers by using the Structured Clinical Interview for DSM-IV (SCID) for diagnosis (Lin et al., 2014).

After presenting a complete description of this study to the subjects, we obtained written informed consents in line with the institutional review board guidelines. This study was approved by the institutional review board of the China Medical University Hospital in Taiwan and was conducted in accordance with the Declaration of Helsinki.

Laboratory Assessments

Plasma G72 protein expression levels were measured by western blotting (Lin et al., 2014). Shortly after 10 mL of blood was collected into EDTA-containing blood collection tubes by using sterile techniques, we processed the blood specimens shortly by using centrifugation at 500 g. After centrifugation, we directly dissected plasma and rapidly stored it at −80°C until western blotting. For western blotting, we depleted 100 μL plasma by using ProteoPrep® Blue Albumin and IgG Depletion Kit. All western blot experiments were repeated for two times.

DAO levels in the serum were measured using commercially available enzyme-linked immunosorbent assay (ELISA) kits according to the manufacture's recommended protocol (Cloud-Clone Corp, Houston, TX, United States). The detailed method has been described elsewhere (Lin et al., 2017b).

Melatonin protein concentrations were measured using commercially available enzyme-linked immunosorbent assay (ELISA) kits according to the manufacture's recommended protocol (MyBioSource, San Diego, CA, United States). Briefly, 100 μL plasma samples and the standard were added to each well of a 96-well plate. The solutions were incubated for 2 h at 37°C. The liquid was then removed. 100 μL Biotin-antibody (1×) was added to each well and incubated for 1 h at 37°C. Each well was washed with buffer for three times. 100 μL HRP-avidin (1×) was added to each well and incubated for 1 h at 37°C. Each well was washed with buffer for five times and then incubated with 90 μL substrate solution for 15–30 min at 37°C with the protection from light. 50 μL stop solution was added to each well, and mixed thoroughly. A Benchmark Plus Microplate Reader (Bio-Rad) was used to read the optical density at 450 nm. The concentrations of melatonin in the samples were determined according to a standard curve.

Statistical Analysis

The Student's *t*-test was conducted to measure the difference in the means of two continuous variables (Lin et al., 2019). We performed the chi-square test for categorical data. The Kruskal-Wallis test was used to determine if there is statistically significant difference between schizophrenia patients and healthy controls on DAO, G72, and melatonin levels. Furthermore, we utilized multivariable logistic regression analysis to assess causal links between DAO, G72, and melatonin levels with adjustment for age and gender. The criterion for significance was set at $P < 0.05$ for all tests. Data are presented as the mean ± standard deviation.

Ensemble Boosting Predictive Models

We employed a key ensemble boosting technique called LogitBoost (Friedman et al., 2000) and utilized the Waikato Environment for Knowledge Analysis (WEKA) software (which is available from https://www.cs.waikato.ac.nz/ml/weka/) (Witten et al., 2005) to carry out the predictive ensemble framework. All the experiments were conducted on a computer with Intel (R) Core (TM) i5-4210U, 4 GB RAM, and Windows 7.

The LogitBoost algorithm is an ensemble boosting approach, which combines the performance of many weak classifiers (also referred to as base classifiers) to achieve a robust classifier with higher accuracy. **Figure 1** shows the illustrative diagram of the ensemble boosting method. The LogitBoost algorithm utilizes a binomial log-likelihood method that changes the classification error linearly so that LogitBoost tends to be robust in handling outliers and noisy data. The base classifier we employed is a decision stump, which is a one-level decision tree (that is, a decision tree with a root node and two leaf nodes). Here, we used the default parameters of WEKA, such as 1.0 for the shrinkage parameter, 100 for the batch size, 3.0 for the Z max threshold, and 10 for the number of iterations.

Furthermore, we utilized a random undersampling technique which eliminates instances in the majority class to balance class distribution (Galar et al., 2011). We further combined the LogitBoost algorithm with the random undersampling technique.

Machine Learning Algorithms for Benchmarking

For the benchmarking task in the present study, we utilized six state-of-the-art machine learning algorithms including SVM, MFNNs, logistic regression, random forests, naive Bayes, and C4.5 decision tree to compare with the ensemble boosting model. We carried out the analyses for these six machine learning algorithms using the WEKA software (Witten et al., 2005) and a computer with Intel (R) Core (TM) i5-4210U, 4 GB RAM, and Windows 7.

The SVM algorithm (Vapnik, 2013) is a popular technique for pattern recognition and classification. Given a training set of instance-label pairs, the SVM algorithm leverages a kernel function to map the training vectors into a higher dimensional space (Lin and Hwang, 2008b; Vapnik, 2013). In this higher dimensional space, the SVM algorithm then finds a linear separating hyperplane with the maximal margin. In this study, we used the Pearson VII function-based universal kernel (Üstün et al., 2006) with the omega value of 1.0 and the sigma value of 0.5.

