Network-Based Approaches towards Understanding Genetics of Metabolism and Regulatory Interactions

Le Chang

Department of Human Genetics

Faculty of Medicine and Health Sciences

McGill University, Montreal

April 2023

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Doctor of Philosophy

Table of Contents

Abstract

The growing availability of large-scale genetic and phenotypic data has facilitated substantial endeavors aimed at understanding the genetic architecture of human diseases and complex traits. However, the mapping of genetic variation onto phenotypes presents several challenges. Integrating multiple types of omics data from diverse sources to reveal functional insights represents a daunting task. Furthermore, the intricate interplay of multiple functional elements within regulatory networks complicates the understanding of the consequences of dysregulation of specific regulators. Additionally, establishing a causal link between risk factors and disease remains a considerable difficulty. This thesis aims to address these challenges by developing bioinformatics applications to support data analytics and hypothesis generation.

First, I developed mGWAS-Explorer, a web-based platform that links SNPs, genes, metabolites, and diseases for functional insights. The tool contains a comprehensive collection of up-to-date significant mGWAS summary statistics with deep annotation on metabolite quantitative trait loci (mQTLs) and advanced network visual analytics support. By integrating multiple knowledgebases, mGWAS-Explorer is able to build SNP-based, gene-based, and metabolite-based networks to facilitate mechanistic insights. The application of the mGWAS-Explorer was demonstrated using COVID-19 and type 2 diabetes case studies.

Next, I developed version 2.0 of miRNet, a web-based platform for miRNA-centric network visual analytics. It integrates data from over 14 different miRNA databases and supports intuitive network analysis and functional enrichment analysis. It supports various inputs and statistics, including a list of miRNAs, SNPs, genes, transcription factors, small molecules, ncRNAs, diseases, epigenetic factors, any combinations, or a data table from microarray, RNAseq, or RT-qPCR experiments.

Lastly, I developed version 2.0 of mGWAS-Explorer, which focuses on inferring causal relationships between metabolites and diseases by leveraging mGWAS summary statistics and two-sample Mendelian randomization approach. It was also able to integrate additional molecular QTLs to facilitate understanding the mechanisms of how genetic variants influence the phenotypes at different omics levels and visualize the results in a network view. Additionally, mGWAS 2.0 allows reproducible and scalable analysis through an underlying R package.

Overall, the web-based platforms described in this dissertation enable the interpretation and functional analysis of omics data and pave the way toward a better understanding of the genetics of metabolism and regulatory relationships.

Résumé

La disponibilité croissante de données génétiques et phénotypiques à grande échelle a facilité les efforts substantiels visant à comprendre l'architecture génétique des maladies humaines et des traits complexes. Cependant, la cartographie de la variation génétique sur les phénotypes présente plusieurs défis. Une proportion importante, environ 90 %, des variants identifiés par les études d'association pangénomique (GWAS) sont situés dans des régions non codantes du génome, ce qui rend difficile la détermination de leur impact fonctionnel sur le phénotype. En outre, l'interaction complexe de multiples éléments fonctionnels au sein des réseaux de régulation complique la compréhension des conséquences de la dérégulation de régulateurs spécifiques. Enfin, l'établissement d'un lien de causalité entre les facteurs de risque et la maladie reste une difficulté considérable. Cette thèse vise à relever ces défis en développant des applications bioinformatiques pour soutenir l'analyse des données et la génération d'hypothèses.

Tout d'abord, j'ai développé mGWAS-Explorer, une plateforme basée sur le web qui relie les SNP, les gènes, les métabolites et les maladies pour obtenir des informations fonctionnelles. L'outil contient une collection complète de statistiques sommaires mGWAS significatives et actualisées, avec une annotation approfondie sur les loci de traits quantitatifs (mQTL) des métabolites et un support avancé d'analyse visuelle des réseaux. En intégrant plusieurs bases de connaissances, mGWAS-Explorer est capable de construire des réseaux basés sur les SNP, les gènes et les métabolites pour faciliter la compréhension des mécanismes. L'application de mGWAS-Explorer a été démontrée à l'aide des études de cas COVID-19 et diabète de type 2.

Ensuite, j'ai développé la version 2.0 de miRNet, une plateforme en ligne pour l'analyse visuelle de réseaux centrés sur les miARN. Elle intègre les données de plus de 14 bases de données miRNA différentes et permet une analyse intuitive des réseaux et une analyse de l'enrichissement

fonctionnel. Elle prend en charge différentes entrées et statistiques, y compris une liste de miRNA, SNP, gènes, facteurs de transcription, petites molécules, ncRNA, maladies, facteurs épigénétiques, toute combinaison, ou un tableau de données provenant d'expériences de microarray, RNAseq, ou RT-qPCR.

Enfin, j'ai développé la version 2.0 de mGWAS-Explorer, qui se concentre sur l'inférence des relations causales entre les métabolites et les maladies en exploitant les statistiques sommaires de mGWAS et l'approche de randomisation mendélienne à deux échantillons. Il a également été en mesure d'intégrer des QTL moléculaires supplémentaires pour faciliter la compréhension des mécanismes d'influence des variants génétiques sur les phénotypes à différents niveaux omiques et de visualiser les résultats dans une vue en réseau. En outre, mGWAS 2.0 permet une analyse reproductible et évolutive grâce à un progiciel R sous-jacent.

Dans l'ensemble, les plateformes web décrites dans cette thèse permettent l'interprétation et l'analyse fonctionnelle des données omiques et ouvrent la voie à une meilleure compréhension de la génétique du métabolisme et des relations de régulation.

List of abbreviations

1000G	1000 genomes project
2SMR	two-sample Mendelian randomization
AA	arachidonic acid
ABO	alpha 1-3-n-acetylgalactosaminyltransferase and alpha 1-3-galactosyltransferase
AI	artificial intelligence
API	application programming interface
BMI	body mass index
CAD	coronary artery disease
CD	Crohn's diseases
CHD	coronary heart disease
ChIP-Seq	chromatin immunoprecipitation sequencing
circRNA	circular RNA
COVID-19	coronavirus disease 2019
CSF	cerebrospinal fluid
DOM	document object model
eaQTLs	enhancer activity quantitative trait locus
ENCODE	encyclopedia of DNA elements
EnrichNet	network-based enrichment analysis
eQTL	expression quantitative trait locus
FAQs	frequently asked questions
FFL	feed-forward loop

FUT2	fucosyltransferase 2
GIM	genetically influenced metabotype
GLP2R	glucagon like peptide 2 receptor
GO	gene ontology
GRCh37	genome reference consortium human build 37
GTEx	genotype-tissue expression project
GWAS	genome-wide association studies
HMDB	human metabolome database
hQTL	histone modification quantitative trait locus
IV	instrumental variable
JSF	Javaserver Faces
KEGG	Kyoto encyclopedia of genes and genomes
LD	linkage disequilibrium
LDL-C	low-density lipoprotein cholesterol
LDSC	linkage disequilibrium score regression
lncRNA	long noncoding RNA
meQTL	methylation quantitative trait locus
mGWAS	metabolome genome-wide association study
miRNA	microRNA
molQTLs	molecular quantitative trait locus
mQTL	metabolite quantitative trait locus
MR	Mendelian randomization

MS	mass spectrometry
ncRNA	noncoding RNA
NetGSA	network-based gene set analysis
NMR	nuclear magnetic resonance
ORA	over representation analysis
PCSF	prize-collecting steiner forest
PPI	protein-protein interaction
pQTL	protein quantitative trait locus
puQTL	promoter usage quantitative trait locus
RCT	randomized controlled trial
RDBMS	relational database management system
RNA	ribonucleic acid
RT-qPCR	quantitative reverse transcription polymerase chain reaction
RWR	random walk with restart
SE	standard error
SemMedDB	the semantic MEDLINE database
SNiPA	an interactive, genetic variant-centered annotation browser
SNP	single-nucleotide polymorphism
SPIA	signaling pathway impact analysis
T2D	type 2 diabetes
TCDB	transporter classification database
TF	transcription factor

UI	user interface
VEP	variant effect predictor

List of figures

Chapter 1

Chapter 2

Chapter 3

Chapter 4

List of tables

Chapter 2

Chapter 3

Chapter 4

Acknowledgements

First of all, I would like to express my sincere gratitude to my supervisor Prof. Jianguo (Jeff) Xia for your mentorship, valuable advice, and encouragements in my entire PhD journey. I am inspired by your passion for science. I am grateful for your thoughtfulness and concrete guidance when I had challenges. Thank you for spending long hours reviewing my code and papers. Thank you for supporting my growth as a scientist, giving me the opportunity to apply for awards and present at conferences. I am certain all of these experiences will be invaluable to me long after my graduation.

I am greatly thankful to my supervisory committee members Profs. Jacek Majewski and Sahir Bhatnagar for your insightful suggestions and discussions during meetings. I would also like to express my appreciation to my thesis examiners and oral defense committee, Profs. Celia Greenwood, Ewy Mathé, Yu (Brandon) Xia, and Danielle Malo. Your thought-provoking questions and comprehensive evaluation have made the defense a truly enriching and rewarding experience.

I am greatly thankful to my lab members: Janice Ou, Yannan Fan, Yao Lu, Farial Karimian, Zhiqiang Pang, Tanisha Shiri, Dr. Dana Praslickova, Dr. Charles Viau, Dr. Orcun Hacariz, Dr. Jessica Ewald, Dr. Xue Gu, Zainab Bello, Tim Yang, Tom Yang, Joseph O'Brien, Dr. Abhishikha Sharma, Dr. Estevan Bruginski, Dr. Mai Yamamoto and Dr. Lei Xu. Special thanks to Drs. Othman Soufan, Peng Liu, Guangyan Zhou, and Jasmine Chong for teaching me coding skills and patiently helping me debugging when I was struggling at the beginning of my PhD.

I would like to acknowledge the financial support from the NSERC-CREATE MATRIX in the past four years. I am thankful to the International Society for Computational Biology, Metabolomics Association of North America, and the Department of Human Genetics for their

travel support to attend conferences. I am thankful to the administrative staff and professors from the department and the local wellness advisors from the student wellness hub. I am also thankful to the reviewers and editors for their constructive feedback on my manuscripts, as well as the bioinformatics, human genetics, and metabolomics research community to their dedications to open science.

I am thankful to myself. Thanks for my resilience and effort during a challenging time.

Finally, I would like to thank all my friends and family for supporting me over the years. To all my friends, thank you for your companionship. Mom and Dad, thank you for encouraging me despite half a globe away. I would not make it without your love and support.

Contribution to original knowledge

From the field of human genetics to metabolomics, I have made the following novel contributions to research:

1. I developed mGWAS-Explorer, a web-based platform linking SNPs, genes, metabolites, and diseases for functional insights. The application was demonstrated using a COVID-19 and a type 2 diabetes case studies. As described in Chapter 2, its key features include:
    - Comprehensive collection with deep annotation of the SNP-metabolite associations based on 65 mGWAS publications.
    - Support for SNP-based, gene-based, and metabolite-based network generation to facilitate interpreting results.
    - Powerful network visual analytics system facilitating interactive exploration and built-in topological and functional enrichment analysis.
2. I enhanced miRNet, a web-based visual analytics platforms for miRNA-centric gene regulatory network analysis. The utility of the tool was illustrated through analyzing TF-miRNA co-regulatory networks in a multiple sclerosis case study. As described in Chapter 3, its key features include:
    - Integrative knowledgebase for understanding miRNA regulatory networks and functions, which contains data from 14 different miRNA databases.
    - Added support for understanding TF-miRNA coregulatory networks and SNPs related to miRNA-binding sites, miRNA processing or target genes.
    - Implemented novel visual analytical functions to support creation and exploration of multipartite networks

3. I developed version 2.0 of mGWAS-Explorer. As described in Chapter 4, its key features include:
    - Implemented two-sample Mendelian randomization strategy to enable the exploration of causal relationships between metabolites and diseases.
    - Support for triangulated evidence from different sources, including statistical associations, biochemical pathways, and causal relationships to uncover mechanistic insights.
    - Improved support for reproducible research with the release of the underlying mGWASR package

Format of the thesis

This thesis follows the manuscript-based format and comprises of three original scholarly manuscripts. Chapters 2 and 3 have been published in peer-reviewed journals. Chapter 4 is a manuscript in preparation. Le Chang is the first author on all three manuscripts.

Published manuscripts:

Chapter 2: Chang, Le et al. "mGWAS-Explorer: Linking SNPs, Genes, Metabolites, and Diseases for Functional Insights." *Metabolites* vol. 12,6 526. 7 Jun. 2022, doi:10.3390/metabo12060526

Chapter 3: Chang, Le et al. "miRNet 2.0: network-based visual analytics for miRNA functional analysis and systems biology." *Nucleic acids research* vol. 48,W1 (2020): W244-W251. doi:10.1093/nar/gkaa467

Manuscript in preparation:

Chapter 4: Chang, Le et al. "mGWAS-Explorer 2.0: a web-based platform for prioritizing metabolites with causal impact on disease phenotypes." (2023)

Contribution of authors

Chapter 2 is a manuscript authored by Le Chang, Guangyan Zhou, Huiting Ou, and Jianguo Xia. It was published in Metabolites in June 2022. Conceptualization, J.X.; methodology, L.C. and J.X.; software, L.C., G.Z. and J.X.; data curation, L.C. and H.O.; writing—original draft preparation, L.C.; writing—review and editing, J.X. and G.Z.; supervision, J.X.; funding acquisition J.X. All authors have read and agreed to the published version of the manuscript.

Chapter 3 is a manuscript authored by Le Chang, Guangyan Zhou, Othman Soufan and Jianguo Xia. It was published in Nucleic Acids Research in June 2020. Conceptualization, J.X.; methodology, L.C. and J.X.; software, L.C., G.Z., O.S. and J.X.; data curation, L.C. and G.Z.; writing—original draft preparation, L.C.; writing—review and editing, J.X., O.S. and G.Z.; supervision, J.X.; funding acquisition J.X. All authors have read and agreed to the published version of the manuscript.

Chapter 4 is a manuscript authored by Le Chang, Guangyan Zhou and Jianguo Xia. Conceptualization, J.X.; methodology, L.C. and J.X.; software, L.C., G.Z., and J.X.; data curation, L.C.; writing—original draft preparation, L.C.; writing—review and editing, J.X., and G.Z.; supervision, J.X.; funding acquisition J.X. All authors have read and agreed to the published version of the manuscript.

Chapter 1: General introduction

The foremost and ongoing question in the field of human genetics pertains to the elucidation of the relationships between genotype and phenotype. With the advent of technological advancement in the generation of data from various levels of biological systems, many experimental and analytical approaches have been developed to define functional elements in the human genome. These include evolutional, biochemical and genetics approaches (1). The evolutionary approach assesses the degree of selective pressure, the biochemical approach assesses molecular activity, while the genetic approach assesses the consequence of alterations on the phenotype (e.g., GWAS). Each of the three methods can provide a wealth of information on the biological significance of a genetic segment. However, there are several challenges in regard to dissecting how genetic variation map into phenotypes. (i) 90% of the variants identified by GWAS locate at the noncoding regions of the genome and their functional consequence on the phenotypes cannot be easily interpreted. (ii) The interaction of multiple functional elements in regulatory networks make it difficult to understand the effect of dysfunction of particular regulators. (iii) Causal inference between the risk factor and disease remains a major hurdle. Therefore, this thesis focuses on developing bioinformatics tools to address these challenges.

INTERMEDIATE MOLECULAR QTLs

The recent advancement in high-throughput technologies has facilitated the generation of multiple types of omics data (Figure 1), thereby enabling the identification and analysis of genetic variants associated with these intermediate molecular phenotypes (molQTLs). Through this, we are able to gain a better understanding of the functional impacts of these genetic variations on biological processes and regulatory networks (1-4). For instance, RNA sequencing quantifies gene

expression, as well as noncoding RNAs and other forms of post-transcriptional regulation (5). Bisulfite sequencing, on the other hand, allows for the profiling of DNA methylation (6). Chromatin immunoprecipitation sequencing (ChIP-Seq) can be utilized to investigate histone modifications, such as H3K4me1 and H3K27ac, as well as the regulatory elements, including transcription factor binding sites and RNA polymerase II-binding sites (7). Furthermore, 16s rRNA and shotgun metagenomics allow the characterization of the microbiome (8), while mass spectrometry is the key technology in profiling proteins and metabolites (9). These advancements have opened up the possibility of identifying a wide range of quantitative trait loci (QTLs), including expression QTLs, epigenetic QTLs, proteomic QTLs, metabolic QTLs, and microbiome QTLs.

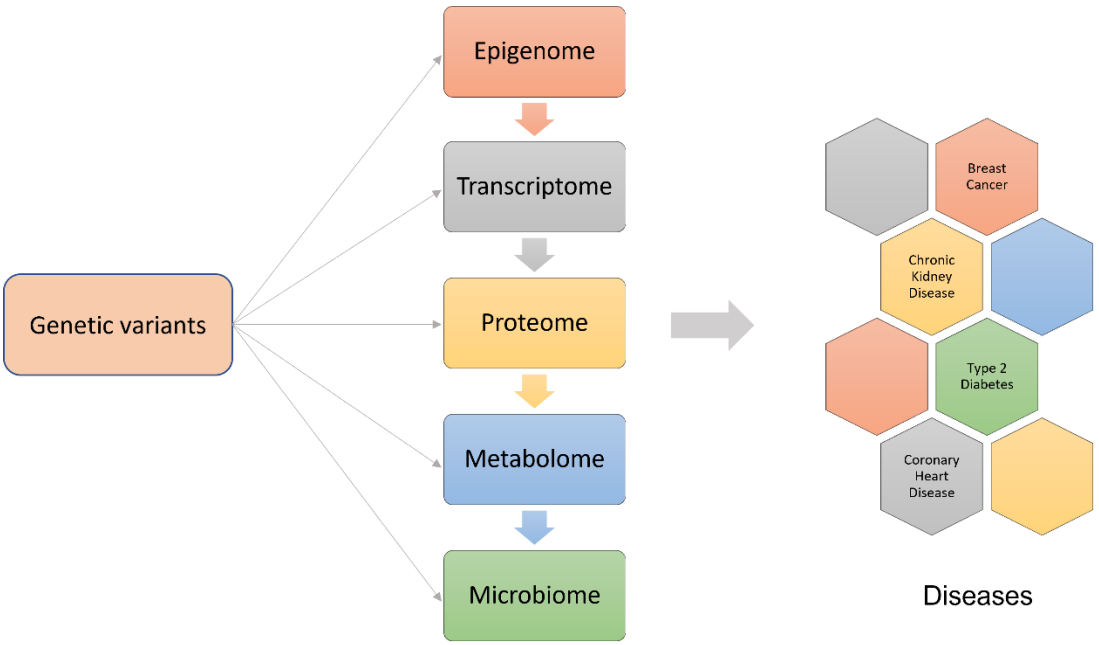

Figure 1. Overview of multi-omics integration in human diseases. The impact of genetic variations from multi-omics molecular quantitative traits (molQTLs) to diseases outcome through various layers of information such as the epigenome, transcriptome, proteome, metabolome, and microbiome.

Expression QTL

eQTL analysis is a method that studies the relationship between a genetic variant and changes in gene expression. This approach has been applied to a wide range of research fields, including the interpretation of GWAS and the fundamental processes of gene regulation (10,11). Studies of eQTL analysis in cell lines or tissues have revealed the existence of shared and context-specific eQTL mapping across a range of cells and tissues (12,13). Additionally, eQTL analysis has facilitated the identification of disease risk loci and genes that may play a role in the modulation of pathogenesis in a variety of diseases, including epilepsy (14), schizophrenia (15), and type 1 diabetes (16). eQTLs can be divided into two subcategories: *cis*-eQTLs and *trans*-eQTLs. *Cis*-eQTLs are genetic variants that are located within 1 megabase of the gene whose expression they affect, whereas trans-eQTLs are those variants that are situated either farther away or on a different chromosome from the gene of interest. It is noted that *cis*-eQTLs and *trans*-eQTLs exhibit distinct characteristics. Studies have shown that the effects of *trans*-eQTLs on gene expression are weaker when compared to those of *cis*-eQTLs (17), and that they are also less replicable across different studies (18).

eQTL in Noncoding RNAs

eQTLs in noncoding RNA (ncRNA) is an emerging area of research that aims to understand how genetic variants affect the expression levels of various types of ncRNA, such as long ncRNA (lncRNA), microRNA (miRNA), and circular RNA (circRNA). lncRNA-eQTL (lncR-eQTL) refers to the effect of genetic variants on lncRNA expression. Thus far, over 2,000 lncR-eQTLs have been identified in four primary tissues, and they have been linked to patient survival and known genetic loci associated with disease (19). miRNA eQTLs (miR-eQTLs) study

the effect of genetic variants on miRNA transcription, maturation and targeting (20), with over 5,200 miR-eQTLs identified in whole blood (21) and more than 90,000 *cis*- and *trans*-miR-eQTLs across 33 cancer types (19). This high number of miR-eQTLs indicates that miRNAs have a significant regulatory impact on gene expression in the context of cancer and may contribute to tumor heterogeneity and cancer progression. Circular RNA-eQTL (circ-eQTL) studies the effect of genetic variants on circRNA expression. Most of the circ-eQTLs were found to be located near the back-splicing sites, suggesting their role in circRNA regulation (22). The increasing knowledge of circRNAs provides opportunities to understand the connection between genetic variations and circRNAs in complex traits and diseases (23-26). The different orders of magnitude observed among the noncoding RNAs could be due to the differences in the focus of the studies, the roles these noncoding RNA molecules play in gene regulation, and the complexity of the molecular mechanisms involved in different biological contexts, such as cancer and whole blood samples. These numbers highlight the importance of further exploring noncoding RNA regulatory mechanisms to better understand their roles in gene expression, disease risk, and phenotypic variation.

Epigenetic QTL

Genetic variations play a crucial role in regulating gene expression through epigenetic marks, such as DNA methylation (27), histone modification (28), and regulatory elements (e.g., promoters and enhancers) (29). (i) DNA methylation QTLs (meQTLs) have been identified in several tissues and organs, including immune cells (30) and brain (31), and have been found to play a functional role in negatively regulating gene expression (30). meQTLs have been associated with human diseases such as post-traumatic stress disorder (32). (ii) Histone modification is

another important epigenetic marker that helps in gene regulation through chromosomal packaging. Different types of histone modification, such as H3K4me3, H3K9ac, and H3K27me3, are associated with different regulatory elements in the human genome (28). The genetic variants associated with histone modifications are defined as histone modification QTLs (hQTLs). (iii) Regulatory elements, such as promoters and enhancers, play a key role in regulating gene expression. The dynamic usage of regulatory elements is defined as promoter usage QTLs (puQTLs) and enhancer activity QTLs (eaQTLs) (29), which are associated with phenotypic traits and diseases such as Alzheimer's disease (33).