An MFNN framework consists of one input layer, one or multiple hidden layers, and one output layer, where connections among neuron structures consist of no directed cycles (Bishop, 1995). In the learning period of the MFNN framework, the back-propagation algorithm (Rumelhart et al., 1996) is leveraged for the learning strategy. In the retrieving period, the MFNN framework repeats via all the structures to perform the retrieval process at the output panel in keeping with the inputs of test patterns (Kung and Hwang, 1998).

We used the logistic regression model, the standard method for classification problems in clinical applications (Witten et al., 2005), as a basis for comparison. In addition, we employed the naive Bayes model that assumes the presence or absence of a particular feature is unrelated to the presence or absence of any other feature (Witten et al., 2005). The naive Bayes model calculates the probability that a given instance belongs to a certain class (that is, "schizophrenia patient" or "healthy control" in this study) by using the Bayes' theorem.

The random forests model is an ensemble learning method that composes a collection of decision trees during training and yields the class that is the mode of the classes among the individual trees (Breiman, 2001). Here, we used the default parameters of WEKA for the random forests model; for example, 100 for the batch size and 100 for the number of iterations.

The C4.5 decision tree model builds decision trees top-down and prunes them using the concept of information entropy (Witten et al., 2005). First, the tree is constructed by finding the root node (for example, protein level) that is most discriminative one for differentiating "schizophrenia patient" from "healthy control." Then, the best single feature test is decided by the information gain and by choosing a feature (for example, protein level) to split the data into subsets. Here, we used the default

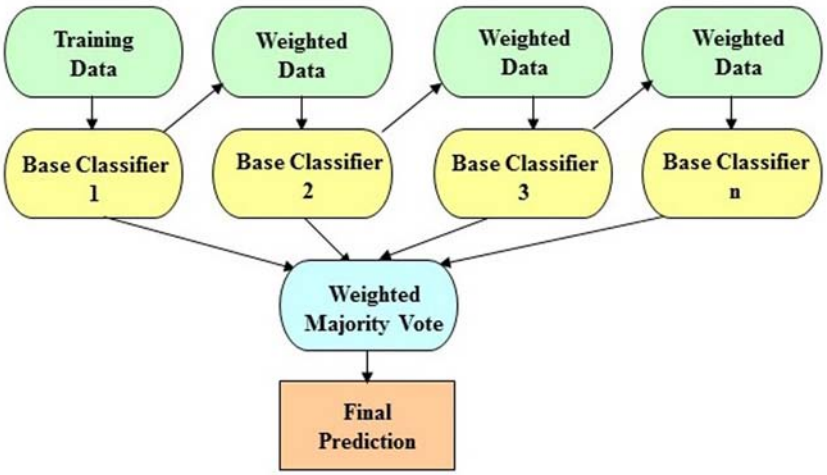


FIGURE 1 | The schematic illustration of the ensemble boosting method. The idea of the ensemble boosting approach is to train weak/base classifiers sequentially in a way that each classifier tries to correct its predecessor. A higher weight is assigned to samples that were incorrectly classified by earlier rounds. That is, week/base classifiers are produced in sequence based on a weighted version of the data during the training phase. The final classification prediction is then produced by a weighted majority vote.

parameters of WEKA, such as 0.25 for the confidence factor and 2 for the minimum number of instances per leaf node (Huang et al., 2009).

Evaluation of the Predictive Performance

In this study, we utilized the receiver operating characteristic (ROC) methodology and determined the area under the ROC curve (AUC) to assess the performance of predictive models (Linden, 2006; Lin and Hwang, 2008b; Huang et al., 2009). The better the prediction model, the higher the AUC (Linden, 2006; Huang et al., 2009). In additional, we calculated sensitivity (that is, the proportion of correctly predicted responders of all tested responders) as:

$$\text{Sensitivity} = \text{True Positive}/(\text{True Positive} + \text{False Negative})$$

and specificity (that is, the proportion of correctly predicted non-responders of all the tested non-responders) as:

$$\text{Specificity} = \text{True Negative}/(\text{True Negative} + \text{False Positive}).$$

Moreover, we utilized the repeated 10-fold cross-validation method and leave-one-out cross-validation method to examine the generalization of predictive models (Huang et al., 2009; Lin and Hsu, 2009).