Proteomic QTL

As fundamental building blocks of life, proteins occupy a crucial position within the biology of organisms, fulfilling various critical functions including enzymatic activity, receptor and transport processes. The genetic variations that influence the expression of these proteins are known as protein quantitative trait loci (pQTLs) (34). These pQTLs have facilitated the identification of shared etiological mechanisms across diseases and have allowed for a prioritization of therapeutic targets. For instance, the FBLN3 protein, which is an extracellular matrix glycoprotein encoded by the *EFEMP1* gene, has been identified as a target in a significant number of diseases and phenotypic conditions, such as whole body fat bass and carpal tunnel syndrome (35). Furthermore, leveraging pQTL in Mendelian randomization analysis, a method for evaluating causality between protein levels and disease risks, holds great promise in the validation of drug targets (36).

Metabolic QTL

The human metabolome is the comprehensive collection of the low-molecular-weight compounds, or "metabolites", found in blood, urine, saliva, or other biofluid and tissues, which reflect the joint effects of genetic and environmental factors (37,38). Genetic variants associated with variations in metabolite levels are referred to as metabolic QTLs (mQTLs) (4). In 2008, first metabolome genome-wide association study (mGWAS) analyzed 363 metabolites in serum of 284 individuals from the KORA study using a targeted approach (39). The authors found associations of SNPs with significant variations in the metabolite concentrations, accounting up to 12% of the observed variance. They were able to determine the presence of four variants in genes responsible for coding enzymes *FADS1*, *LIPC*, *SCAD*, and *MCAD*. These variants were found to exhibit a direct correlation with the corresponding metabolic phenotype, effectively aligning with the established biochemical pathways in which the relevant enzymes react. Furthermore, the study conducted by Suhre et al. provided valuable insights into the intricacies of transport mechanisms in a high-dimensional setting (40). Through the application of a nontargeted mass spectrometry-based approach, the authors quantified hundreds of blood metabolites and identified numerous genetic variants with associations to various metabolic traits. Of particular significance was the observation of genetic variants in the gene *SLC16A9* that were associated with free carnitine concentrations. The results of this study were further validated by experimental confirmation of the role of *SLC16A9*, also known as monocarboxylate transporter 9, as a carnitine efflux transporter using the *Xenopus oocyte* system. Recently, mQTLs have gained increasing attention as a key factor in shaping individual differences in metabolism and health (41-45). The mQTLs are associated with various phenotypic traits, including body mass index (BMI) (46), and have been implicated in the development of several human diseases, including kidney diseases (47) and cardiometabolic diseases (48).

Microbiome QTL

The microbiome has been demonstrated to play a pivotal role in a multitude of host tissues and organs, including the gastrointestinal tract, skin, and respiratory systems (49). Genetic variants associated with microbiome are known as microbiome QTL. A number of studies have aimed to identify these microbiome QTLs in a variety of human tissues, such as gut biopsies (50), feces (51), and skin (52). Additionally, studies have revealed the enrichment of microbiome QTLs in several diseases, including inflammatory bowel disease (53), as well as meningitis and gastrointestinal adenocarcinoma (54). These microbiome QTLs are mapped to genes involved in pathways related to immunity and food metabolism (54) . For example, SNP located in the LCT locus (rs4988235) is associated with the *Bifidobacterium* genus, showing evidence of a gene-diet relationship in the control of *Bifidobacterium* abundance (49).

NETWORK-BASED APPROACHES

Integrating multi-omics in a network biology context holds the promise for advancing our understanding of the complex relationships between the genotype and phenotype (55). The integration of diverse data sources through network analysis can play a crucial role in the identification of functional elements that participate in a multitude of interactions and offer a more extensive comprehension of their cellular functions (56,57). This holistic methodology offers a framework to integrate knowledge, construct meaningful models and extract biological insights at a system level.

The objective of integrating knowledge within network systems is to attain a deeper comprehension of the molecular context that underlies omics datasets. This aim can be realized

through the implementation of these three phases: (i) subjecting individual omics data to data processing and comparative statistical analysis in order to obtain features of significance; (ii) mapping the identified features with the existing knowledge, as represented by the available molecular interaction data, such as protein-protein interactions (PPIs), metabolic pathways, and gene regulatory networks (Figure 2); (iii) visualizing and examining the resulting subnetwork. The resulting subnetwork constitutes a basis for more in-depth subsequent analysis, including the detection of modules (i.e., subnetworks) and functional enrichment analysis. (58,59).

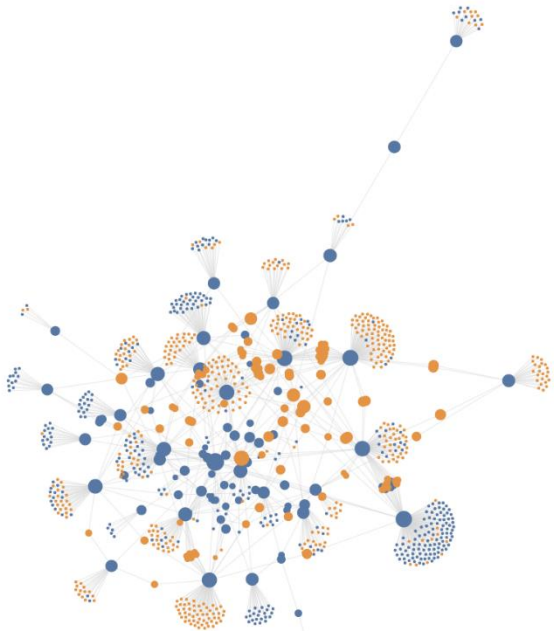

Figure 2. An example of an integrated regulatory network, composed of miRNA-gene interactions and protein-protein interactions, where blue nodes depict the genes/proteins and orange nodes represent miRNAs. This figure is generated with OmicsNet 2.0 web application (60).

Network topology

Interactions between biological components, such as metabolites or genes, can be represented as nodes in a network. These nodes are connected by edges, which depict the

relationships between the nodes. The directionality of the network is determined by the characteristics of the biological data, with undirected networks commonly used for protein-protein or genetic interactions, while directed networks are utilized for transcription factor binding, miRNA targeting, phosphorylation, and metabolic networks (61,62).

Topology plays a critical role in understanding network architecture and function. Some of the most widely used topological features include degree, shortest path length, and betweenness (57). The degree of a node refers to the number of connections it has. Nodes with a high degree are more highly connected within the network and may therefore play a more crucial role. Distance measures the shortest path length between two nodes, with the average distance and diameter of a network indicating the general proximity of nodes. Betweenness represents the number of times it functions as a bridge along the shortest path between two other nodes. In other words, it estimates the amount of traffic passing through a node (63).

Network modules

In recent years, much emphasis has been placed on the examination of the local structural components of biological networks, as opposed to only focusing on the characterization of their global topological structure. These local units, referred to as network modules, are defined as large subnetworks that consist of densely connected nodes (64). Several algorithms have been developed to identify potential network modules. For instance, Prize-Collecting Steiner Forest (PCSF) is an algorithm utilized for the identification of one or multiple subnetworks in a given undirected global network (65,66). The goal of PCSF is to maximize the prizes associated with the input nodes, minimize the costs associated with the edges, and reduce the number of subnetworks. PCSF has been applied to study risk genes for autism spectrum disorders (67) as well as metabolic alterations

in multiple sclerosis (68). Other algorithms include random walks and label propagation (69). Random walk method attempts to locate densely interconnected subnetworks, commonly referred to as communities, within a network through the utilization of random walks. The underlying concept is that random walks of short distance tend to remain within the same community (70). Conversely, label propagation algorithm is a fast approach for identifying community structures within networks. It operates by assigning unique labels to the nodes and then iteratively updating the labels through a majority voting process in the direct neighborhoods of the nodes (71).

Network motifs

The abundance of interaction data has enabled the identification of small, frequent patterns within large networks, referred to as network motifs (72). Each network motif is capable of executing well-defined information processing functions, as has been demonstrated through various experimental studies, particularly in the model organism *Escherichia coli* (73). One such motif, the feed-forward loop (FFL), involves regulation of target genes by a transcription factor (TF) and the regulation of the same genes by a miRNA (62). FFLs can be classified into two types: coherent and incoherent. In coherent FFLs, both direct and indirect regulation have the same effect on gene expression, either activating or repressing. In contrast, incoherent FFLs have opposing effects on gene expression. These two types of FFLs also exhibit distinct patterns of gene expression, with the coherent type showing mutual exclusion of miRNA and target gene expression, and the incoherent type exhibiting co-expression. The role of coherent FFLs may be to prevent the co-expression of miRNA and its targets, while the incoherent type may play a role in precisely modulating target gene expression and controlling biological noise (74). For instance, a recent study identified an incoherent FFL between miR-34a and Numb targeting Notch, which

regulates asymmetric division in early-stage colon cancer stem cells. Disruptions to this FFL resulted in the emergence of an intermediate cell population with plastic properties, highlighting the significance of FFLs in maintaining stem cell proliferation (75).

Network-based enrichment

Network-based enrichment methods leverage the information contained in biological pathways or molecular interaction networks through the use of graph-based statistics. The standard approach to these methods consists of two main stages: (i) mapping the experimental data onto the network and (ii) utilizing "topology-aware" statistics that incorporate structural information to compute pathway enrichment scores (58). A proximity measure is often utilized to establish a relationship between the user input and known biological pathways, with the underlying principle being that nodes with similar functions tend to be situated in close proximity within the network. By using this method, it is possible to identify enriched pathways whose related members do not exactly match the input provided by the user. Additionally, the connectivity pattern can be used in conjunction with conventional enrichment techniques as a complementary factor to improve interpretability and discriminative power in the context of networks (58). Some popular network-based enrichment methods include SPIA (76), NetGSA (77), and EnrichNet (78). For example, EnrichNet is a network-based enrichment analysis method, which utilizes the concept of proximity to quantify the relationship between input genes and reference gene sets (78). This method is comprised of several steps, including the application of the Random Walk with Restart (RWR) algorithm to calculate the distance of seed nodes to all reference gene sets, the conversion of node-level distance scores into distance score vectors for reference gene sets, the aggregation of individual distance vectors to form a background model distribution, and the calculation of

enrichment scores through a measurement of deviation from the background model average distribution.

Network visual analytics

A crucial aspect of network-based approach is the implementation of visualization techniques, which serve to enhance the comprehensibility of the network results (79). Visual analytics allows seamlessly integration of interactive visualization and human judgement within the processes of data analysis (80). Such an approach requires the combination of data analysis methods, data visualization and interactive techniques (e.g., web technologies) in the analytical-reasoning process for decision making. By incorporating this approach, decision makers are enabled to leverage their human creativity and domain knowledge to make sound decisions. The utilization of a visual analytics methodology is highly appropriate for the analysis of omics data due to several compelling factors:1) The integration of statistical analysis algorithms and computational models constitutes a crucial aspect of modern omics data analysis, as it leverages well-established methodologies to enhance the traditional visual assessment techniques. The synergy between these two approaches leads to a more comprehensive and sophisticated analysis, resulting in more informed decision-making. 2) The concept of interactivity plays a crucial role in promoting an iterative approach to data analysis. This iterative process enables the gradual understanding of data, whereby initial hypotheses and insights act as a foundation for more extensive investigations. The incorporation of interactivity in the data analysis process, therefore, fosters incremental data understanding. 3) The analytical process can be enriched through the integration of the domain expertise, intuitive understanding, and creative abilities of the users via the utilization of interactive visualizations. This thesis endeavors to demonstrate the applicability

of visual analytics by presenting two applications. Our contribution underscores the efficacy of visual analytics in the realm of linking genotype to phenotype.

CAUSAL INFERENCE

The inference of causality between risk factors and relevant phenotypes constitutes a major objective in biomedical observational research that holds significant implications for comprehending the etiology of pathological conditions (81). A major challenge in the investigation of causal relationships is confounding, where a variable causally impacts both the risk factor and the outcome. Randomized Controlled Trials (RCTs), frequently held as the gold standard of causality inference, have limitations inherent to their methodology and maybe sometimes impractical and unethical to conduct (82). The limitations of RCTs and challenges of causal inference have led to the development of methodologies to enhance causal inference in observational research. Among these methods, Mendelian randomization (MR) is particularly effective in controlling for confounders (83).

Mendelian randomization

The Mendelian randomization (MR) technique is a scientifically robust methodology that leverages genetic variants that are associated with a given exposure as instrumental variables to calculate the causal effect of the exposure on a specific outcome of interest (83,84). Single nucleotide polymorphisms (SNPs) are commonly used genetic instruments that can act as an anchor for estimating causality. By evaluating the difference in outcome between individuals carrying the risk allele (i.e., exposure) and those without (i.e., control), it is possible to assess the causal effect of a given exposure (85,86) (Figure 3a). Based on Mendel's laws of segregation and

independent assortment, Mendelian randomization (MR) can be considered a natural experiment similar to randomized controlled trials (RCTs) (83,87). By relying on random genetic allocation, MR generates variation in exposure that is not confounded, allowing for a more rigorous examination of causality (81).

Mendelian randomization depends on three key assumptions. Firstly, the instrumental variable must exhibit an association with the risk factor under examination (relevance assumption). Secondly, it must be independent of confounders that may impact the outcome (independence assumption). Lastly, it must affect the outcome only through the risk factor (exclusion restriction assumption) (Figure 3b). It is possible for a genetic variant to fulfill these assumptions if the biological process connecting the variant to the risk factor has been thoroughly understood. However, in numerous instances, Mendelian randomisation investigations includes multiple genetic variants, which can be used in analyses of sensitivity to assess the basic assumptions. Typically, the three conditions must be satisfied for each of the genetic variants utilized (88).

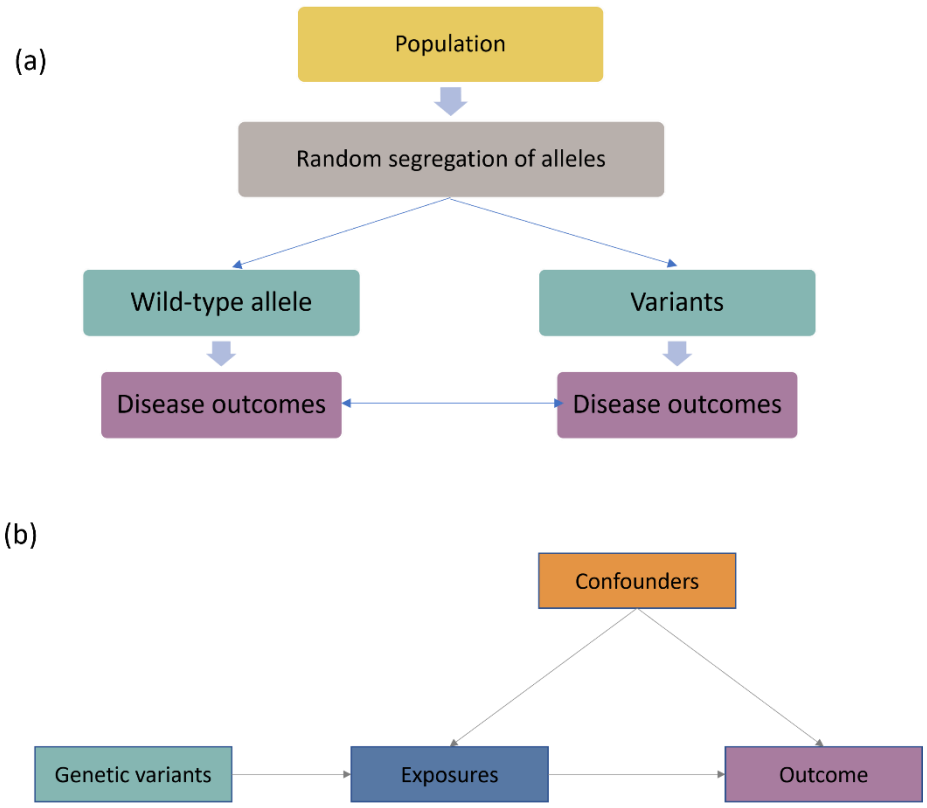

Figure 3. Overview of Mendelian randomization. (a) The design of a Mendelian randomization study. In Mendelian randomization, the random allocation of alleles during meiosis assigns one allele to represent "exposure" and the other to represent "control". The disease outcomes are compared between control group with the wild-type allele and the exposure group with the variant. "Exposure" refers to the presence of a specific variant that is hypothesized to influence a particular risk factor or characteristic. Individuals carrying this allele are considered "exposed" to the risk factor. "Control" refers to individuals who do not carry the specific variant. They represent the comparison group, which is not "exposed" to the potential influence of the genetic variant on the risk factor. (b) The three essential elements of instrumental variable assumptions are: the instrumental variable, which can be genetic variants, must have a relationship with the exposure; the instrumental variable must not be related to any confounders; and there must not be any direct path from the genetic variants to the disease outcome that bypasses the exposure of interest.

The utilization of results from genome-wide association studies (GWAS) in the context of Mendelian randomization (MR) has been recognized as a critical aspect of MR methodology. This strategy, known as two-sample MR (2SMR), was first described by Pierce and Burgess in 2013 (89). The two-sample MR approach allows for the estimation of causal influence between two traits through the use of summary data derived from separate studies, where the SNP-exposure effects and the SNP-outcome effects are obtained independently (90). The advantage of this method is that it enables causal inference to be made between two traits that may not be measured in the same set of samples, thereby leveraging the statistical strength of pre-existing large GWAS. The versatility of 2SMR has broad implications for MR applications, as it can be applied to thousands of potential exposure-outcome associations, including behavioral traits such as alcohol consumption, as well as intermediate molecular phenotypes like miRNAs (91) and metabolites (92).

Triangulation with literature-mined evidence

The Mendelian randomization technique, although effective in addressing the confounding and reverse causality problems that frequently arise in observational studies, nonetheless is not immune to the potential biases that may result from its statistical-based methodology. Hence, it is imperative to adopt a multi-faceted approach to the assessment of causality, through the utilization of what has been referred to as the "triangulation" framework. By combining the results of various approaches and verifying that they concur in their conclusions, the robustness and validity of the results can be enhanced and the confidence in the causal inferences drawn from the data strengthened (93,94). The integration of MR estimates and literature mined experimental results

constitutes a notable exemplar of the utilization of triangulation in the examination of causal relationships. The employment of searching enriched overlapping terms in semantic triples ('subject-predicate-object'), as a means of delineating the intermediate processes that connect risk factors and disease outcomes, holds the potential to significantly enhance our ability to identify the mechanisms that underlie the etiology of disease (95-97). The enrichment analysis is performed utilizing the conventional 2x2 Fisher's Exact Test, which compares the frequency of queries against a reference background. An overlap is considered to occur when the subject of a triple from the set of exposure queries overlaps with the object of the triple from the set of outcome inquiries. Therefore, the process of triangulation, which involves the integration of evidence derived from multiple sources, enables the enhancement of decision-making and the formulation of more robust hypotheses (98).

Rationale and objectives

After decades of reductionist approach, integrated omics analysis and network analysis have started investigating the molecular mechanism at a systems level. Various databases and algorithms have been developed to support such analysis. However, considering the complexity of these methods, the applications to biomedical research are limited. Therefore, user-friendly computational platforms need to be developed to better understand the multiple interactions to obtain a more systems-level understanding of the linking between genotype to phenotype. In particular, this thesis focuses on genetics of metabolism and regulatory interactions.

The objectives of my thesis project are threefold: (i) to enhance our understanding of the functional significance and causal effects underlying the genetic basis of metabolism and disease;

(ii) to improve functional profiling, network visual analytics, and reproducible analysis of miRNA-centric regulatory networks; (iii) to develop bioinformatics applications that can facilitate scalable, transparent, and reproducible research.

Chapter 2: mGWAS-Explorer: linking SNPs, genes, metabolites, and diseases for functional insights

[1]Le Chang, [2]Guangyan Zhou, [2]Huiting Ou and [1,2*]Jianguo Xia

[1]Department of Human Genetics, McGill University, Montreal, QC H3A 0C7, Canada;

[2]Institute of Parasitology, McGill University, Montreal, QC H9X 3V9, Canada

*To whom correspondence should be addressed:

Tel: 1-514-398-8668

Email: jeff.xia@mcgill.ca

Status: Manuscript published in Metabolites: https://doi.org/10.3390/metabo12060526

Abstract

Tens of thousands of single-nucleotide polymorphisms (SNPs) have been identified to be significantly associated with metabolite abundance in over 65 genome-wide association studies with metabolomics (mGWAS) to date. Obtaining mechanistic or functional insights from these associations for translational applications has become a key research area in the mGWAS community. Here, we introduce mGWAS-Explorer, a user-friendly web-based platform to help connect SNPs, metabolites, genes, and their known disease associations via powerful network visual analytics. The application of the mGWAS-Explorer was demonstrated using a COVID-19 and a type 2 diabetes case studies.

Introduction

Genome-wide association studies (GWAS) have identified hundreds of thousands of genetic loci associated with complex diseases. These associations have improved our understanding of the genetic architecture of human diseases [1]. However, translations of these associations into biomedical or pharmaceutical applications have been limited, as the majority of the disease-associated loci reside in the non-coding regions of the genome with no obvious gene targets [2]. Technology advancements in mass spectrometry (MS) and nuclear magnetic resonance (NMR) spectroscopy have allowed GWAS to be carried out with metabolomics (mGWAS) to study genetically influenced metabotypes (GIMs) [3,4]. mGWAS have been very successful in identifying metabolite quantitative trait loci (mQTLs). An mQTL is a locus that is associated with variations in metabolite abundance [3]. In addition to having larger effects compared to loci identified in GWAS of clinical phenotypes in general, many mQTLs can map to genes encoding enzymes or transporters, providing biochemical context for these variations [3,5]. Leveraging these mQTLs to improve our knowledge of metabolism and metabolic disorders for translational applications has become a key research area in the mGWAS community.

mQTLs are characterized by polygenicity and pleiotropy [6,7]. Polygenicity means a single trait is influenced by multiple genes, whereas pleiotropy refers to the phenomenon in which genetic variants affect multiple traits or diseases [8,9]. For instance, one single-nucleotide polymorphism (SNP) can directly affect multiple traits, or different SNPs in high linkage disequilibrium (LD) may exist for more than one trait. Pleiotropy may provide insights into the cause of trait comorbidity and help determine the direction of causal relationships by pointing to shared genetic mechanisms [8,10]. Various strategies have been developed to examine genetic relationships between multiple phenotypes [11,12,13,14,15,16,17,18]. For example, LD score regression is a

popular method to assess the genetic correlations of pairwise traits using GWAS summary statistics [13]. Colocalization is another strategy aiming to identify causal variants at two overlapping association signals [14]. These methods have successfully identified pleiotropic genomic regions and addressed fundamental research questions regarding the polygenicity of traits, but it is challenging to scale up these methods to study hundreds of traits at once.