RESULTS

The Study Cohort in the Taiwanese Population

The participants included 355 schizophrenia patients and 86 unrelated healthy individuals in the Taiwanese population. As shown in **Table 1**, there was no significant difference in gender (P = 0.101) and age (P = 0.136) distributions between the two groups. The mean age (39.6 ± 10.0 years) of schizophrenia patients was older than that of healthy controls (37.8 ± 12.2 years). The mean level of DAO protein in the plasma of schizophrenia patients was considerably higher than that of healthy controls (37.64 ± 14.18 ng/mL vs. 28.03 ± 9.84 ng/mL; $P = 5.55 \times 10^{-9}$) (**Table 1**). In addition, the mean level of G72 protein in the plasma of schizophrenia patients was markedly higher than that of healthy controls (3.24 ± 1.80 ng/μL vs. 1.68 ± 0.81 ng/μL; $P = 4.71 \times 10^{-14}$) (**Table 1**). Moreover, the mean level of melatonin in the plasma of schizophrenia patients was notably higher than that of healthy controls (89.89 ± 46.07 pg/mL vs. 60.04 ± 42.72 pg/mL; $P = 9.75 \times 10^{-7}$) (**Table 1**).

The significant Kruskal-Wallis test was shown for DAO, G72, and melatonin levels ($P = 3.12 \times 10^{-9}$, 2.2×10^{-16}, and 3.35×10^{-6}, respectively) between schizophrenia patients and healthy controls. **Supplementary Figure S1** shows the

TABLE 1 | Demographic characteristics of schizophrenia patients and healthy individuals.

Characteristic	Schizophrenia patients	Healthy individuals	*P*-value[a]
No. of subjects (n)	355	86	
Gender (male)%	61.9%	52.3%	0.101
Age (year)	39.6 ± 10.0	37.8 ± 12.2	0.136
DAO level (ng/mL)	37.64 ± 14.18	28.03 ± 9.84	5.55×10^{-9}
G72 level (ng/μL)	3.24 ± 1.80	1.68 ± 0.81	4.71×10^{-14}
Melatonin level (pg/mL)	89.89 ± 46.07	60.04 ± 42.72	9.75×10^{-7}

[a]*Chi-square test for the categorical data; Student's t-test for continuous variables. Data are presented as mean ± standard deviation. DAO, D-amino acid oxidase; G72 (also known as DAOA), D-amino acid oxidase activator.*

TABLE 2 | The results of repeated 10-fold cross-validation experiments for differentiating schizophrenia patients from healthy individuals using ensemble boosting with random undersampling, ensemble boosting, SVM, MFNNs, logistic regression, random forests, naive Bayes, and C4.5 decision tree with biomarkers such as DAO protein levels, G72 protein levels, melatonin protein levels, age, and gender.

Algorithm	AUC	Sensitivity	Specificity	Number of biomarkers
Ensemble boosting with random undersampling	0.9242 ± 0.0652	0.8580 ± 0.0770	0.8594 ± 0.0760	5
Ensemble boosting	0.9010 ± 0.0464	0.8442 ± 0.0447	0.5803 ± 0.1446	5
SVM	0.6720 ± 0.0837	0.8461 ± 0.0393	0.4979 ± 0.1364	5
MFNN with 1 hidden layer	0.8920 ± 0.0463	0.8343 ± 0.0457	0.5816 ± 0.1340	5
MFNN with 2 hidden layers	0.8949 ± 0.0455	0.8391 ± 0.0515	0.6121 ± 0.1383	5
MFNN with 3 hidden layers	0.8884 ± 0.0507	0.8359 ± 0.0463	0.6312 ± 0.1454	5
Logistic Regression	0.8677 ± 0.0566	0.8497 ± 0.0566	0.5660 ± 0.1295	5
Random Forests	0.8543 ± 0.0627	0.8229 ± 0.0379	0.4197 ± 0.1213	5
naive Bayes	0.8546 ± 0.0628	0.8320 ± 0.0473	0.6611 ± 0.1411	5
C4.5 decision tree	0.7701 ± 0.0721	0.8306 ± 0.0469	0.4526 ± 0.1272	5

AUC, the area under the receiver operating characteristic curve; DAO, D-amino acid oxidase; G72 (also known as DAOA), D-amino acid oxidase activator; MFNNs, multilayer feedforward neural networks; SVM, support vector machine. Data are presented as mean ± standard deviation.

distribution charts of three features (such as DAO, G72, and melatonin levels) and other variables for schizophrenia patients and healthy controls. The distribution charts are grouped separately by two subsets, namely schizophrenia patients (shown in the red color) and healthy controls (shown in the blue color). As shown in **Supplementary Figure S1**, the number of schizophrenia patients was much larger than the number of healthy controls.

Predictive Models for Schizophrenia Disease Status

In this study, we used five biomarkers including DAO levels, G72 levels, melatonin levels, age, and gender to build the predictive models for differentiating schizophrenia patients from healthy individuals by employing the ensemble boosting framework. **Table 2** summarizes the results of repeated 10-fold cross-validation experiments by ensemble boosting (with random undersampling), SVM, MFNNs, logistic regression, random forests, naive Bayes, and C4.5 decision tree using five biomarkers. To measure the performance of prediction models, we used the ROC methodology and calculated the AUC, sensitivity, and specificity for these predictive models using five biomarkers.

Supplementary Figures S2–S4 show plots of ROC, precision-recall, and sensitivity-specificity curves for ensemble boosting with random undersampling using five biomarkers, respectively. **Supplementary Figures S5–S10** show plots of ROC, precision-recall, and sensitivity-specificity curves for ensemble boosting, SVM, MFNNs, logistic regression, random forests, naive Bayes, and C4.5 decision tree using five biomarkers.

As shown in **Supplementary Figure S2**, the lower left point (0, 0) on the ROC curve represents a false positive rate of 0% (that is, no false positive errors) and a true positive rate of 0% (that is, no true positives), indicating never having a positive classification. On the contrary, the upper right point (1, 1) represents a false positive rate of 100% and a true positive rate of 100%, indicating completely having positive classifications. Furthermore, if we assume that the point (0.1406, 0.858) is on the ROC curve, the point (0.1406, 0.858) shows a false positive rate of 14.06% (or specificity of 0.8594) and a true positive rate of 85.8% (or sensitivity of 0.858).

As shown in **Supplementary Figure S3**, if we assume that the point (0.858, 0.8546) is on the precision-recall curve, the point (0.858, 0.8546) shows a true positive rate of 85.8% (or recall/sensitivity of 0. 858) and a precision value of 85.46%. Additionally, as shown in **Supplementary Figure S4**, if we assume that the point (0.8594, 0.858) is on the sensitivity-specificity curve, the point (0.8594, 0.858) shows a true negative rate of 85.94% (or specificity of 0.8594) and a true positive rate of 85.8% (or sensitivity of 0.858).

In addition, **Supplementary Tables S1–S3** summarize the results of repeated 10-fold cross-validation experiments by ensemble boosting (with random undersampling), SVM, MFNNs, logistic regression, random forests, naive Bayes, and C4.5 decision tree using individual features such as DAO (**Supplementary Table S1**), G72 (**Supplementary Table S2**), and melatonin (**Supplementary Table S3**) levels, respectively.

Ensemble Boosting Model for Schizophrenia Disease Status

For the ensemble boosting model for forecasting schizophrenia disease status, we performed a series of different datasets using five biomarkers as well as individual features. As indicated in **Table 2**, the average value of AUC for the ensemble boosting prediction model with random undersampling was 0.9242 ± 0.0652 using five biomarkers including DAO levels, G72 levels, melatonin levels, age, and gender. As indicated in **Supplementary Tables S1–S3**, the average values of AUC for the ensemble boosting prediction model with random undersampling were 0.6471 ± 0.1062, 0.7314 ± 0.1121, and 0.8462 ± 0.0873 using individual features such as DAO levels, G72 levels, and melatonin levels, respectively.

Benchmarking

To evaluate the performance of our approach for predictive models for schizophrenia disease status, we compared the

ensemble boosting model with other state-of-the-art methods, including SVM, MFNNs, logistic regression, random forests, naive Bayes, and C4.5 decision tree.

For MFNN models for forecasting schizophrenia disease status, we performed a series of different architectures containing 1, 2, and 3 hidden layers. **Supplementary Figures S11–S13** show an example of architecture of the MFNN model with 3, 2, and 1 hidden layer(s), respectively. As indicated in **Table 2**, the average values of AUC for the MFNN prediction models of 1, 2, and 3 hidden layers were 0.8920 ± 0.0463, 0.8949 ± 0.0455, and 0.8884 ± 0.0507, respectively. **Supplementary Figures S14–S16** show cost/loss function measurement plots of the MFNN model with 3, 2, and 1 hidden layer(s), respectively. Of all the MFNN prediction models, the MFNN model with 2 hidden layers yielded better performance than the other two models in terms of AUC. Thus, there was no significant improvement in the sensitivity with the increase in hidden layers. Moreover, the specificity was low, indicating that the model provides more false positives. This may have been due to an imbalance in the dataset.

Supplementary Table S4 shows WEKA's hyper-parameters for training the MFNN models with 1–3 hidden layers. For example, we used the following WEKA's parameters for training the MFNN model with one hidden layer: the momentum = 0.01, the learning rate = 0.05, the batch size = 100, and the number of epochs = 500.

As shown in **Table 2**, the ensemble boosting model with random undersampling performed maximally in all cases. The best AUC was 0.9242 ± 0.0652, which was based on the ensemble boosting model with random undersampling (**Table 2**). Our analysis indicated that the ensemble boosting model with random undersampling was well-suited for predictive models for schizophrenia disease status. Furthermore, the ensemble boosting model with random undersampling performed best in both sensitivity (0.8580 ± 0.0770) and specificity (0.8594 ± 0.0760) (**Table 2**).