Comprehensive annotations are necessary in order to gain functional insights into SNP–metabolite associations. Many resources are available to support SNP to gene annotation, such as VEP and SNiPA [19,20]. For metabolite annotation, there is a wealth of biochemical knowledge on enzymatic reactions as well as transporters and their substrates. In addition, mapping GWAS results to the protein–protein interaction (PPI) network can potentially augment the association signals [21].

Recently, cross-phenotype association analysis has gained increasing attention [22,23,24,25,26]. It takes a specific SNP and searches for associations across a range of molecular or disease phenotypes, which allows for elucidations of complex networks between phenotypes and their genetic loci. A variety of databases currently exist to store the genotype–phenotype association datasets, including GWAS Catalog [27], PhenoScanner [28], OpenGWAS [29], Open Targets Genetics [30], PheLiGe [31], DisGeNET [32], as well as specific tools for mGWAS, such as the metabolomics GWAS server [33,34]. Valuable tools currently exist to allow users to perform cross-phenotype analysis [35,36,37,38,39,40]. However, these tools do not offer extra support beyond displaying and visualization, and none of them are dedicated to mGWAS.

There is a clear demand for dedicated bioinformatics resources to support mGWAS data analysis and interpretation. Our overall assumption is that by developing a centralized place for mGWAS datasets and performing deep annotation of the underlying SNPs and metabolites, users

can gain valuable functional insights into the statistical associations identified from mGWAS results.

A network is a valuable approach to depict mGWAS results and allows the dissection of polygenicity and pleiotropy. Heterogeneous networks comprising various types of nodes (e.g., SNPs, genes, metabolites, and diseases) and edges (e.g., statistical or biochemical associations) have been remarkably useful in depicting the complex interplay across biological entities [41]. These network-based approaches have the potential to identify and prioritize therapeutic candidates to generate new hypotheses [42].

Here, we introduce mGWAS-Explorer (https://www.mgwas.ca (accessed on 1 May 2022), a user-friendly web-based platform for network-based integrative analysis and visual exploration of SNPs, genes, metabolites, and diseases. Its key features include:

- Comprehensive collection and deep annotation of SNP–metabolite associations based on data from the 65 mGWAS to date.
- Support for SNP-based, gene-based, and metabolite-based network generation to facilitate interpreting results.
- Powerful network visual analytics system facilitating interactive exploration and built-in topological and functional enrichment analysis.

mGWAS-Explorer also includes a comprehensive list of frequently asked questions (FAQs) and detailed tutorials. Together, these features comprise a powerful platform for functional interpretation and cross-phenotype association analysis of mGWAS datasets.

Results

Overview of the curated mGWAS datasets

Since the first study in 2008 [5], mGWAS with increasing sample sizes and various populations have been conducted, resulting in a continued increase in SNP–metabolite associations. We systematically curated the public mGWAS datasets to date. A summary table of these mGWAS datasets can be found in Table 1. Please note the p-value cutoffs are based on significance thresholds of the original studies, as the p-values and effect sizes of SNP-metabolite associations may differ across different studies due to the differences in sample sizes, population types, or the metabolomics platforms [4,7].

Table 1. A summary of the mGWAS datasets in mGWAS-Explorer.

Sample Type	Study #	* Metabolite #	** Metabolite Ratio #	SNP #	SNP–Metabolite Associations #
Blood	57	3992	1265	67,570	30,3090
Urine	5	271	1123	6877	9647
Saliva	1	14	0	1364	1454
Cerebrospinal fluid (CSF)	1	15	0	1178	1182
Mitochondria	1	0	390	194	404
Sum (unique)	65	4147	2388	73,737	313,720

* Metabolite number includes both targeted (compound names) and untargeted measures (feature IDs, such as '391.2859_3.774J based on mass to charge ratio and retention time. The total number of such feature IDs is 2464).

** Metabolite ratios (metabolite A/metabolite B) can be useful as they may reflect the biochemical conversion of metabolites and thus enhance the association signals. The # sign indicates size or total number [43].

Overview of the mGWAS-Explorer

The main workflow of mGWAS-Explorer is summarized in Figure 1. There are three major steps—data input, network creation, and network visual analytics. To begin, users can enter

through one of the five modules based on input type. The ‘SNP’ module allows users to explore SNP–gene, SNP–metabolite, or SNP–disease networks. We provide support for LD proxy search to maximize the search by looking for SNPs in LDs with the input SNPs. After SNP to gene mapping, users can choose to include PPIs in the networks. The ‘Gene’ module maps genes to SNPs that are significant in the mGWAS datasets, or to metabolites (i.e., through encoding enzymes or transporters), or known associated diseases. The ‘Metabolite’ module maps metabolites to associated SNPs, genes, or diseases. The ‘Search’ module allows users to search known SNP–gene associations in mGWAS datasets, while the ‘Browse’ module allows users to browse individual mGWAS data in a 3D Manhattan plot or a network view. To start the analysis, users must click a circular button from the mGWAS-Explorer homepage to enter the corresponding data upload page. Various functions are available to allow users to refine the networks. In the last step, the results are shown as interactive networks for visual exploration. Users can easily search, explore, highlight, or perform functional enrichment analysis on the nodes of interest. For instance, double-clicking an edge will display the evidence supporting the relationships. The network results can be downloaded in PNG, SVG format, or as graph files.

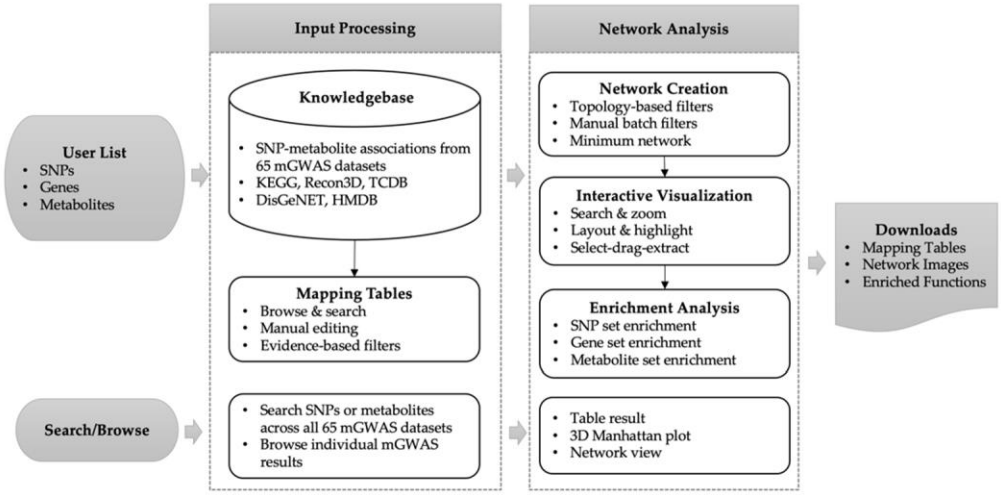

Figure 1. Overview of mGWAS-Explorer workflow. Users can upload different data types. The input will be mapped to the underlying knowledgebases to create mapping tables and networks. The visualization page allows users to intuitively explore the networks to identify important associations as well as to perform topology or functional analysis.

Analysis workflow

There are five modules in mGWAS-Explorer corresponding to the five different types of input. Users can upload a list of SNPs, metabolites, or genes (Figure 2a–c); browse individual mGWAS dataset in a 3D Manhattan plot (Figure 2d); or search significant SNP-metabolite associations across all mGWAS datasets (Figure 2e).

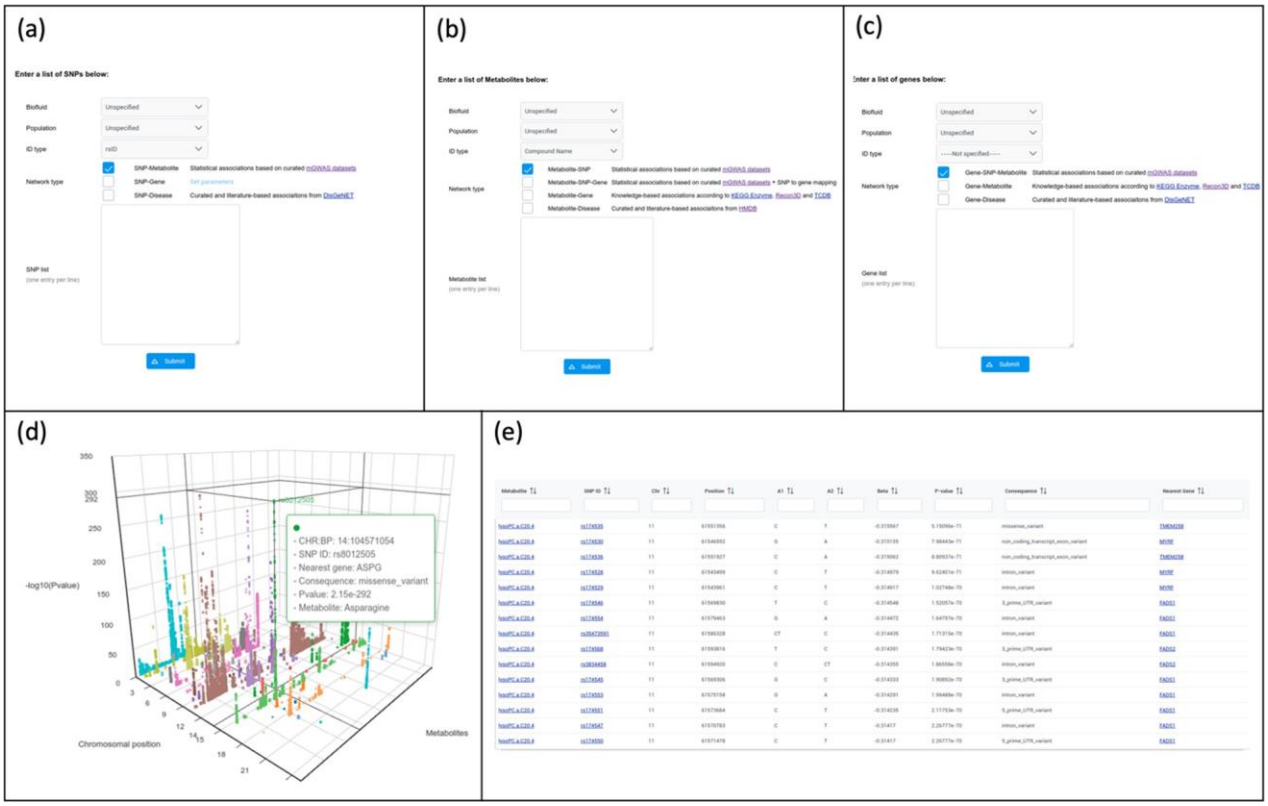

Figure 2. Screenshots of upload pages for SNP (a), metabolite (b), and gene (c) modules. (d) A screenshot of a 3D Manhattan plot from the 'Browse' module based on the data from Lotta et al.

[7]. The x-axis is the position of the SNPs, and the y-axis represents different metabolites, while the z-axis corresponds to the significance of the association; (e) a screenshot showing the table view in the 'Browse' module.

2.3.1. Search and Browse. The 'Search' module supports searching the association results of the curated 65 mGWAS publications. Meanwhile, the 'Browse' module allows users to visually explore the data in a 3D Manhattan plot or network view. The 3D Manhattan plot strengthens the exploration of metabolome-wide pleiotropy at the genome-wide level [39]. Users can mouse-over a dot to see the SNP annotation, including the rsID, CHR:BP, nearest gene, most severe consequence, p-value, and metabolite name (Figure 2d).

2.3.2. From SNPs to Networks. Network-based approaches have become increasingly applied to identify shared genetic underpinnings (i.e., pleiotropy) in GWAS, where nodes are SNPs or phenotypes (e.g., metabolites or diseases) and edges represent significant associations [11]. The 'SNPs' module supports SNP–metabolite (or metabolite ratios), and SNP–disease network analysis. Optionally, LD proxy search (i.e., population type and r2) could be performed to maximize the results. Users also have the flexibility to include PPI networks, as well as to filter on biofluid or population types (Figure 2a).

2.3.3. From Metabolites to Networks. For many GIMs, metabolites can be functionally connected to enzymes, or transporters [3]. The 'Metabolites' module allows users to perform either statistical-based or knowledge-based metabolite-gene associations as well as metabolite-disease associations. Users can upload a list of metabolites from the upload page (Figure 2b). mGWAS-Explorer currently accepts either HMDB ID, KEGG ID or compound name. The uploaded list is then mapped to genes, SNPs, or diseases for network creation and subsequent visualization.

2.3.4. From Genes to Networks. The nearest gene mapping approach is suggested to be an effective indicator of true positive genes for mQTLs [44]. In mGWAS-Explorer, users can upload their gene lists in the 'Genes' module, the reversed nearest-gene mapping will be automatically performed and return SNPs that are significant in the mGWAS (Figure 2c). The network output will include genes, SNPs, and metabolites. Alternatively, users can perform gene–metabolite mapping via biochemical knowledge or to the associated diseases.

Case studies

2.4.1. COVID-19 Case Study. The host genetic variation is known to influence the severity of SARS-CoV-2 infection [45,46,47,48,49] and the blood metabolomics can reveal biomarkers for disease diagnosis and prognosis [50,51]. However, understanding mechanisms that link genetic variation to metabolism and clinical endpoints remains an important challenge. Therefore, we applied mGWAS-Explorer to a list of SNPs identified from a GWAS of severe COVID-19 [47] to provide insights into the shared genetic architecture of diseases and intermediate metabolic phenotypes. We used a suggestive significant association p-value threshold (1×10^{-5}) for mGWAS-Explorer, resulting in 19 SNPs after LD clumping. mGWAS-Explorer revealed that the SNPs at the *ABO* (alpha 1-3-N-acetylgalactosaminyltransferase and alpha 1-3-galactosyltransferase) locus were in high LD ($r^2 > 0.8$) with numerous other SNPs in this region associated with multiple metabolites and other human diseases, such as leucylalanine, citric acid [51], malaria [52], ischemic stroke [53], and venous thrombosis [54]. The blood type locus *ABO* has been linked to the risk of COVID-19 in several studies [47,55]. Multiple hypotheses have been proposed to explain the mechanism, such as anti-A and/or anti-B antibodies against corresponding antigens, or the glycosyltransferase activity [56]. mGWAS-Explorer provided insights into these

possible mechanisms, which identified associations of *ABO* variants with levels and the ratios of fibrinogen A-α peptides (e.g., ADpSGEGDFXAEGGGVR) and venous thromboembolism (Figure 3a). Fibrinogen plays a role in blood clotting [57]. Therefore, the association between *ABO* variants with fibrinogen may suggest that *ABO* influences COVID-19 via regulating thrombosis, which provided a functional explanation for the observed association of *ABO* with COVID-19 risk. Indeed, studies have reported that COVID-19 is associated with an increased risk of thromboembolism [58]. Therefore, we sought to investigate whether the association between fibrinogen A-α peptide-associated loci could provide additional insights into the underlying pathophysiology of COVID-19. Interestingly, mGWAS-Explorer revealed variants in *ENPEP* (glutamyl aminopeptidase) and *FUT2* (fucosyltransferase 2) genes are associated with levels and/or ratios of fibrinogen A-α peptides. Additionally, *FUT2* gene was also identified in the PPI network with the *ABO* gene (Figure 3a). In fact, *ENPEP* was discovered to be a candidate co-receptor for the coronavirus SARS-CoV-2 [59] and individuals with an inactivating *FUT2* mutations were more likely to develop a less severe form of the COVID-19 disease [60]. In summary, mGWAS-Explorer supports the evidence that *ABO*, *ENPEP*, and *FUT2* may be candidate genes and discovered fibrinogen A-α peptides as potential biomarkers for COVID-19 disease.

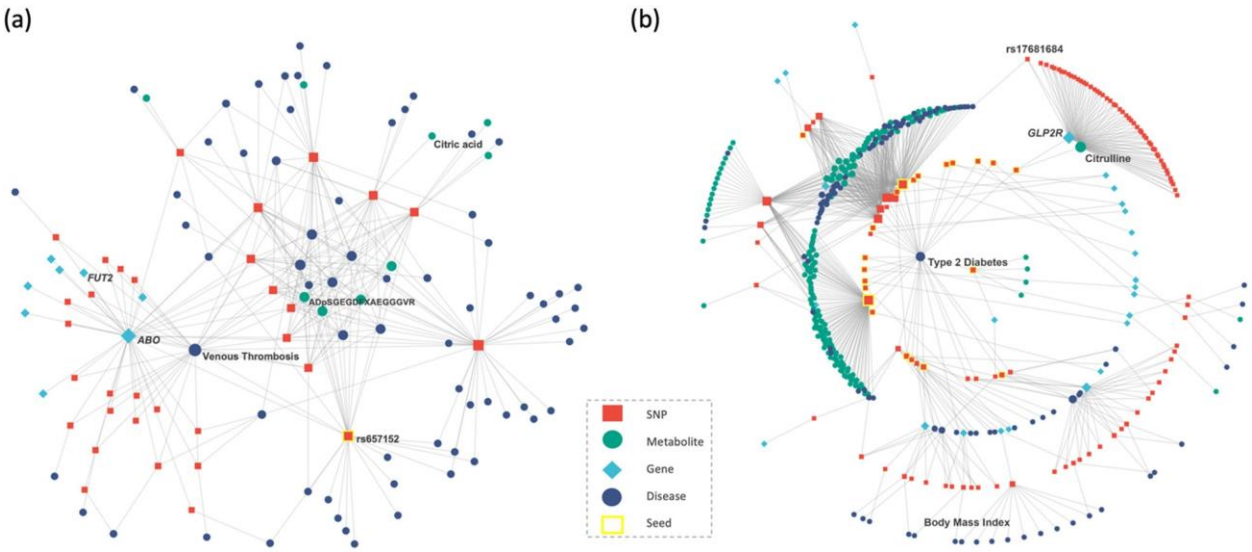

Figure 3. Screenshots of network results for the (a) COVID-19 case study and (b) type 2 diabetes case study. Each node represents either a SNP (orange square), a metabolite (green circle), a gene (blue diamond), a disease (dark blue circle), or a seed node (yellow square outline). Each edge is either an association between one SNP and one metabolite, an association between one SNP and one disease, a positional mapping of SNP to gene, or a protein–protein interaction. The size of a node is proportional to the number of other nodes connected to it.

2.4.2. Type 2 Diabetes Case Study. Around 250 genomic regions have been associated with type 2 diabetes (T2D) susceptibility in genome-wide association studies, some studies have highlighted the link to metabolomic profiles [7,61,62]. We applied mGWAS-Explorer to a list of SNPs from a published GWAS of T2D [63] in an attempt to examine shared genetic signals with circulating metabolites. Notably, mGWAS-Explorer confirmed the associations between citrulline metabolites, T2D, body mass index (Figure 3b), and identified the missense rs17681684 variant for citrulline in the *GLP2R* (glucagon like peptide 2 receptor) gene as reported by Lotta et al. [7]. Additionally, it identified shared genetic signals between T2D, coronary artery disease, and cholesterol levels at the *ABO* locus. Indeed, previous epidemiological studies have reported that

the associations of *ABO* group with coronary artery diseases are mediated by cholesterol [64], although the evidence regarding associations between *ABO* blood group with type 2 diabetes were not consistent [65,66,67,68]. Thus, further studies are required to identify the associations between *ABO* variants, T2D, and cholesterol levels. Furthermore, mGWAS-Explorer also revealed metabolites levels and their ratios identified in the previous COVID-19 case study shared associations with T2D loci. In fact, multiple studies have reported the comorbidity of T2D and COVID-19 [69,70,71]. In brief, analyzing the T2D cross-phenotype associations with metabolites and other diseases highlighted comorbid conditions with shared genetic signals, illustrating the usefulness of mGWAS-Explorer.

Comparison with other tools

Table 2 provides detailed comparisons of mGWAS-Explorer with several bioinformatics resources that can be used for mGWAS, including Metabolomics GWAS Server [33,34], PheWeb [35], NETMAGE [37], and GePhEx [72]. The metabolomics GWAS server supports searching the results of two genome-wide association studies on the blood and urine metabolome in 7824 and 3861 individuals with European ancestry [33,34]. PheWeb is an excellent tool for developers to build a website to explore and visualize large-scale genetic associations [35]. NETMAGE focuses on visualizing disease–disease networks from summary statistics [37], and GePhEx allows visualization and interpretation of relationships across multiple traits with genetic associations evidence [72].

Table 2. Comparison of the main features of mGWAS-Explorer with other web-based tools. Symbols used for feature evaluations with '√' for present, '−' for absent, and '+' for a more quantitative assessment (more '+' symbols indicate better support).

Tool Name	mGWAS-Explorer	Metabolomics GWAS Server	PheWeb	NETMAGE	GePhEx
Data input and processing					
SNP	√	√	√	√	√
LD proxy search	√	√	−	−	√
Gene	√	√	√	−	√
Metabolite	√	√	√	−	√
Enrichment analysis					
SNP-set	√	−	−	−	−
Gene-set	√	−	−	−	√
Metabolite-set	√	−	−	−	−
Cross-phenotype exploration	√	√	√	√	√
Visual analytics					
Network visualization	+++	−	−	+	−
Network customization	+++	−	−	+	−
Integration with PPI network	√	−	−	−	−
Subnetwork extraction	√	−	−	√	−
Topology-based filtering	√	−	−	−	−
3D Manhattan plot	√	−	−	−	−

Discussion

Establishing meaningful connections between diseases and deciphering molecular mechanisms that underpin shared genetic architectures are among the key objectives of GWAS. Our work shows that integrating mGWAS summary statistics, LD proxy search, and visual analytics can rapidly reveal multiple associations across metabolites and diseases, which can be utilized to better understand the ongoing global health crisis, such as the COVID-19 pandemic and type 2 diabetes.