Leave-One-Out Cross-Validation Experiments

In this study, we also explored the generalization of predictive models using the leave-one-out cross-validation method. **Supplementary Table S5** summarizes the results of leave-one-out cross-validation experiments by ensemble boosting (with random undersampling), SVM, MFNNs, logistic regression, random forests, naive Bayes, and C4.5 decision tree using five biomarkers such as DAO levels, G72 levels, melatonin levels, age, and gender. In addition, **Supplementary Tables S6–S8** summarize the results of leave-one-out cross-validation experiments by ensemble boosting (with random undersampling), SVM, MFNNs, logistic regression, random forests, naive Bayes, and C4.5 decision tree using individual features such as DAO (**Supplementary Table S6**), G72 (**Supplementary Table S7**), and melatonin (**Supplementary Table S8**) levels, respectively.

As indicated in **Supplementary Table S5**, the AUC value for the ensemble boosting prediction model with random undersampling was 0.937 using five biomarkers including DAO levels, G72 levels, melatonin levels, age, and gender. As indicated in **Supplementary Tables S6–S8**, the AUC values for the ensemble boosting prediction model with random undersampling were 0.603, 0.610, and 0.826 using individual features such as DAO levels, G72 levels, and melatonin levels, respectively.

As shown in **Supplementary Table S5**, the best AUC was 0.937, which was based on the ensemble boosting model with random undersampling using five biomarkers such as DAO levels, G72 levels, melatonin levels, age, and gender. Furthermore, the ensemble boosting model with random undersampling performed best in both sensitivity (0.855) and specificity (0.855) (**Supplementary Table S5**).

Causal Links Between Protein Levels

Finally, we assessed causal links among DAO levels, G72 levels, and melatonin levels in predicting schizophrenia disease status with age and sex as covariates. In our analysis, there was a significant causal link involving DAO levels and G72 levels (P = 0.0036) in influencing schizophrenia disease status. However, there were no causal links either between DAO levels and melatonin levels or between G72 levels and melatonin levels.

DISCUSSION

To our knowledge, this is the first study to date to leverage an ensemble boosting approach with random undersampling for building predictive models of schizophrenia disease status among Taiwanese individuals. Moreover, we performed the first study to predict schizophrenia disease status by utilizing protein data in both the NMDAR and tryptophan catabolic pathways. The findings pinpointed that the ensemble boosting model with random undersampling using five biomarkers outperformed other state-of-the-art predictive models in terms of AUC for distinguishing schizophrenia patients from healthy controls. The five biomarkers encompassed DAO levels, G72 levels, melatonin levels, age, and gender. In addition, we found that a significant causal link between DAO and G72 protein levels possessed a strong potential to reflect schizophrenia disease status. By leveraging the molecular data in the NMDAR and tryptophan catabolic pathways, we establish the predictive models of schizophrenia disease status by using the ensemble boosting framework with random undersampling. Our data also suggest that our ensemble boosting models with random undersampling may provide a suitable approach to create predictive models for forecasting schizophrenia disease status with clinically meaningful accuracy. Therefore, the ensemble boosting approach with random undersampling in this study is a proof of concept of a machine learning predictive tool for discriminating schizophrenia patients from healthy individuals.

Remarkably, an intriguing finding was that we further inferred the causal link between DAO and G72 protein levels in influencing schizophrenia disease status. To our knowledge,

scanty human studies have been conducted to evaluate causal links between DAO and G72 protein levels. The biological mechanisms of these causal links in schizophrenia disease status remain to be elucidated. In line with our results, an *in vitro* study identified a physical interaction between DAO and G72 proteins using yeast two-hybrid experiments (Chumakov et al., 2002). Moreover, a recent study found a putative correlation between DAO and G72 protein expressions in the brain regions such as the brainstem, cerebellum, amygdala, and thalamus (except for the frontal cortex) by using post-mortem brain samples in normal human subjects (Jagannath et al., 2017).

In this study, the dataset is highly imbalanced because the class of schizophrenia patients is significantly larger in terms of instances than the class of healthy controls. To overcome this limitation, we employed the random undersampling method to balance class distribution. Without random undersampling, the predictive models tend to have lower specificity values. In line with previous findings (Chawla et al., 2004; Galar et al., 2011), we found that the ensemble boosting model with random undersampling is highly suitable for handling class imbalances. It has also been suggested to use more accurate measures such as AUC to evaluate predictive models in the case of class imbalances (Chawla et al., 2004).