When looking at the cross-phenotype associations between metabolites and diseases, it is important to investigate the shared SNPs identified in the mGWAS-Explorer output to examine

where the SNPs are located on the genome and the extent of overlapping of the SNPs. In fact, we consider mGWAS-Explorer as the initial stage in a pipeline for an in-depth mechanistic understanding of mGWAS before moving on to similarity analysis [26], colocalization analysis [14] or Mendelian randomization studies [73] to further investigate shared genetic signals in the same locus and to identify causal links. Ultimately, experimental studies in model organisms and human clinical studies are required to test the generated hypothesis to fully understand the mechanisms.

While our first case study highlighted shared genetic variants regulating metabolite abundance (e.g., citric acid and fibrinogen A-α peptides) and COVID-19 at *ABO*, much work needs to be done to fully understand the underlying mechanisms. Citric acid acts as a bridge between carbohydrate and fatty acid metabolism, promoting the growth and development of immune cells [74]. Additionally, citric acid is an important component in the TCA cycle. TCA cycle metabolites play key roles in signaling regulations of the innate and adaptive immune systems [75], which may be involved in COVID-19 pathogenesis. mGWAS-Explorer was also able to identify *ENPEP* and *FUT2* as potential candidate genes for COVID-19, although the association signal of these two genes were below the genome-wide significance threshold in the original study [47]. Many follow-up studies have reported *ABO* and *ENPEP* as COVID-19 risk genes; however, the evidence for the *FUT2* gene is conflicting [60,76,77]. Indeed, a recent whole-genome sequencing study identified variants in *FUT2* associated with critical COVID-19 diseases [45]. *FUT2* is responsible for the expression of histo-blood group antigens on the mucosal surface of gastrointestinal, genitourinary, and respiratory tracts. Inactivating *FUT2* mutations lead to a non-secretor status, which confer resistance to norovirus and rotavirus gastrointestinal infections [78,79]. Furthermore, the minor allele frequency of the stop–gain variant rs601338 in the *FUT2* gene is drastically different

between the European population (0.441) and the East Asian population (0.004) [80,81], which might explain the differences in host response to SARS-CoV-2 among different populations. However, the mechanism by which secretor status influencing COVID-19 pathogenesis is not fully understood. Therefore, it may be valuable to perform colocalization analysis and Mendelian randomization studies to identify the causal link between *FUT2* variants, fibrinogen A-α peptides, and COVID-19.

Increases in throughput and decreases in cost will enable a growing number of mGWAS to be conducted in the near future. The web server will be regularly updated to incorporate the most up-to-date mGWAS datasets, disease associations, and additional SNP annotation data (e.g., eQTLs or chromatin interactions) to serve as a valuable bioinformatics platform for mGWAS researchers. With this context, we also intend to add support for peak annotations of untargeted metabolomics data obtained from high-resolution mass spectrometry.

Materials and methods

Knowledgebase curation

(a) mGWAS papers were searched from PubMed, Web of Science, bioRxiv, and medRxiv, resulting in 65 publications as of December 2021. The summary statistics were either downloaded from public databases or supplementary data of the original publications. Statistical associations between metabolites and SNPs were summarized and pre-filtered using study-specific significance thresholds. In addition to p-values and effect sizes of SNP–metabolite associations, we have included metadata from each publication, such as the type of biofluid, sample size, population type, genotyping platform, metabolomics platform, etc. (b) For SNP annotation, three options are provided, including HaploReg [82], PhenoScanner [28], and VEP [20] by using the Application

Programming Interface (API) service of each database. For the first two options, users can also perform an LD proxy search based on different populations and r2 values. With VEP, users can select either a specific distance or the nearest number of genes for SNP annotation. (c) SNP–disease and gene–disease associations were downloaded from the DisGeNET database [32]. HMDB database was used to obtain metabolite–disease associations [83]. (d) KEGG, Recon3D, and Transporter Classification Database (TCDB) were used to curate knowledge-based gene–metabolite association information [84,85,86]. (e) The protein–protein interaction information is based on several well-established PPI databases [87,88,89]. (f) The libraries for enrichment analysis were curated from seven well-known databases, including GO, Reactome, KEGG, Orphanet, DrugMatrix, DisGeNET and DSigDB. The detailed description of these databases and their links can be found in the Supplementary Material.

Input processing and connection identification

SNPs are identified by rsIDs, genes are identified by Entrez IDs, and metabolites are identified by HMDB IDs, platform-specific IDs, or feature tags (m/z_retention_time). Additional identifiers have also been included, such as genomic coordinates for SNPs (after lifting all SNPs to GRCh37 assembly using LiftOver [90]), Ensemble ID, and gene symbol for genes, as well as KEGG ID and common name for metabolites.

There are two general types of relationships, including inter-omics and phenotype-specific links. Inter-omics connections are based on statistical associations (based on mGWAS), or knowledge-based associations (based on positional mapping for SNP–gene annotation or encoding enzymes/transporters for metabolite–gene connections). Phenotype-specific links allow users to identify variants that are associated with disease phenotypes. This information is obtained from

DisGeNET [32] based on case-control genome-wide association studies or via text mining from the literature.

Implementation

The backend analysis was implemented using the R programming language (version 4.1.3). The whole framework was built based on the PrimeFaces component library (version 11.0.0). The integrated data are stored in a relational database using SQLite. The interactive visualization was developed based on the sigma.js and echarts.js JavaScript libraries for network view and 3D Manhattan plot, respectively.

Data collection for case studies

The datasets of COVID-19 and type 2 diabetes case studies were downloaded from their respective original publications [47,63]. LD clumping were performed to identify the independent signals by using the ieugwasr package with default parameters [29] prior to the analysis. Specifically, the SNPs with the lowest p-value are retained, where SNPs in LD within a certain window are removed in LD clumping [91]. In both case studies, European population and r2 = 0.8 were set as input parameters for LD proxy search.

Network visual analytics

4.5.1. Network Creation and Customization. The default networks are built by querying for the direct mapping from the knowledgebase. Optionally, users can choose to expand the network by including PPIs in the SNP module. However, the result may suffer from the 'hairball' effect, which severely limits the usefulness and interpretability. Therefore, mGWAS-Explorer offers

support to refine large networks based on node degree or betweenness values, batch filtering, or the shortest paths, as well as by computing minimum subnetworks based on the prize-collecting Steiner Forest (PCSF) algorithm [92]. The detailed instructions on how to navigate the network visualization system can be found in our Supplementary Material.

4.5.2. Functional Enrichment Analysis. The combination of network visualization and functional enrichment analysis is a valuable tool for gaining key biological insights. For SNP input, two types of enrichment approaches have been implemented—(1) directly testing in SNP-set library, or (2) testing on mapped genes for enrichment using hypergeometric tests. When the input is a gene or metabolite, the associated gene-set or metabolite-set enrichment analysis can be performed. The result tables will be displayed under the Function Explorer panel. Notably, clicking a row of the table will highlight the nodes contained in the corresponding function/pathway within the network. In addition, mGWAS-Explorer also permits enrichment analysis on the selected nodes of interests, for instance, from the batch selection panel.

4.5.3. Other Advanced Features. The Network Viewer page contains multiple advanced features for network visual exploration, including Network Layout, Global Node Styles, Module Explorer, Batch Selection, and Path Finder. Ten different network layout algorithms are available, including Force-Atlas, Fruchterman–Reingold, Circular, Graphopt, Large Graph, Random, Circular Bipartite/Tripartite, Linear Bipartite/Tripartite, Concentric, and Backbone layout. Three module detection algorithms are offered in the Module Explorer, including the WalkTrap, InfoMap, and the Label Propagation algorithms based on the igraph R package [93]. These options can be combined to obtain a better visualization experience. Users can find details of these algorithms on our FAQs page.

Conclusions

We have developed mGWAS-Explorer to allow users to easily explore the published mGWAS datasets, and to provide contextualized analysis for a given list of SNPs, genes, or metabolites. As demonstrated by our case studies of COVID-19 and type 2 diabetes, mGWAS-Explorer can facilitate hypothesis generation and reveal functional insights into the genetic basis of human metabolism to permit translational discoveries.

Supplementary materials

Program description and methods

1. KNOWLEDGEBASE CREATION

1.1 Knowledgebase for network creation

1.1.1 SNP-metabolite association (mGWAS)

As of December 2021, 65 mGWAS papers had been found after searching PubMed, Web of Science, bioRxiv, and medRxiv, after a thorough literature research. The summary statistics were either collected from publicly available databases or supplementary data from the original publications. Study-specific significant thresholds were used to pre-filter statistical associations between metabolites and SNPs. Details of the curated mGWAS dataset can be found in our Resources page: https://www.mgwas.ca/mGWAS/faces/Secure/Resources.xhtml

1.1.2 SNP to gene mapping

Three options are provided for SNP to gene mapping, including HaploReg [1], PhenoScanner [2], and VEP [3] by using the Application Programming Interface (API) service of each database.

The Ensembl Variant Effect Predictor (VEP) is a comprehensive suite of tools for analyzing, annotating, and prioritizing genomic variants in both coding and non-coding regions. https://rest.ensembl.org/documentation/info/vep_id_get

PhenoScanner is a curated database of results from large-scale genetic association studies. https://github.com/phenoscanner/phenoscanner

HaploReg is a database for mining putative causal variants, cell types, regulators, and target genes for human diseases and complex traits.

https://github.com/izhbannikov/haploR

1.1.3 LD proxy search

mGWAS-Explorer allow users to search for metabolites/diseases associations with proxies for SNPs of interest using the HaploReg API or PhenoScanner API. The LD information is based on the 1000 Genomes Project.

1.1.4 SNP-disease association

DisGeNET was used to obtain SNP-disease associations, which contains both curated and literature data [4]. The curated data include SNP-disease associations from UniProt [5], ClinVar [6], GWAS Catalog [7], and GWASdb [8].

https://www.disgenet.org/downloads

1.1.5 Gene-metabolite association

Knowledge-based gene-metabolite association information was curated using KEGG, Recon3D, and the Transporter Classification Database (TCDB) [9-11].

KEGG: metabolite-gene associations based on KEGG reaction. https://www.genome.jp/kegg/

Recon3D: high-quality genome-scale metabolic reconstruction. https://www.vmh.life/

TCDB: transporter classification database for transporter protein information. https://www.tcdb.org/

1.1.6 Protein-protein interaction

Information on protein-protein interaction was taken from four well-established PPI databases, including InnateDB [12], STRING [13], HuRI [14], and Rolland et al [15].

InnateDB contains literature-curated data on protein-protein interactions. https://www.innatedb.com/index.jsp

STRING is a comprehensive database for protein-protein interaction networks. We have filtered on medium (400) to high (900) confidence score.

https://string-db.org/

HuRI is a reference interactome map of human binary protein interactions. http://www.interactome-atlas.org/

Rolland et al. contains experimentally validated binary human PPI data. http://interactome.dfci.harvard.edu/H_sapiens/

1.1.7 Gene-disease association

DisGeNET was also used to obtain gene-disease associations. https://www.disgenet.org/downloads

1.1.8 Metabolite-disease association

HMDB was used to retrieve metabolite-disease associations. https://hmdb.ca/downloads

1.2 Knowledgebase for network interpretation

For network analysis, it is crucial to be able to interpret the results in addition to visualization. In this regard, enrichment analysis plays a key role. Thus, we have implemented three types of enrichment analysis, including SNP-set, gene-set and metabolite-set enrichment analysis.

1.2.1 SNP-sets

DisGeNET was used for SNP-sets, where diseases having three or more SNPs are taken into account when creating a SNP-set.

https://www.disgenet.org/downloads

1.2.2 Gene-sets

Gene Ontology (GO) was used to obtain gene-sets for biological processes, molecular functions, and cellular components.

http://geneontology.org/

Reactome and KEGG was used to get gene-sets for pathways. https://reactome.org/download-data

https://www.kegg.jp/

Orphanet was used to obtain gene-sets for rare diseases. http://www.orphadata.org/cgi-bin/index.php

DrugMatrix and DSigDB were used to obtain gene-sets that related drugs and their target genes. https://ntp.niehs.nih.gov/data/drugmatrix/

http://dsigdb.tanlab.org/DSigDBv1.0/

1.2.3 Metabolite-sets

KEGG was used to retrieve metabolite-sets for KEGG pathways. https://www.kegg.jp/

2. DATA SEARCH

The ‘Search’ module allows users to search the results for significant SNP-metabolite associations from the curated mGWAS datasets. Users can enter the rsID or Common Name in the search bar, where autocomplete is supported. The search results will return in the table below.

3. DATA BROWSE

The ‘Browse’ module allows users to visually examine the summary statistics from individual mGWAS datasets in a 3D Manhattan plot. Users can zoom in/out or rotate the 3D plot and mouse over on the dots to see detailed information. Meanwhile, table view and network view are also provided.

4. DATA UPLOAD AND PROCESSING

4.1 Data inputs

The flexible interface allows users to start from SNPs, genes, or metabolites. The input can be uploaded by entering a list of IDs (SNPs, genes, metabolites). Users can refer to the relevant FAQs and tutorials or see our test examples for more details.

mGWAS-Explorer currently supports rsID for SNPs, Ensembl gene ID, Entrez ID and official gene symbol for genes, HMDB ID, KEGG ID and compound name for metabolites. Additionally, users can filter on “Biofluid” or “Population”, the results will return SNP-metabolite associations on the metabolites measured in the chosen biofluid and population.

5. NETWORK CREATION AND REFINEMENT

The input data from users will be searched against our knowledgebase. The resulting pair-wise tables are used to generate the default networks. Since not all nodes will be connected, the results may return many networks, usually one large network with a few smaller ones. Table summary and network summary are displayed (Figure S1), indicating the statistics for nodes and edges to allow users to have an overview of the networks. We recommend that users keep their networks below 2000 nodes for practical reasons, since large networks will induce a ‘hairball’ effect, which make it difficult to comprehend the results. mGWAS-Explorer has built-in network tools that allow users to refine networks according to topological measurements (degree, betweenness, and shortest path), batch filtering, and computing minimum subnetworks based on the prize-collecting Steiner Forest (PCSF) algorithm.

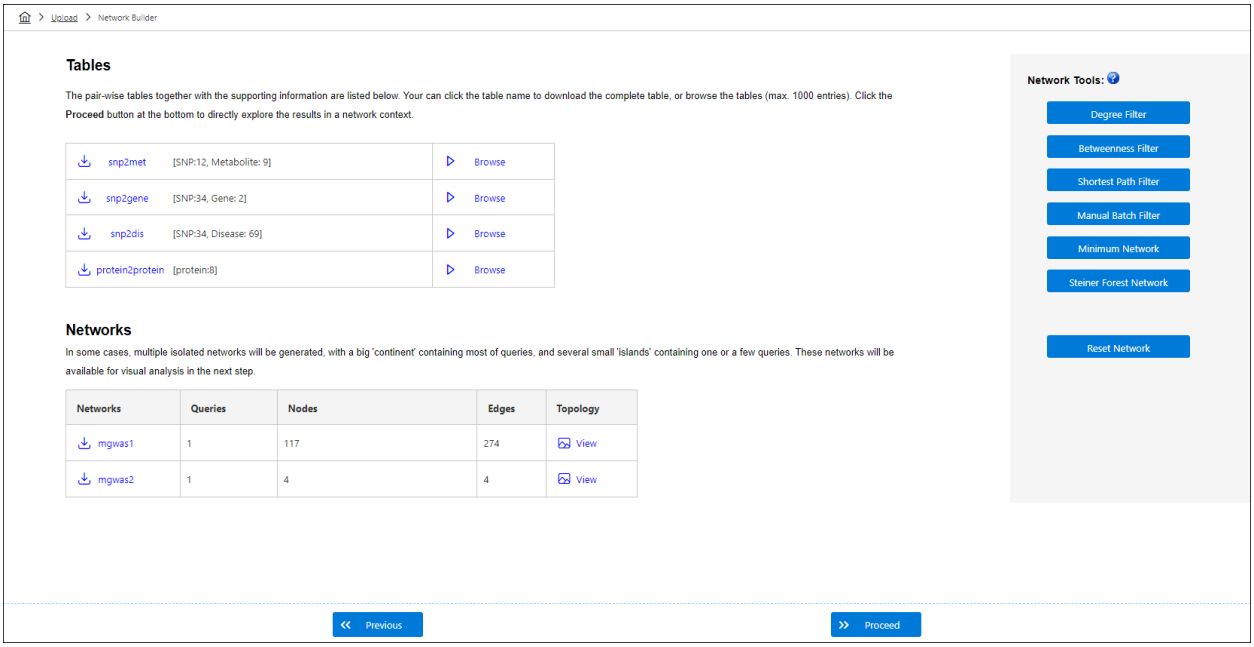

Figure S1. A screenshot of the Network Builder page, including table statistics, network statistics and network tools.

6. NETWORK VISUALIZATION AND FUNCTIONAL ANALYSIS

The HTML5 canvas and JavaScript were used to develop the network visualization system. Figure S2 shows a screenshot of the network visualization interface. The network visualization system comprises four main components: the top menu bar, the left node table, the center network viewing area, and the right panel.

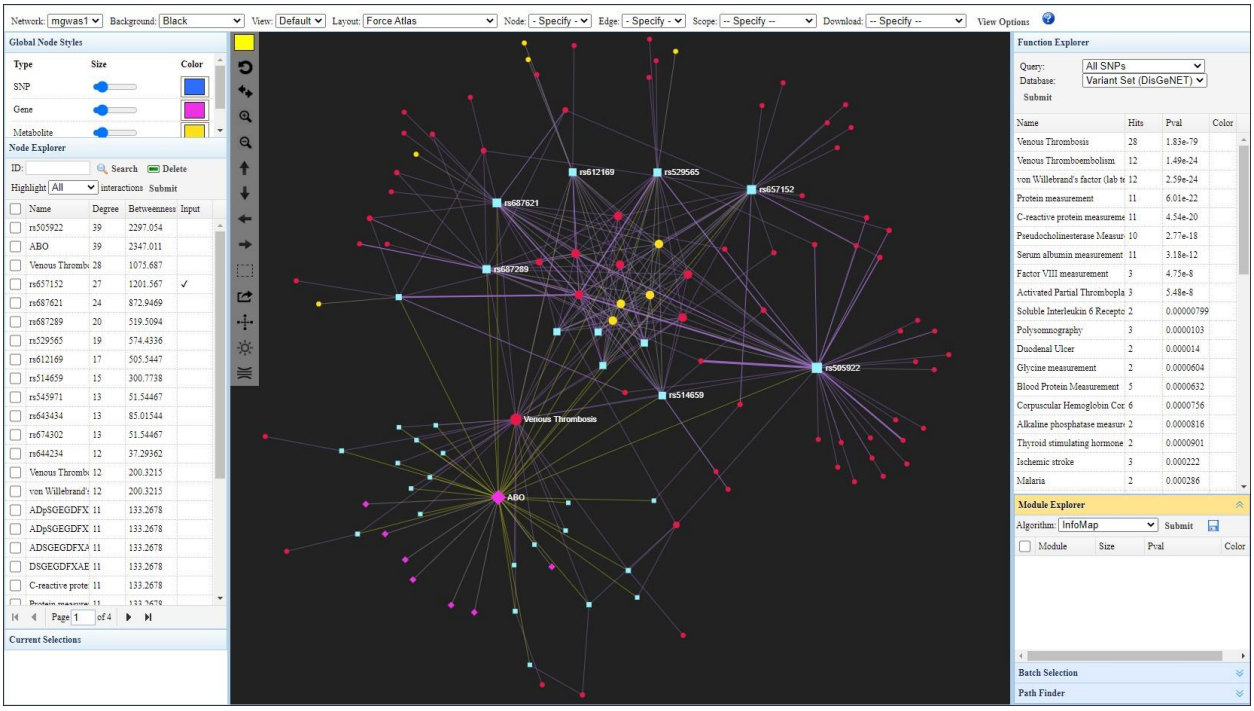

Figure S2. A screenshot of the mGWAS-Explorer network visual analytic system. The system comprises four main components: the top menu bar, the left node table, the center network viewing area, and the right panel. Users can easily highlight and arrange nodes based on their connectivity patterns or enriched functions.

6.1 General network customization

The top menu bar offers common functions to customize the network, such as changing the background color, adjusting the characteristics of the nodes (label, color, size, shape) and edges (opacity, thickness, color), or downloading the results. Users can select the preferred network layout by using the 'Layout' option. The 'Scope' option allows users to specify the mouse operation range during drag-and-drop, either 'Single node', 'Node-neighbors', 'All highlights', 'Current highlights', or 'Node type'.

6.2 Node searching and edge viewing

The nodes of the network are displayed in the Node Explorer on the left side, along with their degree and betweenness measurements. A checkmark will show in the 'Input' column if it is a seed node provided by the user. Users can click on a row or type the ID in the search bar to view the node of interest. The network will zoom to the selected node automatically. Alternatively,

multiple nodes can be selected by clicking the checkboxes. Users can decide to highlight 'All' or the 'Shared' nodes accordingly. Meanwhile, double clicking an edge will show the evidence that supports the connection between the nodes.

6.3 Functional enrichment analysis

The combination of network visualization and functional enrichment analysis can provide valuable biological insights. mGWAS-Explorer supports over representation analysis (ORA) [16]. ORA is a widely used method to assess whether known biological functions or pathways are over-represented (i.e., enriched) in a list of interest (e.g., SNP, gene, or metabolite). Hypergeometric tests are used to calculate the p-values.

6.4 Other Advanced Features

The bottom right panel contains three tabs – Module Explorer, Batch Selection, and Path Finder. The Module Explorer tab provides three different approaches for module detection – the WalkTrap, InfoMap, and Label propagation algorithms. Users can perform module detection to identify tightly clustered subnetworks with more internal connections than would be expected at random in the whole network. The Batch Selection tab allows users to highlight or exclude a list of nodes. The Path Finder tab allows users to find the shortest path between any two nodes.

7. IMPLEMENTATION

The backend of mGWAS-Explorer was implemented using the R programming language (version 4.1.3; https://www.r-project.org/). The whole framework was built based on the JavaServer Faces technology using the PrimeFaces component library (version 11.0; https://www.primefaces.org/). The integrated data is stored in a relational database using SQLite (https://www.sqlite.org/index.html). Network visualization and analysis are based on jquery (https://jquery.com/) for general-purpose scripting, sigma.js (https://www.sigmajs.org/) for network display and interactions, and igraph (https://igraph.org/) for network analysis and layout. 3D Manhattan plot is built on ECharts-GL, an extension pack of Apache ECharts, which provides 3D plots and WebGL acceleration (https://echarts.apache.org/en/index.html).