Furthermore, it is worthwhile to bring the discussion on the random undersampling method for dealing with the imbalanced data (that is, the bigger number of schizophrenia patients vs. the smaller number of healthy controls) in our study. Due to the imbalanced data, the models without the random undersampling method showed predictions that were clearly biased toward higher sensitivity and lower specificity. For example, without random undersampling, sensitivity was around 80% and specificity was around 50–60% for the models using the combined biomarkers of DAO, G72, and melatonin protein levels (**Table 2**). On the contrary, ensemble boosting with random undersampling had sensitivity of 85.8% and specificity of 85.94% for the combined biomarkers (**Table 2**). The models with individual biomarkers were also in the similar situation (**Supplementary Tables S1–S3**). For instance, without random undersampling, sensitivity was around 80% and specificity was around 40% for the models using individual melatonin protein levels (**Supplementary Table S3**). On the other hand, ensemble boosting with random undersampling had sensitivity of 77.19% and specificity of 77.44% for melatonin protein levels (**Supplementary Table S3**). Therefore, predictions were no longer biased toward higher sensitivity and lower specificity by using ensemble boosting with random undersampling. Our improved results demonstrate that the ensemble boosting model with random undersampling provides an effective way to solve the imbalanced data problem in our study.

CONCLUSION

In conclusion, we created an ensemble boosting predictive framework with random undersampling for estimating schizophrenia disease status in Taiwanese subjects by using DAO and G72 protein datasets in the NMDAR pathway as well as by using melatonin dataset in the tryptophan catabolic pathway. The analysis indicates that our ensemble boosting framework with random undersampling could contribute a conceivable way to construct predictive algorithms for determining schizophrenia disease status in terms of clinically purposeful performance. Consequently, we would foresee that the findings of this study may be generalized for genomic medicine studies in precision psychiatry to forecast disease status and treatment response for psychiatric disorders. Furthermore, the findings may be potentially adopted to provide molecular diagnostic and prognostic tools in the coming years. It is indispensable to unfold further discoveries into the role of the machine learning predictive framework explored in this study by using replication studies with independent samples.

DATA AVAILABILITY STATEMENT

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

ETHICS STATEMENT

This study was approved by the Institutional Review Board of China Medical University Hospital, Taiwan and complies with the Declaration of Helsinki. Informed written consent was obtained from all participants.

AUTHOR CONTRIBUTIONS

EL, C-HL, and H-YL designed the study and revised the manuscript. C-HL, C-CH, and H-YL conducted the study. EL analyzed the data and drafted the manuscript. All authors provided the final approval of the version to be published.

FUNDING

This work was supported by the National Health Research Institutes, Taiwan (NHRI-EX109-10731NI), the Ministry of Science and Technology in Taiwan (MOST 108-2314-B-039-002, MOST 109-2312-B-039-001, and MOST 109-2622-B-039-001-CC2), and China Medical University Hospital, Taiwan (DMR-109-246).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fbioe.2020.00569/full#supplementary-material

REFERENCES

Anderson, G., and Maes, M. (2012). Melatonin: an overlooked factor in schizophrenia and in the inhibition of anti-psychotic side effects. *Metab. Brain Dis.* 27, 113–119. doi: 10.1007/s11011-012-9307-9

Bishop, C. M. (1995). *Neural Networks for Pattern Recognition*. Oxford: Clarendon Press.

Breiman, L. (2001). Random forests. *Machine Learn.* 45, 5–32.

Chang, C.-H., Lane, H.-Y., Tseng, P.-T., Chen, S.-J., Liu, C.-Y., and Lin, C.-H. (2019). Effect of N-methyl-D-aspartate-receptor-enhancing agents on cognition in patients with schizophrenia: a systematic review and meta-analysis of double-blind randomised controlled trials. *J. Psychopharmacol.* 33, 436–448. doi: 10.1177/0269881118822157

Chawla, N. V., Japkowicz, N., and Kotcz, A. (2004). Special issue on learning from imbalanced data sets. *SIGKDD Explor.* 6, 1–6. doi: 10.1145/1007730.1007733

Chumakov, I., Blumenfeld, M., Guerassimenko, O., Cavarec, L., Palicio, M., Abderrahim, H., et al. (2002). Genetic and physiological data implicating the new human gene G72 and the gene for D-amino acid oxidase in schizophrenia. *Proc. Natl. Acad. Sci. U.S.A.* 99, 13675–13680.