Funding

Genome Canada: Genome Quebec, Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant, Canada Research Chairs Program. L. Chang was supported by a PhD Scholarship from the NSERC-MATRIX program.

Data availability statement

The data is available from https://www.mgwas.ca/mGWAS/faces/Secure/Resources.xhtml (accessed on 1 May 2022).

References

1. Visscher, P.M.; Wray, N.R.; Zhang, Q.; Sklar, P.; McCarthy, M.I.; Brown, M.A.; Yang, J. 10 years of GWAS discovery: Biology, function, and translation. Am. J. Hum. Genet. 2017, 101, 5–22.

2. Cano-Gamez, E.; Trynka, G. From GWAS to Function: Using Functional Genomics to Identify the Mechanisms Underlying Complex Diseases. Front. Genet. 2020, 11, 424.

3. Kastenmüller, G.; Raffler, J.; Gieger, C.; Suhre, K. Genetics of human metabolism: An update. Hum. Mol. Genet. 2015, 24, R93–R101.

4. Hagenbeek, F.A.; Pool, R.; van Dongen, J.; Draisma, H.H.M.; Jan Hottenga, J.; Willemsen, G.; Abdellaoui, A.; Fedko, I.O.; den Braber, A.; Visser, P.J.; et al. Heritability estimates for 361 blood metabolites across 40 genome-wide association studies. Nat. Commun. 2020, 11, 39.

5. Gieger, C.; Geistlinger, L.; Altmaier, E.; De Angelis, M.H.; Kronenberg, F.; Meitinger, T.; Mewes, H.-W.; Wichmann, H.-E.; Weinberger, K.M.; Adamski, J. Genetics meets metabolomics: A genome-wide association study of metabolite profiles in human serum. PLoS Genet. 2008, 4, e1000282.

6. Gallois, A.; Mefford, J.; Ko, A.; Vaysse, A.; Julienne, H.; Ala-Korpela, M.; Laakso, M.; Zaitlen, N.; Pajukanta, P.; Aschard, H. A comprehensive study of metabolite genetics reveals strong pleiotropy and heterogeneity across time and context. Nat. Commun. 2019, 10, 4788.

7. Lotta, L.A.; Pietzner, M.; Stewart, I.D.; Wittemans, L.B.L.; Li, C.; Bonelli, R.; Raffler, J.; Biggs, E.K.; Oliver-Williams, C.; Auyeung, V.P.W.; et al. A cross-platform approach identifies genetic regulators of human metabolism and health. Nat. Genet. 2021, 53, 54–64.

8. Solovieff, N.; Cotsapas, C.; Lee, P.H.; Purcell, S.M.; Smoller, J.W. Pleiotropy in complex traits: Challenges and strategies. Nat. Rev. Genet. 2013, 14, 483.

9. Visscher, P.M.; Yang, J. A plethora of pleiotropy across complex traits. Nat. Genet. 2016, 48, 707.

10. Watanabe, K.; Stringer, S.; Frei, O.; Umićević Mirkov, M.; de Leeuw, C.; Polderman, T.J.C.; van der Sluis, S.; Andreassen, O.A.; Neale, B.M.; Posthuma, D. A global overview of pleiotropy and genetic architecture in complex traits. Nat. Genet. 2019, 51, 1339–1348.

11. Weighill, D.; Jones, P.; Bleker, C.; Ranjan, P.; Shah, M.; Zhao, N.; Martin, M.; DiFazio, S.; Macaya-Sanz, D.; Schmutz, J.; et al. Multi-Phenotype Association Decomposition: Unraveling Complex Gene-Phenotype Relationships. Front. Genet. 2019, 10, 417.

12. Julienne, H.; Laville, V.; McCaw, Z.R.; He, Z.; Guillemot, V.; Lasry, C.; Ziyatdinov, A.; Nerin, C.; Vaysse, A.; Lechat, P.; et al. Multitrait GWAS to connect disease variants and biological mechanisms. PLoS Genet. 2021, 17, e1009713.

13. Bulik-Sullivan, B.K.; Loh, P.R.; Finucane, H.K.; Ripke, S.; Yang, J.; Patterson, N.; Daly, M.J.; Price, A.L.; Neale, B.M. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. Nat. Genet. 2015, 47, 291–295.

14. Giambartolomei, C.; Vukcevic, D.; Schadt, E.E.; Franke, L.; Hingorani, A.D.; Wallace, C.; Plagnol, V. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. PLoS Genet. 2014, 10, e1004383.

15. Majumdar, A.; Haldar, T.; Bhattacharya, S.; Witte, J.S. An efficient Bayesian meta-analysis approach for studying cross-phenotype genetic associations. PLoS Genet. 2018, 14, e1007139.

16. Han, B.; Pouget, J.G.; Slowikowski, K.; Stahl, E.; Lee, C.H.; Diogo, D.; Hu, X.; Park, Y.R.; Kim, E.; Gregersen, P.K.; et al. A method to decipher pleiotropy by detecting underlying heterogeneity driven by hidden subgroups applied to autoimmune and neuropsychiatric diseases. Nat. Genet. 2016, 48, 803–810.

17. Trochet, H.; Pirinen, M.; Band, G.; Jostins, L.; McVean, G.; Spencer, C.C.A. Bayesian meta-analysis across genome-wide association studies of diverse phenotypes. Genet. Epidemiol. 2019, 43, 532–547.

18. Li, X.; Zhu, X. Cross-Phenotype Association Analysis Using Summary Statistics from GWAS. Methods Mol. Biol. 2017, 1666, 455–467.

19. Arnold, M.; Raffler, J.; Pfeufer, A.; Suhre, K.; Kastenmüller, G. SNiPA: An interactive, genetic variant-centered annotation browser. Bioinformatics 2014, 31, 1334–1336.

20. McLaren, W.; Gil, L.; Hunt, S.E.; Riat, H.S.; Ritchie, G.R.; Thormann, A.; Flicek, P.; Cunningham, F. The ensembl variant effect predictor. Genome Biol. 2016, 17, 122.

21. Carlin, D.E.; Fong, S.H.; Qin, Y.; Jia, T.; Huang, J.K.; Bao, B.; Zhang, C.; Ideker, T. A Fast and Flexible Framework for Network-Assisted Genomic Association. iScience 2019, 16, 155–161.

22. Bastarache, L.; Denny, J.C.; Roden, D.M. Phenome-Wide Association Studies. JAMA 2022, 327, 75–76.

23. Tanha, H.M.; Sathyanarayanan, A.; Nyholt, D.R. Genetic overlap and causality between blood metabolites and migraine. Am. J. Hum. Genet. 2021, 108, 2086–2098.

24. Kaur, Y.; Wang, D.X.; Liu, H.Y.; Meyre, D. Comprehensive identification of pleiotropic loci for body fat distribution using the NHGRI-EBI Catalog of published genome-wide association studies. Obes. Rev. 2019, 20, 385–406.

25. Guo, Y.; Rist, P.M.; Daghlas, I.; Giulianini, F.; Kurth, T.; Chasman, D.I. A genome-wide cross-phenotype meta-analysis of the association of blood pressure with migraine. Nat. Commun. 2020, 11, 3368.

26. George, G.; Huang, Y.; Gan, S.; Nar, A.S.; Ha, J.; Venkatesan, R.; Mohan, V.; Wang, H.; Brown, A.; Palmer, C.N.A.; et al. iPheGWAS: An intelligent computational framework to integrate and visualise genome-phenome wide association results. bioRxiv 2022.

27. Buniello, A.; MacArthur, J.A.L.; Cerezo, M.; Harris, L.W.; Hayhurst, J.; Malangone, C.; McMahon, A.; Morales, J.; Mountjoy, E.; Sollis, E.; et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res. 2019, 47, D1005–D1012.

28. Kamat, M.A.; Blackshaw, J.A.; Young, R.; Surendran, P.; Burgess, S.; Danesh, J.; Butterworth, A.S.; Staley, J.R. PhenoScanner V2: An expanded tool for searching human genotype-phenotype associations. Bioinformatics 2019, 35, 4851–4853.

29. Elsworth, B.; Lyon, M.; Alexander, T.; Liu, Y.; Matthews, P.; Hallett, J.; Bates, P.; Palmer, T.; Haberland, V.; Smith, G.D.; et al. The MRC IEU OpenGWAS data infrastructure. bioRxiv 2020.

30. Ghoussaini, M.; Mountjoy, E.; Carmona, M.; Peat, G.; Schmidt, E.M.; Hercules, A.; Fumis, L.; Miranda, A.; Carvalho-Silva, D.; Buniello, A.; et al. Open Targets Genetics: Systematic identification of trait-associated genes using large-scale genetics and functional genomics. Nucleic Acids Res. 2021, 49, D1311–D1320.

31. Shashkova, T.I.; Pakhomov, E.D.; Gorev, D.D.; Karssen, L.C.; Joshi, P.K.; Aulchenko, Y.S. PheLiGe: An interactive database of billions of human genotype-phenotype associations. Nucleic Acids Res. 2021, 49, D1347–D1350.

32. Piñero, J.; Ramírez-Anguita, J.M.; Saüch-Pitarch, J.; Ronzano, F.; Centeno, E.; Sanz, F.; Furlong, L.I. The DisGeNET knowledge platform for disease genomics: 2019 update. Nucleic Acids Res 2020, 48, D845–D855.

33. Shin, S.-Y.; Fauman, E.B.; Petersen, A.-K.; Krumsiek, J.; Santos, R.; Huang, J.; Arnold, M.; Erte, I.; Forgetta, V.; Yang, T.-P. An atlas of genetic influences on human blood metabolites. Nat. Genet. 2014, 46, 543.

34. Raffler, J.; Friedrich, N.; Arnold, M.; Kacprowski, T.; Rueedi, R.; Altmaier, E.; Bergmann, S.; Budde, K.; Gieger, C.; Homuth, G.; et al. Genome-Wide Association Study with Targeted and Non-targeted NMR Metabolomics Identifies 15 Novel Loci of Urinary Human Metabolic Individuality. PLoS Genet. 2015, 11, e1005487.

35. Gagliano Taliun, S.A.; VandeHaar, P.; Boughton, A.P.; Welch, R.P.; Taliun, D.; Schmidt, E.M.; Zhou, W.; Nielsen, J.B.; Willer, C.J.; Lee, S.; et al. Exploring and visualizing large-scale genetic associations by using PheWeb. Nat. Genet. 2020, 52, 550–552.

36. Wang, L.; Balmat, T.J.; Antonia, A.L.; Constantine, F.J.; Henao, R.; Burke, T.W.; Ingham, A.; McClain, M.T.; Tsalik, E.L.; Ko, E.R.; et al. An atlas connecting shared genetic architecture of human diseases and molecular phenotypes provides insight into COVID-19 susceptibility. Genome Med. 2021, 13, 83.

37. Sriram, V.; Shivakumar, M.; Jung, S.H.; Nam, Y.; Bang, L.; Verma, A.; Lee, S.; Choe, E.K.; Kim, D. NETMAGE: A human disease phenotype map generator for the network-based visualization of phenome-wide association study results. Gigascience 2022, 11, giac002.

38. Strayer, N.; Shirey-Rice, J.K.; Shyr, Y.; Denny, J.C.; Pulley, J.M.; Xu, Y. PheWAS-ME: A web-app for interactive exploration of multimorbidity patterns in PheWAS. Bioinformatics 2021, 37, 1778–1780.

39. George, G.; Gan, S.; Huang, Y.; Appleby, P.; Nar, A.S.; Venkatesan, R.; Mohan, V.; Palmer, C.N.A.; Doney, A.S.F. PheGWAS: A new dimension to visualize GWAS across multiple phenotypes. Bioinformatics 2020, 36, 2500–2505.

40. Zhu, Z.; Anttila, V.; Smoller, J.W.; Lee, P.H. Statistical power and utility of meta-analysis methods for cross-phenotype genome-wide association studies. PLoS ONE 2018, 13, e0193256.

41. Lee, B.; Zhang, S.; Poleksic, A.; Xie, L. Heterogeneous Multi-Layered Network Model for Omics Data Integration and Analysis. Front. Genet. 2019, 10, 1381.

42. Sadegh, S.; Skelton, J.; Anastasi, E.; Bernett, J.; Blumenthal, D.B.; Galindez, G.; Salgado-Albarrán, M.; Lazareva, O.; Flanagan, K.; Cockell, S.; et al. Network medicine for disease module identification and drug repurposing with the NeDRex platform. Nat. Commun. 2021, 12, 6848.

43. Petersen, A.-K.; Krumsiek, J.; Wägele, B.; Theis, F.J.; Wichmann, H.-E.; Gieger, C.; Suhre, K. On the hypothesis-free testing of metabolite ratios in genome-wide and metabolome-wide association studies. BMC Bioinform. 2012, 13, 120.

44. Stacey, D.; Fauman, E.B.; Ziemek, D.; Sun, B.B.; Harshfield, E.L.; Wood, A.M.; Butterworth, A.S.; Suhre, K.; Paul, D.S. ProGeM: A framework for the prioritization of candidate causal genes at molecular quantitative trait loci. Nucleic Acids Res. 2018, 47, e3.

45. Kousathanas, A.; Pairo-Castineira, E.; Rawlik, K.; Stuckey, A.; Odhams, C.A.; Walker, S.; Russell, C.D.; Malinauskas, T.; Wu, Y.; Millar, J.; et al. Whole genome sequencing reveals host factors underlying critical COVID-19. Nature 2022.

46. Pairo-Castineira, E.; Clohisey, S.; Klaric, L.; Bretherick, A.D.; Rawlik, K.; Pasko, D.; Walker, S.; Parkinson, N.; Fourman, M.H.; Russell, C.D.; et al. Genetic mechanisms of critical illness in COVID-19. Nature 2021, 591, 92–98.

47. Ellinghaus, D.; Degenhardt, F.; Bujanda, L.; Buti, M.; Albillos, A.; Invernizzi, P.; Fernández, J.; Prati, D.; Baselli, G.; Asselta, R.; et al. Genomewide Association Study of Severe COVID-19 with Respiratory Failure. N. Engl. J. Med. 2020, 383, 1522–1534.

48. Mapping the human genetic architecture of COVID-19. Nature 2021, 600, 472–477.

49. Zhang, Q.; Bastard, P.; Liu, Z.; Le Pen, J.; Moncada-Velez, M.; Chen, J.; Ogishi, M.; Sabli, I.K.D.; Hodeib, S.; Korol, C.; et al. Inborn errors of type I IFN immunity in patients with life-threatening COVID-19. Science 2020, 370, eabd4570.

50. Sindelar, M.; Stancliffe, E.; Schwaiger-Haber, M.; Anbukumar, D.S.; Adkins-Travis, K.; Goss, C.W.; O'Halloran, J.A.; Mudd, P.A.; Liu, W.C.; Albrecht, R.A.; et al. Longitudinal metabolomics of human plasma reveals prognostic markers of COVID-19 disease severity. Cell Rep. Med. 2021, 2, 100369.

51. Shi, D.; Yan, R.; Lv, L.; Jiang, H.; Lu, Y.; Sheng, J.; Xie, J.; Wu, W.; Xia, J.; Xu, K.; et al. The serum metabolome of COVID-19 patients is distinctive and predictive. Metab. Clin. Exp. 2021, 118, 154739.

52. Timmann, C.; Thye, T.; Vens, M.; Evans, J.; May, J.; Ehmen, C.; Sievertsen, J.; Muntau, B.; Ruge, G.; Loag, W.; et al. Genome-wide association study indicates two novel resistance loci for severe malaria. Nature 2012, 489, 443–446.

53. Li, H.; Cai, Y.; Xu, A.D. Association study of polymorphisms in the ABO gene and their gene-gene interactions with ischemic stroke in Chinese population. J. Clin. Lab. Anal. 2018, 32, e22329.

54. Germain, M.; Saut, N.; Oudot-Mellakh, T.; Letenneur, L.; Dupuy, A.M.; Bertrand, M.; Alessi, M.C.; Lambert, J.C.; Zelenika, D.; Emmerich, J.; et al. Caution in interpreting results from imputation analysis when linkage disequilibrium extends over a large distance: A case study on venous thrombosis. PLoS ONE 2012, 7, e38538.

55. Zhao, J.; Yang, Y.; Huang, H.; Li, D.; Gu, D.; Lu, X.; Zhang, Z.; Liu, L.; Liu, T.; Liu, Y.; et al. Relationship between the ABO Blood Group and the Coronavirus Disease 2019 (COVID-19) Susceptibility. Clin. Infect. Dis. 2021, 73, 328–331.

56. Goel, R.; Bloch, E.M.; Pirenne, F.; Al-Riyami, A.Z.; Crowe, E.; Dau, L.; Land, K.; Townsend, M.; Jecko, T.; Rahimi-Levene, N.; et al. ABO blood group and COVID-19: A review on behalf of the ISBT COVID-19 Working Group. Vox Sang. 2021, 116, 849–861.

57. Kattula, S.; Byrnes, J.R.; Wolberg, A.S. Fibrinogen and Fibrin in Hemostasis and Thrombosis. Arterioscler. Thromb. Vasc. Biol. 2017, 37, e13–e21.

58. Malas, M.B.; Naazie, I.N.; Elsayed, N.; Mathlouthi, A.; Marmor, R.; Clary, B. Thromboembolism risk of COVID-19 is high and associated with a higher risk of mortality: A systematic review and meta-analysis. EClinicalMedicine 2020, 29, 100639.

59. Lange, C.; Wolf, J.; Auw-Haedrich, C.; Schlecht, A.; Boneva, S.; Lapp, T.; Horres, R.; Agostini, H.; Martin, G.; Reinhard, T.; et al. Expression of the COVID-19 receptor ACE2 in the human conjunctiva. J. Med. Virol. 2020, 92, 2081–2086.

60. Mankelow, T.J.; Singleton, B.K.; Moura, P.L.; Stevens-Hernandez, C.J.; Cogan, N.M.; Gyorffy, G.; Kupzig, S.; Nichols, L.; Asby, C.; Pooley, J.; et al. Blood group type A secretors are associated with a higher risk of COVID-19 cardiovascular disease complications. EJHaem 2021, 2, 175–187.

61. Langenberg, C.; Lotta, L.A. Genomic insights into the causes of type 2 diabetes. Lancet 2018, 391, 2463–2474.

62. Fuchsberger, C.; Flannick, J.; Teslovich, T.M.; Mahajan, A.; Agarwala, V.; Gaulton, K.J.; Ma, C.; Fontanillas, P.; Moutsianas, L.; McCarthy, D.J.; et al. The genetic architecture of type 2 diabetes. Nature 2016, 536, 41–47.

63. Scott, R.A.; Scott, L.J.; Mägi, R.; Marullo, L.; Gaulton, K.J.; Kaakinen, M.; Pervjakova, N.; Pers, T.H.; Johnson, A.D.; Eicher, J.D.; et al. An Expanded Genome-Wide Association Study of Type 2 Diabetes in Europeans. Diabetes 2017, 66, 2888–2902.

64. Chen, Y.; Chen, C.; Ke, X.; Xiong, L.; Shi, Y.; Li, J.; Tan, X.; Ye, S. Analysis of circulating cholesterol levels as a mediator of an association between ABO blood group and coronary heart disease. Circulation. Cardiovasc. Genet. 2014, 7, 43–48.

65. Li, S.; Schooling, C.M. A phenome-wide association study of ABO blood groups. BMC Med. 2020, 18, 334.

66. Meo, S.A.; Rouq, F.A.; Suraya, F.; Zaidi, S.Z. Association of ABO and Rh blood groups with type 2 diabetes mellitus. Eur. Rev. Med. Pharmacol. Sci. 2016, 20, 237–242.

67. Fagherazzi, G.; Gusto, G.; Clavel-Chapelon, F.; Balkau, B.; Bonnet, F. ABO and Rhesus blood groups and risk of type 2 diabetes: Evidence from the large E3N cohort study. Diabetologia 2015, 58, 519–522.

68. Yahaya, T.O.; Oladele, E.O.; Mshelia, M.B.; Sifau, M.O.; Fashola, O.D.; Bunza, M.; Nathaniel, J. Influence of ABO blood groups and demographic characteristics on the prevalence of type 2 diabetes in Lagos, southwest Nigeria. Bull. Natl. Res. Cent. 2021, 45, 144.

69. Zhu, L.; She, Z.G.; Cheng, X.; Qin, J.J.; Zhang, X.J.; Cai, J.; Lei, F.; Wang, H.; Xie, J.; Wang, W.; et al. Association of Blood Glucose Control and Outcomes in Patients with COVID-19 and Pre-existing Type 2 Diabetes. Cell Metab. 2020, 31, 1068–1077.e1063.

70. Metwally, A.A.; Mehta, P.; Johnson, B.S.; Nagarjuna, A.; Snyder, M.P. COVID-19-Induced New-Onset Diabetes: Trends and Technologies. Diabetes 2021, 70, 2733–2744.

71. Rajpal, A.; Rahimi, L.; Ismail-Beigi, F. Factors leading to high morbidity and mortality of COVID-19 in patients with type 2 diabetes. J. Diabetes 2020, 12, 895–908.

72. Farré, X.; Spataro, N.; Haziza, F.; Rambla, J.; Navarro, A. Genome-phenome explorer (GePhEx): A tool for the visualization and interpretation of phenotypic relationships supported by genetic evidence. Bioinformatics 2019, 36, 890–896.

73. Hemani, G.; Zheng, J.; Elsworth, B.; Wade, K.H.; Haberland, V.; Baird, D.; Laurin, C.; Burgess, S.; Bowden, J.; Langdon, R.; et al. The MR-Base platform supports systematic causal inference across the human phenome. eLife 2018, 7, e34408.

74. Williams, N.C.; O'Neill, L.A.J. A Role for the Krebs Cycle Intermediate Citrate in Metabolic Reprogramming in Innate Immunity and Inflammation. Front. Immunol. 2018, 9, 141.

75. Martínez-Reyes, I.; Chandel, N.S. Mitochondrial TCA cycle metabolites control physiology and disease. Nat. Commun. 2020, 11, 102.

76. Delanghe, J.R.; De Buyzere, M.L.; Speeckaert, M.M. Genetic Polymorphisms in the Host and COVID-19 Infection. Adv. Exp. Med. Biol. 2021, 1318, 109–118.