Coyle, J. T., Tsai, G., and Goff, D. (2003). Converging evidence of NMDA receptor hypofunction in the pathophysiology of schizophrenia. *Ann. N. Y. Acad. Sci.* 1003, 318–327. doi: 10.1196/annals.1300.020

Ermilov, M., Gelfin, E., Levin, R., Lichtenberg, P., Hashimoto, K., Javitt, D. C., et al. (2013). A pilot double-blind comparison of d-serine and high-dose olanzapine in treatment-resistant patients with schizophrenia. *Schizophr. Res.* 150, 604–605. doi: 10.1016/j.schres.2013.09.018

Friedman, J., Hastie, T., and Tibshirani, R. (2000). Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). *Ann. Stat.* 28, 337–407. doi: 10.1214/aos/1016218223

Galar, M., Fernandez, A., Barrenechea, E., Bustince, H., and Herrera, F. (2011). A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches. *IEEE Trans. Syst. Man Cybern. C Appl. Rev.* 42, 463–484. doi: 10.1109/tsmcc.2011.2161285

Goff, D. C. (2012). D-cycloserine: an evolving role in learning and neuroplasticity in schizophrenia. *Schizophr. Bull.* 38, 936–941. doi: 10.1093/schbul/sbs012

Hashimoto, K., Fukushima, T., Shimizu, E., Komatsu, N., Watanabe, H., Shinoda, N., et al. (2003). Decreased serum levels of D-serine in patients with schizophrenia: evidence in support of the N-methyl-D-aspartate receptor hypofunction hypothesis of schizophrenia. *Arch. Gen. Psychiatry* 60, 572–576.

Huang, L. C., Hsu, S. Y., and Lin, E. (2009). A comparison of classification methods for predicting chronic fatigue syndrome based on genetic data. *J. Transl. Med.* 7:81. doi: 10.1186/1479-5876-7-81

Jagannath, V., Marinova, Z., Monoranu, C.-M., Walitza, S., and Grünblatt, E. (2017). Expression of D-amino acid oxidase (DAO/DAAO) and D-amino acid oxidase activator (DAOA/G72) during development and aging in the human post-mortem brain. *Front. Neuroanat.* 11:31. doi: 10.3389/fnana.2017.00031

Javitt, D. C. (2012). Twenty-five years of glutamate in schizophrenia: are we there yet? *Schizophr. Bull.* 38, 911–913. doi: 10.1093/schbul/sbs100

Katsanis, S. H., Javitt, G., and Hudson, K. (2008). Public health. A case study of personalized medicine. *Science* 320, 53–54. doi: 10.1126/science.1156604

Kung, S. Y., Hwang, J. N. (1998). Neural networks for intelligent multimedia processing. *Proc. IEEE* 86, 1244–1272. doi: 10.1109/5.687838

Lane, H. Y., Lin, C. H., Green, M. F., Hellemann, G., Huang, C. C., Chen, P. W., et al. (2013). Add-on treatment of benzoate for schizophrenia: a randomized, double-blind, placebo-controlled trial of D-amino acid oxidase inhibitor. *JAMA Psychiatry* 70, 1267–1275. doi: 10.1001/jamapsychiatry.2013.2159

Lane, H. Y., Tsai, G. E., and Lin, E. (2012). Assessing gene-gene interactions in pharmacogenomics. *Mol. Diagn. Ther.* 16, 15–27. doi: 10.2165/11597270-000000000-00000

Lin, C. H., Chang, H. T., Chen, Y. J., Lin, C. H., Huang, C. H., Tun, R., et al. (2014). Distinctively higher plasma G72 protein levels in patients with schizophrenia than in healthy individuals. *Mol. Psychiatry* 19, 636–637. doi: 10.1038/mp.2013.80

Lin, C.-H., and Lane, H.-Y. (2019). Early identification and intervention of schizophrenia: insight from hypotheses of glutamate dysfunction and oxidative stress. *Front. Psychiatry* 10:93. doi: 10.3389/fpsyt.2019.00093

Lin, C. H., Lin, C. H., Chang, Y. C., Huang, Y. J., Chen, P. W., Yang, H. T., et al. (2017a). Sodium benzoate, a D-amino acid oxidase inhibitor, added to clozapine for the treatment of schizophrenia: a randomized, double-blind, placebo-controlled trial. *Biol. Psychiatry* 84, 422–432. doi: 10.1016/j.biopsych.2017.12.006

Lin, C.-H., Lin, C.-H., Chang, Y.-C., Huang, Y.-J., Chen, P.-W., Yang, H.-T., et al. (2018). Sodium benzoate, a D-amino acid oxidase inhibitor, added to clozapine for the treatment of schizophrenia: a randomized, double-blind, placebo-controlled trial. *Biol. Psychiatry* 84, 422–432.

Lin, C.-H., Yang, H.-T., Chiu, C.-C., and Lane, H.-Y. (2017b). Blood levels of D-amino acid oxidase vs. D-amino acids in reflecting cognitive aging. *Sci. Rep.* 7:14849.