77. Matzhold, E.M.; Berghold, A.; Bemelmans, M.K.B.; Banfi, C.; Stelzl, E.; Kessler, H.H.; Steinmetz, I.; Krause, R.; Wurzer, H.; Schlenke, P.; et al. Lewis and ABO histo-blood types and the secretor status of patients hospitalized with COVID-19 implicate a role for ABO antibodies in susceptibility to infection with SARS-CoV-2. Transfusion 2021, 61, 2736–2745.

78. Lindesmith, L.; Moe, C.; Marionneau, S.; Ruvoen, N.; Jiang, X.; Lindblad, L.; Stewart, P.; LePendu, J.; Baric, R. Human susceptibility and resistance to Norwalk virus infection. Nat. Med. 2003, 9, 548–553.

79. Payne, D.C.; Currier, R.L.; Staat, M.A.; Sahni, L.C.; Selvarangan, R.; Halasa, N.B.; Englund, J.A.; Weinberg, G.A.; Boom, J.A.; Szilagyi, P.G.; et al. Epidemiologic Association between FUT2 Secretor Status and Severe Rotavirus Gastroenteritis in Children in the United States. JAMA Pediatrics 2015, 169, 1040–1045.

80. Cunningham, F.; Allen, J.E.; Allen, J.; Alvarez-Jarreta, J.; Amode, M.R.; Armean, I.M.; Austine-Orimoloye, O.; Azov, A.G.; Barnes, I.; Bennett, R.; et al. Ensembl 2022. Nucleic Acids Res. 2021, 50, D988–D995.

81. Ferrer-Admetlla, A.; Sikora, M.; Laayouni, H.; Esteve, A.; Roubinet, F.; Blancher, A.; Calafell, F.; Bertranpetit, J.; Casals, F. A natural history of FUT2 polymorphism in humans. Mol. Biol. Evol. 2009, 26, 1993–2003.

82. Ward, L.D.; Kellis, M. HaploReg v4: Systematic mining of putative causal variants, cell types, regulators and target genes for human complex traits and disease. Nucleic Acids Res. 2015, 44, D877–D881.

83. Wishart, D.S.; Guo, A.; Oler, E.; Wang, F.; Anjum, A.; Peters, H.; Dizon, R.; Sayeeda, Z.; Tian, S.; Lee, B.L.; et al. HMDB 5.0: The Human Metabolome Database for 2022. Nucleic Acids Res. 2022, 50, D622–D631.

84. Kanehisa, M. Enzyme annotation and metabolic reconstruction using KEGG. Protein Funct. Predict. Methods Protoc. 2017, 1611, 135–145.

85. Saier, M.H.; Reddy, V.S.; Moreno-Hagelsieb, G.; Hendargo, K.J.; Zhang, Y.; Iddamsetty, V.; Lam, K.J.K.; Tian, N.; Russum, S.; Wang, J.; et al. The Transporter Classification Database (TCDB): 2021 update. Nucleic Acids Res. 2021, 49, D461–D467.

86. Brunk, E.; Sahoo, S.; Zielinski, D.C.; Altunkaya, A.; Dräger, A.; Mih, N.; Gatto, F.; Nilsson, A.; Preciat Gonzalez, G.A.; Aurich, M.K.; et al. Recon3D enables a three-dimensional view of gene variation in human metabolism. Nat. Biotechnol. 2018, 36, 272–281.

87. Szklarczyk, D.; Gable, A.L.; Lyon, D.; Junge, A.; Wyder, S.; Huerta-Cepas, J.; Simonovic, M.; Doncheva, N.T.; Morris, J.H.; Bork, P.; et al. STRING v11: Protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res. 2019, 47, D607–D613.

88. Breuer, K.; Foroushani, A.K.; Laird, M.R.; Chen, C.; Sribnaia, A.; Lo, R.; Winsor, G.L.; Hancock, R.E.; Brinkman, F.S.; Lynn, D.J. InnateDB: Systems biology of innate immunity and beyond—Recent updates and continuing curation. Nucleic Acids Res. 2013, 41, D1228–D1233.

89. Rolland, T.; Taşan, M.; Charloteaux, B.; Pevzner, S.J.; Zhong, Q.; Sahni, N.; Yi, S.; Lemmens, I.; Fontanillo, C.; Mosca, R.J.C. A proteome-scale map of the human interactome network. Cell 2014, 159, 1212–1226.

90. Lee, B.T.; Barber, G.P.; Benet-Pagès, A.; Casper, J.; Clawson, H.; Diekhans, M.; Fischer, C.; Gonzalez, J.N.; Hinrichs, A.S.; Lee, C.M.; et al. The UCSC Genome Browser database: 2022 update. Nucleic Acids Res. 2022, 50, D1115–D1122.

91. Purcell, S.; Neale, B.; Todd-Brown, K.; Thomas, L.; Ferreira, M.A.; Bender, D.; Maller, J.; Sklar, P.; de Bakker, P.I.; Daly, M.J.; et al. PLINK: A tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. 2007, 81, 559–575.

92. Akhmedov, M.; Kedaigle, A.; Chong, R.E.; Montemanni, R.; Bertoni, F.; Fraenkel, E.; Kwee, I. PCSF: An R-package for network-based interpretation of high-throughput data. PLoS Comput. Biol. 2017, 13, e1005694.

93. Csardi, G.; Nepusz, T.J.I. The igraph software package for complex network research. InterJournal Complex Syst. 1695, 5, 1–9.

94. UniProt Consortium, T. UniProt: The universal protein knowledgebase. Nucleic Acids Res. 2018, 46, 2699.

95. Landrum, M.J.; Kattman, B.L. ClinVar at five years: Delivering on the promise. Hum. Mutat. 2018, 39, 1623–1630.

96. MacArthur, J.; Bowler, E.; Cerezo, M.; Gil, L.; Hall, P.; Hastings, E.; Junkins, H.; McMahon, A.; Milano, A.; Morales, J.; et al. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). Nucleic Acids Res. 2017, 45, D896–D901.

97. Li, M.J.; Liu, Z.; Wang, P.; Wong, M.P.; Nelson, M.R.; Kocher, J.P.; Yeager, M.; Sham, P.C.; Chanock, S.J.; Xia, Z.; et al. GWASdb v2: An update database for human genetic variants identified by genome-wide association studies. Nucleic Acids Res. 2016, 44, D869–D876.

98. Luck, K.; Kim, D.K.; Lambourne, L.; Spirohn, K.; Begg, B.E.; Bian, W.; Brignall, R.; Cafarelli, T.; Campos-Laborie, F.J.; Charloteaux, B.; et al. A reference map of the human binary protein interactome. Nature 2020, 580, 402–408.

99. Boyle, E.I.; Weng, S.; Gollub, J.; Jin, H.; Botstein, D.; Cherry, J.M.; Sherlock, G. GO::TermFinder—Open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. Bioinformatics 2004, 20, 3710–3715.

Bridging statement to Chapter 3

In transitioning from Chapter 2 to Chapter 3, I build upon our understanding of the associations between genetic variants and metabolites by delving deeper into gene regulation. As we move from the metabolic level to the regulatory level, Chapter 3 introduces miRNet 2.0, a miRNA-centric network visual analytics platform. This platform enables us to explore complex regulatory interaction networks, integrating both experimentally validated and computationally predicted evidence. While Chapter 2 provided insights into the direct associations between genetic variants and metabolites, Chapter 3 will enhance our understanding of the broader regulatory landscape, including miRNA-target interactions and miRNA-disease associations. In doing so, miRNet 2.0 offers a complementary perspective to our investigations in Chapter 2, enriching our overall understanding of the genetics of metabolism and regulatory interactions.

Chapter 3: miRNet 2.0: network-based visual analytics for miRNA functional analysis and systems biology

[1]Le Chang, [2]Guangyan Zhou, [2]Othman Soufan and [1,2*]Jianguo Xia

[1]Department of Human Genetics, McGill University, Montreal, QC H3A 0C7, Canada;

[2]Institute of Parasitology, McGill University, Montreal, QC H9X 3V9, Canada

*To whom correspondence should be addressed:

Tel: 1-514-398-8668

Email: jeff.xia@mcgill.ca

Status: Manuscript published in Nucleic Acids Research: https://doi.org/10.1093/nar/gkaa467

Abstract

miRNet is an easy-to-use, web-based platform designed to help elucidate microRNA (miRNA) functions by integrating users' data with existing knowledge via network-based visual analytics. Since its first release in 2016, miRNet has been accessed by >20 000 researchers worldwide, with ~100 users on a daily basis. While version 1.0 was focused primarily on miRNA-target gene interactions, it has become clear that in order to obtain a global view of miRNA functions, it is necessary to bring other important players into the context during analysis. Driven by this concept, in miRNet version 2.0, we have (i) added support for transcription factors (TFs) and single nucleotide polymorphisms (SNPs) that affect miRNAs, miRNA-binding sites or target genes, whilst also greatly increased (>5-fold) the underlying knowledgebases of miRNAs, ncRNAs and disease associations; (ii) implemented new functions to allow creation and visual exploration of multipartite networks, with enhanced support for in situ functional analysis and (iii) revamped the web interface, optimized the workflow, and introduced microservices and web application programming interface (API) to sustain high-performance, real-time data analysis. The underlying R package is also released in tandem with version 2.0 to allow more flexible data analysis for R programmers. The miRNet 2.0 website is freely available at https://www.mirnet.ca.

Introduction

Gene expression in eukaryotes is a complex process controlled by many factors functioning at epigenetic, transcriptional or post-transcriptional levels (1,2). Over the past two decades, the broad applications of comprehensive molecular profiling technologies have enabled us to study gene expression in various biological processes and disease conditions. However, our understanding of the underlying regulatory mechanisms remains incomplete. It has become clear that in order to address this issue, it is critical to adopt a systems biology approach to integrate all important players (3–5). Network-based approaches have received wide attention as they can abstract and integrate different types of information in a format that is often intuitive and interpretable (6–8). Based on this concept, we developed miRNet version 1.0 to help illustrate the 'multiple-to-multiple' relationships (i.e. one miRNA can regulate multiple genes and one gene can be regulated by multiple miRNAs) through network-based visualization of miRNA-target gene interactions coupled with improved functional analysis (9,10). However, the interplays between miRNAs and target genes represent only the starting points toward understanding the roles that miRNAs play at cellular level. In particular, miRNAs can regulate gene regulatory networks through feedback or feedforward loops (11), for instance, by adjusting expression of transcription factors (TFs) which in turn exert effects on their corresponding target genes. Such higher-order interactions are not captured in miRNet version 1.0.

The past few years have witnessed several trends in miRNA research. A growing number of studies have employed systems biology approaches either experimentally by employing multi-omics measurements or computationally by including other key factors such as miRNA–lncRNA–gene (12), miRNA–TF–gene (4) or miRNA–gene–disease (13) to better understand miRNA regulatory mechanisms. Another growing area of research is precision medicine, in which the

characteristic gene expression patterns of a particular patient can be interpreted by his or her own genetic mutations to inform treatment or prevention plan (14). For instance, SNPs in miRNA and miRNA-binding sites have been found to be associated with several diseases (15). The complex interplays amongst different functional elements can be captured using multipartite networks to reveal a more holistic picture. However, integrating multiple data types and interpreting these results at a systems level is challenging (16). Building such networks requires manual curation of data from multiple databases and powerful network visualization support to aid researchers in navigation and understanding.

Since the release of miRNet version 1.0, many new features and components have been gradually introduced based on users' feedback and developments in the field. For instance, tissue and cellular contexts are important for interpretation of miRNA-gene interactions. To support this need, we have implemented tissue-specific filters based on their expression profiles (17). In addition, current bioinformatics tools focus primarily on human and other model organisms. To facilitate miRNA research in species important for agriculture and veterinary medicine, we have added support for cows, pigs and chickens following well-established protocols (18). For researchers interested in exploring potential regulatory roles of miRNAs derived from pathogens such as parasitic helminths (19), viruses (20) or other sources, we have added support for reported or putative xeno-miRNAs (21). A continuous effort has been to keep current with new releases of its underlying databases as well as to maintain backward compatibility. This effort has triggered several rounds of code refactoring to achieve a more robust and modular design, with computational intensive tasks distributed among different servers through microservices (22). The latter technique also helps address computational bottlenecks with bigger databases and growing user traffic. The user-friendly web interface is mainly used by clinicians and bench biologists with

little to no programming skills. While for bioinformaticians or tool developers, it may be more meaningful to directly access miRNet's functionality through its underlying R code or a well-defined application programming interface (API).

To address these emerging bioinformatics needs and challenges, we developed miRNet version 2.0 to allow users to easily create complex miRNA-centric networks for systems-level interpretation of miRNA functions and gene regulations. The 2.0 release captures all the aforementioned updates since 2016 and represents a solid step toward network-based data integration for miRNA systems biology. A more detailed description of each of these updates and changes in miRNet 2.0 is given below.

Program description and methods

Overview of miRNet 2.0 framework

The main workflow of miRNet 2.0 is summarized in Figure 1. There are three main steps—data input, network creation and network visual analytics. To maintain a flexible and modular design, we have organized the main functions into 12 modules based on input types. The 'miRNAs' module allows users to connect miRNAs with target genes, TFs, ncRNAs etc.; the 'Genes' and 'TFs' modules link the corresponding inputs to their partners within the context of known interactions among miRNAs, genes and TFs; the 'SNPs' module maps SNPs to the above key players themselves or their binding sites. The remaining modules follow a similar procedure by mapping users' inputs to their corresponding miRNA associated interaction partners. To start, users must click a circular button from the miRNet homepage to enter the corresponding data upload page. Two general data formats are accepted: a list of miRNAs, SNPs, genes, small molecules etc., or an expression table generated from qPCR, microarray or RNAseq experiments.

In the latter case, well-established differential expression analysis will be applied to identify significant miRNAs or genes as new input lists. In the second step, the input lists will be mapped to the underlying knowledgebases to create one or more interaction tables and networks. Many functions are available to allow users to further customize or refine the networks. In the third step, the results are presented as interactive networks for visual exploration. Users can easily search, zoom, highlight or perform functional enrichment analysis on selected regions of interest. In the following sections, we will focus primarily on the new and improved features introduced in version 2.0. Other features can be found in our prior publications (9,10,21).

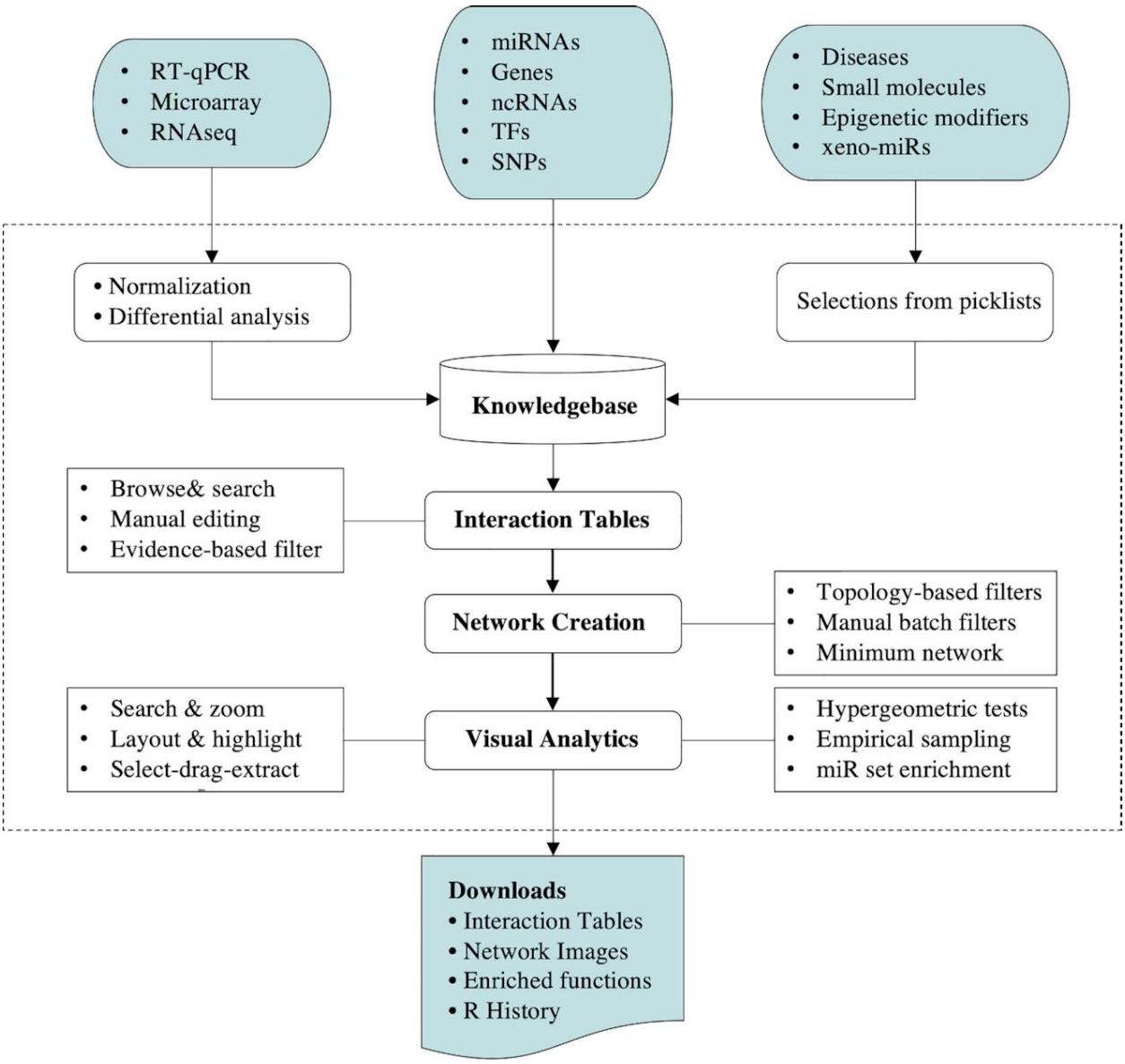

Figure 1. Overview of miRNet 2.0 workflow. Users can upload different data types or select queries from built-in databases to start analysis. The input will be mapped to the underlying knowledgebases to create interaction tables and networks. The visualization page allows users to intuitively explore the networks using different layout algorithms as well as to perform topology or functional analysis.

Knowledgebase update and creation

Knowledgebase for network creation. We have put considerable efforts into keeping miRNet's underlying knowledgebases up to date. miRNet 2.0 can automatically recognize different versions of miRBase IDs, as well as link pre-miRNAs to their mature forms based on the miRBaseConverter R package (23). We have updated the miRNA interaction knowledgebase based on the latest releases from major miRNA annotation databases including miRBase (24), miRTarBase (25), TarBase (26), HMDD (27) etc. The human tissue-specific miRNA annotations are based on TSmiR (28) and IMOTA (17) databases, and the human exosomal miRNA annotations are from ExoCarta (29). The interactions among miRNAs, TFs and genes are obtained from TransmiR 2.0 (30), ENCODE (31), JASPAR (32) and ChEA (33). For miR-SNPs, we have used ADmiRE (34), PolymiRTS (35) and SNP2TFBS (36) to obtain SNP information in miRNA genes, miRNA-binding sites and TF-binding sites. We have also systematically collected the reported xeno-miRNAs together with their putative targeted genes into xeno-miRNet (21), which is now integrated in miRNet 2.0. Finally, we have expanded the miRNA-lncRNA interactions to include all other major ncRNAs including circRNA, ceRNA, pseudogene and sncRNA based on starBase (37). These data can be downloaded from the miRNet 'Resources' page as plain text files.

Knowledgebase for network interpretation. For network analysis, it is important to be able to interpret the interactions in addition to their visualization. Enrichment analysis plays a significant role in this respect. Applying conventional enrichment analyses such as hypergeometric tests on target genes are known to be biased (38,39). In miRNet 1.0, we implemented an algorithm based on empirical sampling for enrichment analysis using GO, KEGG or Reactome pathways (38). Another effective approach is to perform enrichment analysis directly at miRNA levels (39). To support this type of analysis, we have added six miRNA-set libraries including miRNA–function, miRNA–disease, miRNA–TF, miRNA–cluster, miRNA–family and miRNA–tissue

based on TAM 2.0 (40). In summary, miRNet 2.0 provides four query types (all genes, highlighted genes, all miRNAs, highlighted miRNAs), two enrichment algorithms (hypergeometric tests and empirical sampling), nine annotation libraries (three gene-set libraries and six miRNA-set libraries), representing the most comprehensive support to understand collective functions of miRNAs. Their potential applications are showcased in recent studies to compare miRNA changes specific to different tissues in pancreatic ductal adenocarcinoma (41) and to identify enriched miRNA families in a study comparing genetic variants between Alzheimer's disease and cancers (42).

Enabling flexible user input

Significant efforts have been made to provide an intuitive interface that permits the integration of miRNAs into different types of interaction networks. From the homepage, users can enter their queries by: (a) uploading a list of miRNAs, ncRNAs, genes, TFs or SNPs; (b) selecting a list from our built-in databases such as diseases, small compounds, epigenetic modifiers etc. (c) uploading a miRNA or gene expression table generated from RT-qPCR, microarray or RNAseq or (d) uploading multiple queries of different input types. Here, we will introduce new features for several common scenarios.

From miRNAs to networks. In miRNet 1.0, miRNA–targets mapping was limited to target genes based on experimentally validated interaction information. However, increasing evidence has shown that miRNAs participate in complex networks through interactions with other functional elements to exert effects on cell biology and human diseases (12). For instance, lncRNAs can act as miRNA ‘sponge’ and compete with target mRNAs, thus increasing the expression level of mRNAs (43). In version 2.0, users can select one or multiple targets from the

‘Targets’ dropdown list and miRNet will automatically map miRNAs to those selected targets. Users can further include protein-protein interactions (PPI) in the target networks based on several well-established PPI databases (44–46).

From TFs to networks. miRNAs and TFs can cooperate to tune gene expression, or mutually regulate each other in feedback loops (4,47). Consequently, we have added a new module to allow users to include TFs into analysis. Users can simply upload their TF list, miRNet will automatically map the TFs to all potential targets (miRNAs and/or genes) and return as TF–miRNA and/or TF–gene interaction tables. The interactions will then be further integrated into networks for visual exploration. With the updated miRNA module and the addition of the TF module, miRNet 2.0 allows users to easily create miRNA-TF coregulatory networks from either a list of miRNAs or a list of TFs of interest.

From SNPs to networks. Mutations in mature miRNAs or their binding sites could significantly change their targeting abilities and dysregulate the expression of many genes simultaneously, whereas variations in primary or precursor miRNAs could alter the expression levels of mature miRNAs by affecting miRNA processing (48,49). In miRNet 2.0, we have added a new module to support the analysis of SNPs within the context of miRNA-target gene interactions. Users can upload a list of SNPs from the SNPs upload page. miRNet currently accepts either rsIDs or genomic coordinates based on the human reference genome build GRCh37. The uploaded lists are then mapped to miRNAs and/or their target genes. Following this step, users can visually explore their data in the network visualization page.

Uploading multiple queries. The Multiple Query Types module complements miRNet's single type analysis modules by permitting the identification of novel connections amongst multiple types of user input. The module currently supports ten input types shown in a dialog when

users click the central circular button at the home page. After selecting the input types of interest, users simply copy-and-paste their query lists (miRNAs, genes, TFs, lncRNAs, pseudogenes, circRNAs, sncRNAs) or select from picklists (diseases, small compounds and epigenetic modifiers). The uploaded lists are then mapped to the internal knowledgebases and proceed with the workflow as described in other modules.

Enhancing network visual analytics

Network creation and customization. The default networks are created by searching for direct interaction partners in the interaction knowledgebases. These are generally known as first-order interaction networks. When there is a large number of queries (seeds), it is reasonable to focus only on the interactions among those seeds (i.e. zero-order networks). However, many seeds could become orphan nodes when switching directly to zero-order networks. A 'gentle' approach is to extract, from the first-order network, a minimal subnetwork that maximally connects those seeds. In miRNet 2.0, we have added the support for computing minimum subnetworks based on the prize-collecting Steiner Forest (PCSF) algorithm (50), as well as several other empirical refining methods (available under 'Network Tools') based on shortest paths, batch filtering, node degree or betweenness values. The results can be downloaded as pair-wise interaction tables or graph files.

Network visualization and layout. miRNet 2.0 provides a wide array of options to help improve visual exploration of miRNA-centric interaction networks. During the network creation stage, users can refine the network by applying different filters on interaction tables or networks. At the network visualization page, users can specify node styles based on their types, reduce node overlap, or perform edge bundling etc. The resulting network can be further improved using

different layout algorithms. Over ten network layout algorithms have been implemented, including Force-Atlas, Fruchterman-Reingold, Circular, Graphopt, Large Graph, Random, Circular Bipartite/Tripartite, Linear Bipartite/Tripartite, Concentric and Backbone. The latter four algorithms are designed for complex networks consisting of multiple node types (miRNAs, genes, TFs etc.). The bipartite/tripartite layout provides a straightforward abstraction of the relationships between different types of molecular entities by emphasizing the data type of each node (51). When there are multiple node types, we recommend visualizing the network in either circular bipartite/tripartite (Figure 2A) or linear bipartite/tripartite layout (Figure 2B) followed by applying the 'reduce node overlap' algorithm. To enable better understanding of a particular key node, we have added the Concentric layout (52). This layout arranges nodes in concentric circles around a node of interest (i.e. the focal node) in the middle (Figure 2C). The order of the circles represents the degree level of their interactions. By arranging nodes in this fashion, it enables a better understanding of how the focal node relates to the rest of the graph. By default, the focal node is the node with the highest degree value. Users can manually specify the key node by selecting it in the Node Explorer table or by double clicking on it in the network. Another new addition is the Backbone layout which is very effective in revealing hidden patterns in medium and large networks. The algorithm calculates layout after applying sparsification on the network by only including the most embedded edges (53). This process helps uncover hidden modules based on edge density by putting more emphasis on the structure of graph layout (Figure 2D).

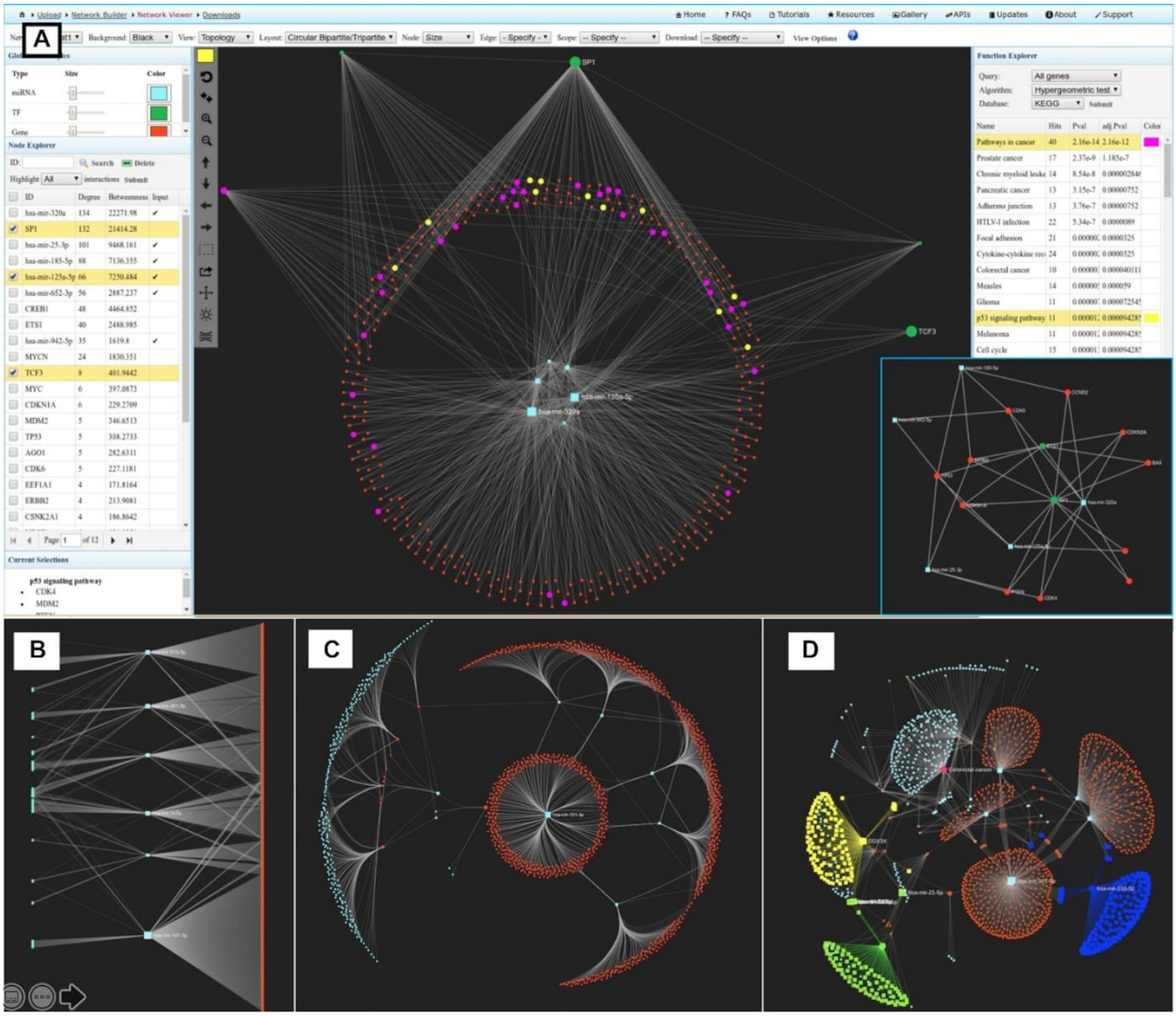

Figure 2. Screenshots of the Network Visualization page showing the main features and several network layouts. (A) A typical view of the page. The central panel shows a network in Circular-tripartite layout, and the surrounding panels provide functions for network analysis and customization. For instance, users can perform enrichment analysis or module analysis on this network. An extracted network module was displayed at bottom right. (B) Linear-tripartite layout. (C) Concentric layout with edge bundling. (D) Backbone layout with several modules highlighted in different colors. More details of each layout are described in the main text.

Improving transparency/reproducibility and web APIs

Except for the interactive visualization step, which is executed on users' browsers, all other data analysis steps including mapping, filtering, network creation and customization are performed by the corresponding R functions on our cloud server. To enable more transparent data analysis, we have released the underlying R package (https://github.com/xia-lab/miRNetR), and added a 'Download' page in the web application to allow users to download the R command history and results tables generated during their analysis sessions. The R history contains all function calls with user-selected parameters. We hope that the R package together with the R command history will allow users to track each step of their analysis in a form (R script) that can be easily shared and reproduced, complementing the web-based platform. We have also implemented RESTful APIs to allow tool developers to submit their query lists programmatically as external requests. While offering open access to miRNet 2.0 resources, APIs give a level of abstraction and hide complexity from programmers. The currently available APIs are shown in Table 1. More APIs will be added based on users' feedback.

Table 1. List of APIs and programmatic access endpoints on the miRNet server. The API base for miRNet 2.0 is http://api.mirnet.ca, which can be visited to view a detailed documentation

Endpoint	HTTP method	Input	Description
base/table/mir	POST	Organism, miRNA ID type, target type, miRNA list	Get experimentally validated table results of the miRNA-target interactions (forward mapping)
base/table/gene	POST	Organism, gene ID type, gene list	Get experimentally validated table results of the miRNA-gene (mRNA, TF, lncRNA) interactions (reverse mapping)

base/function/mir	POST	Organism, miRNA ID type, target type, miRNA list, algorithm, database	Get functional enrichment results
base/function/gene	POST	Organism, gene ID type, gene list, algorithm, database	Get functional enrichment results
base/graph/mir	POST	Organism, miRNA ID type, target type, miRNA list	Get graph of miRNA-target interactions (json format)
base/graph/gene	POST	Organism, gene ID type, gene list	Get graph of miRNA-target interactions (json format)

Use cases

To further demonstrate the utility of these new features in miRNet 2.0, we used data from a multiple sclerosis (MS) study aiming to identify the role of miRNA and TF co-regulatory networks in the pathogenesis of MS (54). In this study, the miRNA target analysis and TF target analysis were performed by searching several miRNA–gene, TF–gene and TF–miRNA databases. The network was manually built by using Cytoscape (55). We reconstructed and visualized the network in miRNet 2.0 by using the miRNA module. The resulting network is comprised of 2414 nodes (TF: 5; Gene: 2403; miRNA: 6) and 2798 edges. For better visual exploration, a degree cutoff 1.0 was applied. As shown in Figure 2A, the TF-miRNA co-regulatory networks is displayed at the center of the Network Viewer page in Circular Tripartite layout. It illustrates various interactions between miRNAs (inner zone), genes (middle layer) as well as TFs (outer layer). The nodes are sorted by degree centrality measures in the Node Explorer table. In this case, miRNet 2.0 confirmed the detection of important nodes according to their degree measures. Among the top nodes, hsa-miR-125a-5p (degree = 66) has been frequently associated with MS, while SP1 (degree = 132) and TCF3 (degree = 8) have been reported in the transcriptional regulations of MS (54). We also performed functional enrichment and module analysis on the

whole network. The results of functional enrichment analysis using KEGG database are displayed in the Function Explorer table. For instance, cytokine-cytokine receptor interaction pathway (adj. P-value = 3.25×10^{-5}) and p53 signaling pathway (adj. P-value = 9.43×10^{-5}) were significantly enriched, which were not reported by the original study but the results have been supported by other publications (56). Figure 2A (lower right corner) shows an example of an extracted module (p53 signaling pathway) after manual increase of the edge thickness. Compared to the original network, this module is much more digestible while keeping the important nodes and connections (e.g. hsa-miR-125a-5p and SP1). This use case highlights that with only a few mouse clicks, users can easily create comprehensive regulatory networks to gain a more holistic view of miRNA-mediated regulations as well as to extract important modules for more in-depth analysis.

Comparison with miRNet 1.0 and other web-based tools

Several excellent web-based tools have been developed for miRNA network analysis, including miRTargetLink (57), MIENTURNET (58), Arena-Idb (59), and starBase (37). Detailed comparison between these tools and miRNet 2.0, as well as its previous version is shown in Table 2. Particularly, miRTargetLink, MIENTURNET and Arena-Idb can assist researchers in understanding miRNAs and their targets through a network-based visualization method based on predicted or experimentally validated miRNA–target interactions; while starBase is the most comprehensive miRNA–mRNA and miRNA–ncRNA interaction database based on CLIP-Seq experiments. In comparison, miRNet 2.0 is a high-performance, easy-to-use web application which offers the most comprehensive support for real-time, interactive miRNA network analytics in ways that no other tools currently can. More than 15 databases and over 10 graph layout algorithms have been integrated to facilitate knowledge discovery and hypothesis generation. The companion R

package and APIs have been developed to permit transparent and reproducible analysis as well as to reach a broader user base. In summary, miRNet 2.0 caters for both bench researchers as well as bioinformaticians by providing an interactive and integrative platform for miRNA-centric systems biology.

Table 2. Comparison of the main features of miRNet (versions 1.0–2.0) with other web-based or web-enabled tools. Symbols used for feature evaluations with '√' for present, '−' for absent, and '+' for a more quantitative assessment (more '+' indicate better support)

Tool name	miRNet		miRTargetLink	MIENTURNET	Arena-Idb	starBase
	2.0	1.0				
Data processing						
Species #	10	7	1	6	1	23
Target genes						
Experimental	+++	++	++	+	++	+++
Predicted	√	√	√	√	√	√
Others						
miR-SNPs	√	-	-	-	-	-
TFs	√	-	-	-	-	-
ncRNAs	+++	+	-	-	+	++++
xeno-miRNAs	√	-	-	-	-	-
Diseases	+++	++	-	-	++	++
Epigenetic modifiers	√	√	-	-	-	-
Small compounds	√	√	-	-	-	-
Expression profiling	√	√	-	-	-	-
Enrichment analysis						
Hypergeometric tests	√	√	√	√	-	√
Empirical sampling	√	√	-	-	-	-
miR-set enrichment	√	-	-	-	-	-
Network visual analytics						
Multiple query types	√	-	-	-	-	-
Integration with PPI network	√	-	-	-	-	-
Multipartite network visualization	√	-	-	-	-	-
Subnetwork extraction	√	-	-	-	-	-

URL links:

miRTargetLink: https://ccb-web.cs.uni-saarland.de/mirtargetlink/

MIENTURNET: http://userver.bio.uniroma1.it/apps/mienturnet/

Arena-Idb: http://ncrnadb.scienze.univr.it/sites/arenaidb/

starBase: http://starbase.sysu.edu.cn/index.php

Conclusions

Over the past few years, miRNA research has gradually evolved from target identification toward understanding the regulatory mechanisms underlying their systems level effects. However, very few user-friendly bioinformatics tools are available to support this objective. To address this gap, we have developed miRNet version 2.0 to assist researchers to easily create miRNA-centric multiplex networks integrating key players involved in gene regulation as well as other molecules of interest. During this process, we have greatly expanded the underlying knowledgebases and added new libraries on TFs, SNPs, ncRNAs and PPIs to provide a rich context for analysis, hypothesis generation and mechanistic insights. We have also implemented new graph mining functions and layout algorithms tailored to complex multipartite network creation, customization and visualization. To sustain real-time intuitive data analysis, we have completely revamped the web interface, optimized the workflow, introduced APIs and microservices to enable high-performance computing and visualization. A limitation of miRNet is its static and qualitative nature of the current network analysis. It is important to keep in mind that miRNA functions are highly dependent on the context (abundance, location, cell type, cell state etc.) and the effects can be dynamic and transient to confer robustness to biological processes (60). We believe that miRNet

2.0 will continue to be an invaluable bioinformatics asset for researchers in miRNA systems biology.

Data availability

The miRNet 2.0 web server can be freely accessed at https://www.mirnet.ca. The web APIs can be accessed from http://api.mirnet.ca. The miRNetR is available on Github (https://github.com/xia-lab/miRNetR).

Funding

Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant, NSERC-CREATE-MATRIX Scholarship, Genome Canada, Genome Quebec and Canada Research Chairs (CRC) Program. Funding for open access charge: Genome Canada.

Conflict of interest statement. None declared.

References

1. Orphanides,G. and Reinberg,D. (2002) A unified theory of gene expression. Cell, 108, 439–451.

2. Herranz,H. and Cohen,S.M. (2010) MicroRNAs and gene regulatory networks: managing the impact of noise in biological systems. Genes Dev., 24, 1339–1344.

3. Arora,S., Rana,R., Chhabra,A., Jaiswal,A. and Rani,V. (2013) miRNA-transcription factor interactions: a combinatorial regulation of gene expression. Mol. Genet. GENOMics: MGG, 288, 77–87.

4. Bracken,C.P., Scott,H.S. and Goodall,G.J. (2016) A network-biology perspective of microRNA function and dysfunction in cancer. Nat. Rev. Genet., 17, 719–732.

5. Zhang,H.M., Kuang,S., Xiong,X., Gao,T., Liu,C. and Guo,A.Y. (2015) Transcription factor and microRNA co-regulatory loops: important regulatory motifs in biological processes and diseases. Brief. Bioinform., 16, 45–58.

6. Barabasi,A.L., Gulbahce,N. and Loscalzo,J. (2011) Network medicine: a network-based approach to human disease. Nat. Rev. Genet., 12, 56–68.

7. Macneil,L.T. and Walhout,A.J. (2011) Gene regulatory networks and the role of robustness and stochasticity in the control of gene expression. Genome Res., 21, 645–657.

8. Barabasi,A.L. and Oltvai,Z.N. (2004) Network biology: understanding the cell's functional organization. Nat. Rev. Genet., 5, 101–113.

9. Fan,Y., Siklenka,K., Arora,S.K., Ribeiro,P., Kimmins,S. and Xia,J. (2016) miRNet - dissecting miRNA-target interactions and functional associations through network-based visual analysis. Nucleic Acids Res., 44, W135–W141.

10. Fan,Y. and Xia,J. (2018) miRNet-Functional analysis and visual exploration of miRNA-Target interactions in a network context. Methods Mol. Biol., 1819, 215–233.

11. Lai,X., Wolkenhauer,O. and Vera,J. (2016) Understanding microRNA-mediated gene regulatory networks through mathematical modelling. Nucleic Acids Res., 44, 6019–6035.

12. Anastasiadou,E., Jacob,L.S. and Slack,F.J. (2018) Non-coding RNA networks in cancer. Nat. Rev. Cancer, 18, 5–18.

13. Jin,S., Zeng,X., Fang,J., Lin,J., Chan,S.Y., Erzurum,S.C. and Cheng,F. (2019) A network-based approach to uncover microRNA-mediated disease comorbidities and potential pathobiological implications. NPJ Syst. Biol. Applic., 5, 41.

14. Detassis,S., Grasso,M., Del Vescovo,V. and Denti,M.A. (2017) microRNAs make the call in cancer personalized medicine. Front. Cell Dev. Biol., 5, 86.

15. Fehlmann,T., Sahay,S., Keller,A. and Backes,C. (2019) A review of databases predicting the effects of SNPs in miRNA genes or miRNA-binding sites. Brief. Bioinform., 20, 1011–1020.

16. Lai,X., Eberhardt,M., Schmitz,U. and Vera,J. (2019) Systems biology-based investigation of cooperating microRNAs as monotherapy or adjuvant therapy in cancer. Nucleic Acids Res., 47, 7753–7766.

17. Palmieri,V., Backes,C., Ludwig,N., Fehlmann,T., Kern,F., Meese,E. and Keller,A. (2018) IMOTA: an interactive multi-omics tissue atlas for the analysis of human miRNA-target interactions. Nucleic Acids Res., 46, D770–D775.

18. 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.

19. Hoy,A.M., Lundie,R.J., Ivens,A., Quintana,J.F., Nausch,N., Forster,T., Jones,F., Kabatereine,N.B., Dunne,D.W., Mutapi,F. et al. (2014) Parasite-derived microRNAs in host serum as novel biomarkers of helminth infection. PLoS Negl. Trop. Dis., 8, e2701.

20. Cullen,B.R. (2013) MicroRNAs as mediators of viral evasion of the immune system. Nat. Immunol., 14, 205–210.

21. Fan,Y., Habib,M. and Xia,J. (2018) Xeno-miRNet: a comprehensive database and analytics platform to explore xeno-miRNAs and their potential targets. PeerJ, 6, e5650.

22. Williams,C.L., Sica,J.C., Killen,R.T. and Balis,U.G. (2016) The growing need for microservices in bioinformatics. J. Pathol. Informatics, 7, 45.

23. Xu,T., Su,N., Liu,L., Zhang,J., Wang,H., Zhang,W., Gui,J., Yu,K., Li,J. and Le,T.D. (2018) miRBaseConverter: an R/Bioconductor package for converting and retrieving miRNA name, accession, sequence and family information in different versions of miRBase. BMC Bioinformatics, 19, 514.

24. Kozomara,A., Birgaoanu,M. and Griffiths-Jones,S. (2019) miRBase: from microRNA sequences to function. Nucleic Acids Res., 47, D155–D162.

25. Huang,H.Y., Lin,Y.C., Li,J., Huang,K.Y., Shrestha,S., Hong,H.C., Tang,Y., Chen,Y.G., Jin,C.N., Yu,Y. et al. (2020) miRTarBase 2020: updates to the experimentally validated microRNA-target interaction database. Nucleic Acids Res., 48, D148–D154.

26. Karagkouni,D., Paraskevopoulou,M.D., Chatzopoulos,S., Vlachos,I.S., Tastsoglou,S., Kanellos,I., Papadimitriou,D., Kavakiotis,I., Maniou,S., Skoufos,G. et al. (2018) DIANA-TarBase v8: a decade-long collection of experimentally supported
miRNA-gene interactions. Nucleic Acids Res., 46, D239–D245.

27. Huang,Z., Shi,J., Gao,Y., Cui,C., Zhang,S., Li,J., Zhou,Y. and Cui,Q. (2019) HMDD v3.0: a database for experimentally supported human microRNA-disease associations. Nucleic Acids Res., 47, D1013–D1017.

28. Guo,Z., Maki,M., Ding,R., Yang,Y., Zhang,B. and Xiong,L. (2014) Genome-wide survey of tissue-specific microRNA and transcription factor regulatory networks in 12 tissues. Sci. Rep., 4, 5150.

29. Mathivanan,S. and Simpson,R.J. (2009) ExoCarta: a compendium of exosomal proteins and RNA. Proteomics, 9, 4997–5000.

30. Tong,Z., Cui,Q., Wang,J. and Zhou,Y. (2019) TransmiR v2.0: an updated transcription factor-microRNA regulation database. Nucleic Acids Res., 47, D253–D258.