Lin, E., and Chen, P. S. (2008a). Pharmacogenomics with antidepressants in the STAR*D study. *Pharmacogenomics* 9, 935–946. doi: 10.2217/14622416.9.7.935

Lin, E., and Hsu, S. Y. (2009). A Bayesian approach to gene-gene and gene-environment interactions in chronic fatigue syndrome. *Pharmacogenomics* 10, 35–42. doi: 10.2217/14622416.10.1.35

Lin, E., and Hwang, Y. (2008b). A support vector machine approach to assess drug efficacy of interferon-alpha and ribavirin combination therapy. *Mol. Diagn. Ther.* 12, 219–223. doi: 10.1007/bf03256287

Lin, E., Kuo, P.-H., Liu, Y.-L., Yang, A., and Tsai, S.-J. (2019). Association and interaction effects of interleukin-12 related genes and physical activity on cognitive aging in old adults in the Taiwanese population. *Front. Neurol.* 10:1065. doi: 10.3389/fneur.2019.01065

Lin, E., Kuo, P.-H., Liu, Y.-L., Yu, Y. W.-Y., Yang, A. C., and Tsai, S.-J. (2018a). A deep learning approach for predicting antidepressant response in major depression using clinical and genetic biomarkers. *Front. Psychiatry* 9:290. doi: 10.3389/fpsyt.2018.00290

Lin, E., and Lane, H. Y. (2015). Genome-wide association studies in pharmacogenomics of antidepressants. *Pharmacogenomics* 16, 555–566. doi: 10.2217/pgs.15.5

Lin, E., and Lane, H.-Y. (2017). Machine learning and systems genomics approaches for multi-omics data. *Biomarker Res.* 5:2. doi: 10.1186/s40364-017-0082-y

Lin, E., Lin, C.-H., Lai, Y.-L., Huang, C.-H., Huang, Y.-J., and Lane, H.-Y. (2018b). Combination of G72 genetic variation and G72 protein level to detect schizophrenia: machine learning approaches. *Front. Psychiatry* 9:566. doi: 10.3389/fpsyt.2018.00566

Linden, A. (2006). Measuring diagnostic and predictive accuracy in disease management: an introduction to receiver operating characteristic (ROC) analysis. *J. Eval. Clin. Pract.* 12, 132–139. doi: 10.1111/j.1365-2753.2005.00598.x

Moghaddam, B., and Javitt, D. (2012). From revolution to evolution: the glutamate hypothesis of schizophrenia and its implication for treatment. *Neuropsychopharmacology* 37, 4–15. doi: 10.1038/npp.2011.181

Morera-Fumero, A. L., and Abreu-Gonzalez, P. (2013). Role of melatonin in schizophrenia. *Int. J. Mol. Sci.* 14, 9037–9050. doi: 10.3390/ijms14059037

Rumelhart, D. E., Hinton, G. E., and William, R. J. (1996). "Learning internal representation by error propagation," in *Parallel Distributed Processing: Explorations in the Micro-Structure of Cognition*, eds D. E. Rumelhart and J. L. McClelland (Cambridge, MA: MIT Press).

Sacchi, S., Bernasconi, M., Martineau, M., Mothet, J.-P., Ruzzene, M., Pilone, M. S., et al. (2008). pLG72 modulates intracellular D-serine levels through its interaction with D-amino acid oxidase EFFECT ON SCHIZOPHRENIA SUSCEPTIBILITY. *J. Biol. Chem.* 283, 22244–22256. doi: 10.1074/jbc.M709153200

Snyderman, R. (2012). Personalized health care: from theory to practice. *Biotechnol. J.* 7, 973–979. doi: 10.1002/biot.201100297

Üstün, B., Melssen, W. J., and Buydens, L. M. (2006). Facilitating the application of support vector regression by using a universal Pearson VII function based kernel. *Chemometr. Intell. Lab. Syst.* 81, 29–40. doi: 10.1016/j.chemolab.2005.09.003

Vapnik, V. (2013). *The Nature of Statistical Learning Theory*. Berlin: Springer science & business media.

Witten, I. H., Frank, E., and Hall, M. (2005). *Data Mining: Practical Machine Learning Tools and Techniques*. Francisco, CA: Morgan Kaufmann Publishers.

Wulff, K., Dijk, D.-J., Middleton, B., Foster, R. G., and Joyce, E. M. (2012). Sleep and circadian rhythm disruption in schizophrenia. *Br. J. Psychiatry* 200, 308–316. doi: 10.1192/bjp.bp.111.096321

Yang, P., Hwa Yang, Y., Zhou, B., and Zomaya, A. (2010). A review of ensemble methods in bioinformatics. *Curr. Bioinform.* 5, 296–308. doi: 10.2174/157489310794072508

Zhang, L., Yu, G., Xia, D., and Wang, J. (2019). Protein–protein interactions prediction based on ensemble deep neural networks. *Neurocomputing* 324, 10–19. doi: 10.1016/j.neucom.2018.02.097

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.

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: info@frontiersin.org | +41 21 510 17 00

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