31. ENCODE Consortium (2004) The ENCODE (ENCyclopedia Of DNA Elements) Project. Science (New York, N.Y.), 306, 636–640.

32. Fornes,O., Castro-Mondragon,J.A., Khan,A., van der Lee,R., Zhang,X., Richmond,P.A., Modi,B.P., Correard,S., Gheorghe,M., Baranasic,D. et al. (2020) JASPAR 2020: update of the open-access database of transcription factor binding profiles. Nucleic Acids Res., 48, D87–D92.

33. Lachmann,A., Xu,H., Krishnan,J., Berger,S.I., Mazloom,A.R. and Ma'ayan,A. (2010) ChEA: transcription factor regulation inferred from integrating genome-wide ChIP-X experiments. Bioinformatics, 26, 2438–2444.

34. Oak,N., Ghosh,R., Huang,K.L., Wheeler,D.A., Ding,L. and Plon,S.E. (2019) Framework for microRNA variant annotation and prioritization using human population and disease datasets. Hum. Mutat., 40, 73–89.

35. Bhattacharya,A., Ziebarth,J.D. and Cui,Y. (2014) PolymiRTS Database 3.0: linking polymorphisms in microRNAs and their target sites with human diseases and biological pathways. Nucleic Acids Res., 42, D86–D91.

36. Kumar,S., Ambrosini,G. and Bucher,P. (2017) SNP2TFBS - a database of regulatory SNPs affecting predicted transcription factor binding site affinity. Nucleic Acids Res., 45, D139–D144.

37. Li,J.H., Liu,S., Zhou,H., Qu,L.H. and Yang,J.H. (2014) starBase v2.0: decoding miRNA-ceRNA, miRNA-ncRNA and protein-RNA interaction networks from large-scale CLIP-Seq data. Nucleic Acids Res., 42, D92–D97.

38. Bleazard,T., Lamb,J.A. and Griffiths-Jones,S. (2015) Bias in microRNA functional enrichment analysis. Bioinformatics, 31, 1592–1598.

39. Godard,P. and van Eyll,J. (2015) Pathway analysis from lists of microRNAs: common pitfalls and alternative strategy. Nucleic Acids Res., 43, 3490–3497.

40. Li,J., Han,X., Wan,Y., Zhang,S., Zhao,Y., Fan,R., Cui,Q. and Zhou,Y. (2018) TAM 2.0: tool for MicroRNA set analysis. Nucleic Acids Res., 46, W180–W185.

41. Chhatriya,B., Mukherjee,M., Ray,S., Sarkar,P., Chatterjee,S., Nath,D., Das,K. and Goswami,S. (2019) Comparison of tumour and serum specific microRNA changes dissecting their role in pancreatic ductal adenocarcinoma: a meta-analysis. BMC Cancer, 19, 1175.

42. Pathak,G.A., Zhou,Z., Silzer,T.K., Barber,R.C. and Phillips,N.R. (2020) Two-stage Bayesian GWAS of 9576 individuals identifies SNP regions that are targeted by miRNAs inversely expressed in Alzheimer's and cancer. Alzheimer's Dementia, 16, 162–177.

43. Paci,P., Colombo,T. and Farina,L. (2014) Computational analysis identifies a sponge interaction network between long non-coding RNAs and messenger RNAs in human breast cancer. BMC Syst. Biol., 8, 83.

44. Szklarczyk,D., Morris,J.H., Cook,H., Kuhn,M., Wyder,S., Simonovic,M., Santos,A., Doncheva,N.T., Roth,A., Bork,P. et al. (2017) The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible.
Nucleic Acids Res., 45, D362–D368.

45. Breuer,K., Foroushani,A.K., Laird,M.R., Chen,C., Sribnaia,A., Lo,R., Winsor,G.L., Hancock,R.E., Brinkman,F.S. and Lynn,D.J. (2013) InnateDB: systems biology of innate immunity and beyond–recent updates and continuing curation. Nucleic Acids Res., 41, D1228–D1233.

46. Rolland,T., Tasan,M., Charloteaux,B., Pevzner,S.J., Zhong,Q., Sahni,N., Yi,S., Lemmens,I., Fontanillo,C., Mosca,R. et al. (2014) A proteome-scale map of the human interactome network. Cell, 159, 1212–1226.

47. John,J.P., Thirunavukkarasu,P., Ishizuka,K., Parekh,P. and Sawa,A. (2019) An in-silico approach for discovery of microRNA-TF regulation of DISC1 interactome mediating neuronal migration. NPJ Syst. Biol. Applic., 5, 17.

48. Roden,C., Gaillard,J., Kanoria,S., Rennie,W., Barish,S., Cheng,J., Pan,W., Liu,J., Cotsapas,C., Ding,Y. et al. (2017) Novel determinants of mammalian primary microRNA processing revealed by systematic evaluation of hairpin-containing transcripts and human genetic variation. Genome Res., 27, 374–384.

49. Ryan,B.M., Robles,A.I. and Harris,C.C. (2010) Genetic variation in microRNA networks: the implications for cancer research. Nat. Rev. Cancer, 10, 389–402.

50. Akhmedov,M., Kedaigle,A., Chong,R.E., Montemanni,R., Bertoni,F., Fraenkel,E. and Kwee,I. (2017) PCSF: An R-package for network-based interpretation of high-throughput data. PLoS Comput. Biol., 13, e1005694.

51. Pavlopoulos,G.A., Kontou,P.I., Pavlopoulou,A., Bouyioukos,C., Markou,E. and Bagos,P.G. (2018) Bipartite graphs in systems biology and medicine: a survey of methods and applications. GigaScience, 7, giy014.

52. Brandes,U. and Pich,C. (2011) More flexible radial layout. J. Graph Algorithms Appl., 15, 157–173.

53. Nocaj,A., Ortmann,M. and Brandes,U.J. (2015) Untangling the hairballs of multi-centered, small-world online social media networks. J. Graph Algorithms Appl., 19, 595–618.

54. Nuzziello,N., Vilardo,L., Pelucchi,P., Consiglio,A., Liuni,S., Trojano,M. and Liguori,M. (2018) Investigating the role of MicroRNA and transcription factor Co-regulatory networks in multiple sclerosis pathogenesis. Int. J. Mol. Sci., 19, 3652.

55. Shannon,P., Markiel,A., Ozier,O., Baliga,N.S., Wang,J.T., Ramage,D., Amin,N., Schwikowski,B. and Ideker,T. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res., 13, 2498–2504.

56. Luo,D. and Fu,J. (2018) Identifying characteristic miRNAs-genes and risk pathways of multiple sclerosis based on bioinformatics analysis. Oncotarget, 9, 5287–5300.

57. Hamberg,M., Backes,C., Fehlmann,T., Hart,M., Meder,B., Meese,E. and Keller,A. (2016) MiRTargetLink–miRNAs, genes and interaction networks. Int. J. Mol. Sci., 17, 564.

58. Licursi,V., Conte,F., Fiscon,G. and Paci,P. (2019) MIENTURNET: an interactive web tool for microRNA-target enrichment and network-based analysis. BMC Bioinformatics, 20, 545.

59. Bonnici,V., Caro,G., Constantino,G., Liuni,S., D'Elia,D., Bombieri,N., Licciulli,F. and Giugno,R. (2018) Arena-Idb: a platform to build human non-coding RNA interaction networks. BMC Bioinformatics, 19, 350.

60. Ebert,M.S. and Sharp,P.A. (2012) Roles for microRNAs in conferring robustness to biological processes. Cell, 149, 515–524.

Bridging statement to Chapter 4

Building upon my exploration of genetics of metabolism and gene regulations in Chapters 2&3, Chapter 4 takes a further step by focusing on the potential causal relationships between metabolites and diseases. I present mGWAS-Explorer 2.0, an upgraded platform that leverages mGWAS and GWAS summary statistics to identify these relationships. This chapter integrates the knowledgebase acquired from Chapter 2 and the visual analytics system from Chapter 3. The mGWAS-Explorer 2.0 implements the two-sample Mendelian randomization approach, using genetic variants as instrumental variables. This method allows us to examine causal relationships between exposure risk factors, such as metabolites, and outcomes in observational studies. With this platform, we can now choose robustly associated instrumental variables and select the outcome data to further our understanding of the interplay between metabolites and diseases.

Chapter 4: mGWAS-Explorer 2.0: a web-based platform for prioritizing metabolites with causal impact on disease phenotypes

[1]Le Chang, [2]Guangyan Zhou and [1,2*]Jianguo Xia

[1]Department of Human Genetics, McGill University, Montreal, QC H3A 0C7, Canada;

[2]Institute of Parasitology, McGill University, Montreal, QC H9X 3V9, Canada

*To whom correspondence should be addressed:

Tel: 1-514-398-8668

Email: jeff.xia@mcgill.ca

Status: Manuscript in preparation

Abstract

mGWAS-Explorer is an integrated web-based platform developed to explore results from metabolome genome-wide association studies (mGWAS) via network visual analytics. While version 1.0 was primarily focused on linking single-nucleotide polymorphisms (SNPs), metabolites, genes and diseases for hypothesis generation, it has become obvious that establishing causal relationships between metabolites and diseases is crucial for understanding molecular etiology of diseases. Motivated by this concept, three major enhancements have been introduced in mGWAS-Explorer version 2.0, which include: (i) implemented two-sample Mendelian randomization analysis for causal inference; (ii) curated comprehensive molecular QTL data and semantic triples evidence for improved annotation; (iii) released the underlying R package for reproducible data analysis. The utility of mGWAS-Explorer 2.0 is demonstrated in two case studies.

Introduction

The circulating metabolites can act as inputs, mediators or products in the metabolic networks and play key roles in human health [1]. Over the past 15 years, wide applications of the metabolomics technologies in genome-wide association studies (mGWAS) have revealed a wealth of statistical associations between metabolites and single-nucleotide polymorphisms (SNPs) [2-5]. Meanwhile, a variety of genotype-phenotype association data and biochemistry knowledge are now readily available for understanding the impact of SNPs as well as related enzymes or transporters for metabolites. However, it is challenging to harmonize and integrate data from difference sources to interpret the results at a system level and gain mechanistic insights. To address this challenge, we developed mGWAS-Explorer version 1.0 to allow users to visually explore known connections among SNPs, genes, metabolites and diseases and to perform cross-phenotype association analysis for functional insights [6]. However, this is only the first step toward a thorough understanding of mGWAS and does not necessarily imply causal relationships.

In recent years, causal inference has emerged as a popular method within the mGWAS community, with Mendelian randomization (MR) becoming a well-established technique for this purpose [7-9]. MR leverages genetic variants as anchors to assess causal relationships between an exposure and an outcome, offering a reduced susceptibility to confounding and reverse causality compared to traditional observational research [10]. This approach is predicated on several assumptions, including that genetic variants do not influence the disease outcome through any process other than the relevant exposure [11]. As MR analysis has grown in prominence, numerous computational methods, databases, and tools have been developed to support this work [12-15], with two-sample Mendelian randomization (2SMR) emerging as a particularly useful method for inferring causal links from GWAS summary statistics [16]. This approach relies on separate data

sources for associations between genetic variant-exposure and genetic variant-outcome, with databases and analytical platforms like IEU OpenGWAS and MR-Base web application facilitating rapid application of two-sample MR approaches using large-scale GWAS study results [12-17].

MR has also proven effective in estimating the causal effects of metabolites on diseases or other phenotypes by using metabolite quantitative trait loci (mQTLs) as genetic instruments [10,12,18]. For example, MR studies have identified the causal role of low-density lipoprotein cholesterol (LDL-C) in coronary artery disease (CAD), leading to the discovery of LDL-C-lowering drugs [19,20]. Despite these advances, there is currently no dedicated bioinformatics tool for metabolome-wide MR mapping, necessitating additional resources for researchers seeking causal insights between genetically influenced metabolites and diseases. This gap highlights the urgent need for accessible bioinformatics tools to support MR analysis in mGWAS.

Addressing the challenges of interpreting causal assessments by MR methods can be achieved by combining causal estimates with information derived from semantic triples (subject-predicate-object), such as “homocysteine – predisposes – coronary arteriosclerosis” [21]. This triangulation approach strengthens the evidence for causation when different sources are in agreement and facilitates more robust causal interpretations by leveraging literature-mined knowledge from resources like SemMedDB and MELODI Presto [22,23,24].

Progressive developments in the field have led to the addition of new features and components in mGWAS-Explorer since version 1.0. High-throughput technologies have enabled systemic discovery of various molecular quantitative trait loci (QTLs), and integrative analysis of QTLs from multi-omics data is crucial for understanding functional impacts of genetic variants on phenotypes at different molecular levels [25-28]. In response, features for mapping SNPs to

eGenes (eQTLs) and SNPs to proteins (pQTLs) have been added, along with ongoing efforts to enhance transparency, scalability, reproducibility, and user support through the development of a corresponding R package.

We now introduce mGWAS-Explorer version 2.0 to address the evolving bioinformatics needs and challenges of mGWAS research. Compared to version 1.0, mGWAS-Explorer 2.0 features three key improvements:

- Implementation of a two-sample Mendelian randomization strategy for exploring causal relationships between metabolites and disease phenotypes;
- Addition of eQTL and pQTL data, as well as semantic triples evidence, for improved annotation and uncovering of mechanistic insights;
- Enhanced support for reproducible research through the release of the mGWASR package.

Results

Metabolome-wide Mendelian randomization

Utilizing genetic associations derived from large-scale metabolome-wide genome-wide association studies (mGWAS), we can systematically investigate the potential causal relationships between numerous metabolite compounds (targeted) or metabolite features (untargeted) and human diseases via Mendelian Randomization (MR). In our newly implemented “MR Analysis” module, we support two-sample MR analysis between metabolites and disease phenotypes.

To conduct a two-sample MR analysis, users must first select metabolites as exposures and designate disease outcomes on the data upload page. The mGWAS-Explorer 2.0 enables users to identify significant SNPs associated with metabolites (exposure), and subsequently extract these instrumental SNPs from the disease (outcome) GWAS. The information on SNP-metabolite

associations is derived from curated significant mGWAS summary statistics. Meanwhile, the data on disease outcomes is obtained through querying the API of the IEU OpenGWAS database [17], which houses complete GWAS summary statistics. After acquiring summary statistics for both exposure and outcome, users can harmonize datasets to ensure consistency in genetic instruments, effect sizes, and effect alleles.

Subsequently, the parameter page enables users to perform clumping and linkage disequilibrium pruning, retaining only independent genetic variants for MR estimation. Our platform offers eighteen distinct MR analysis methods, such as MR Egger, weighted median, and inverse variance-weighted methods, to estimate causal effects.

Additionally, mGWAS-Explorer 2.0 automatically performs sensitivity assessments and heterogeneity tests to evaluate potential violations of MR assumptions and the robustness of causal estimates. Upon completion, the MR results are displayed in a summary table view and illustrated in four types of plots for results interpretation. Users have the option to customize these plots in terms of format, resolution, or size for downloading purposes, ensuring a comprehensive and user-friendly experience that covers all key steps of two-sample MR analysis.

Triangulating evidence from semantic triples

Incorporating evidence from diverse research methods can mitigate biases and produce more reliable findings in response to research questions. This approach, referred to as "triangulation," has gained attention in epidemiological research, particularly for causal inference [22]. Triangulation increases confidence in results when multiple sources converge on the same conclusion.

In our study, we utilized triangulation to support causal estimates derived from Mendelian Randomization (MR) by examining millions of semantic triples extracted from scientific literature. The MR module's results page enables users to examine these semantic triples, which consist of subject-predicate-object relationships pertinent to the exposure (metabolite) and outcome (disease). Users can search for commonly enriched terms that might suggest an association. An overlap is identified when the object of the exposure query corresponds with the subject of the outcome query [24]. This rapid process typically takes a few seconds, with the findings presented in a data table or as a network diagram for users to investigate. Using this triangulation method offers a more in-depth understanding of the associations between exposures and outcomes.

Enhancing SNP-gene-metabolite network

The SNP-gene-metabolite network help gain potential functional insights of the SNP-metabolite statistical associations. In SNP-metabolite network, SNP nodes connect to metabolite nodes with which they have shown significant statistical associations. We have curated significant associations based on study-specific genome-wide cut-off threshold in version 1.0 of mGWAS-Explorer. However, it is limited to only include first-step neighbors of the input SNPs (e.g., SNP-metabolite statistical associations). We have added support for network expansion and network enrichment in the SNP module. mGWAS-explorer 2.0 offers SNP-gene-metabolite networks by introducing a new node type (genes) and new edge types (SNP-gene, gene-metabolites) based on SNP annotation and known gene/protein - metabolite relationships (i.e., enzymatic reactions or transporters). Afterwards, enriched networks are generated with new edges between SNPs, genes and metabolites, representing new relationship discoveries.

Enabling joint SNP/metabolite user input

Integrative SNP-metabolite analysis is now a valuable approach for identifying connections that cross the boundaries of traditional metabolic pathways [29]. We have added a new module to allow users to jointly provide SNPs and metabolite as input. For the SNP input, we currently support four types of mapping, including the SNP-metabolite statistical associations, SNP to gene mapping based on nearest gene or eQTL, SNP to protein based on pQTL, as well as SNP to disease mapping. For the metabolite input, metabolite-SNP mapping, knowledge-based metabolite to gene mapping, as well as metabolite-disease mappings are supported. Users can enter either a SNP, a metabolite or both. For metabolites, mGWAS-Explorer currently accepts compound names, HMDB IDs or KEGG IDs. For SNPs, rsID is currently supported. The uploaded SNP and metabolite are then mapped to the internal databases of mGWAS-Explorer. Following this step, users can filter the networks based on topological measures (degree and betweenness), shortest paths, to compute minimum network or to manually filter the network based on a given list. For instance, by using a degree cutoff as 1.0, we can exclude all terminal nodes to better visualize the nodes and edges with higher importance (i.e., higher degree levels).

Improving transparency/reproducibility

Transparent analytical procedures are essential for reproducible research since they increase the validity of research findings and scientific claims. Open-source software and in-depth documentation are crucial steps toward more reproducible analysis in the field of bioinformatics [30]. In mGWAS-Explorer 2.0, we have compiled and re-leased the underlying R functions as the mGWASR package (https://github.com/xia-lab/mGWASR). We have also added a Result Download page in each module to allow users to obtain all results tables and images generated

during the analysis, as well as the R command history. As a complement to the web-based platform, we anticipate that the R package and the R command history will enable users to track each stage of their analysis in an easily sharable and reproducible format (i.e., R script). Additionally, we have migrated all our frequently asked questions (FAQs) to the OmicsForum (https://omicsforum.ca/).

Case Studies

Crohn's Disease Case Study: To showcase the utility of the MR feature in mGWAS-Explorer 2.0, we leveraged a recent study integrating metabolomics, genomics and immune phenotypes to discover causal effects of metabolites in disease [31]. Specifically, we will demonstrate the application to Crohn's disease, which is a complex disorder with poorly understanding of the disease pathogenesis. Previous studies have shown that arachidonic acid shares common genetic variants with Crohn's disease through colocalization analysis [31]. Therefore, we sought to investigate the causal effect of the arachidonic acid on Crohn's disease using the summary statistics for both traits in mGWAS-Explorer 2.0 [4,32]. Using 24 independent genetic instruments (i.e., SNPs), the results of four commonly used MR methods (inverse-variance weighted, MR Egger, weighted median estimator and weighted mode estimator) consistently illustrate that the decrease in arachidonic acid levels had a causal effect on Crohn's disease (Figure 1), which is consistent to the findings reported by Chu et al [31]. This use case highlights how users can quickly and easily perform MR analysis by leveraging our comprehensive knowledgebase of mGWAS summary statistics as well as an easy-to-use interface to test the hypothesis of plausible causal role of metabolites on diseases.

(a)

Methods	MR Results				Heterogeneity Tests			Horizontal Pleiotropy		
	Number of SNPs	Beta	SE	P value	Q	Q_df	Q_pval	Egger Intercept	SE	P value
Inverse variance weighted	24	-0.91	0.313	0.0036	42.1	23	0.00893	-	-	-
MR Egger	24	-1.39	0.523	0.0146	39.7	22	0.0116	0.00993	0.00877	0.27
Simple mode	24	-0.128	0.711	0.858	-	-	-	-	-	-
Weighted median	24	-1.5	0.323	3.46e-06	-	-	-	-	-	-
Weighted mode	24	-1.48	0.315	9.73e-05	-	-	-	-	-	-

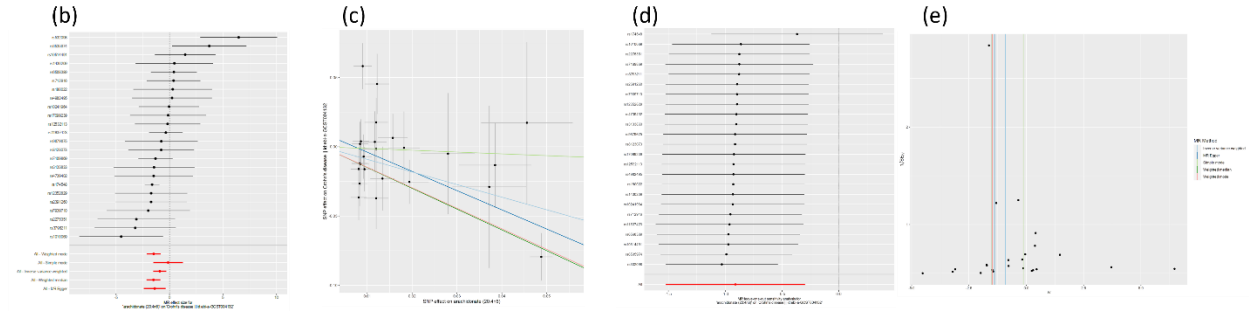

Figure 1 Mendelian randomization case study of the effect of arachidonic acid levels on Crohn's disease. (a) A summary table displaying the results of MR analysis, heterogeneity and horizontal pleiotropy tests; (b) a forest plot, comparing the causal effects calculated using the methods include all the SNPs to using each SNP separately; (c) a scatter plot, showing the relationships between SNP effects on arachidonic acids against the SNP effects on the Crohn's disease, with slope indicating the causal association; (d) a leave-one-out plot, which determine whether a single SNP is having a disproportionately larger impact on an association; (e) a funnel plot, showing the relationships between the causal effect of arachidonic acids on Crohn's disease calculated using each individual SNP as a separate instrument against the inverse of the standard error of the causal estimate.

Coronary Heart Disease Case Study: To demonstrate triangulating casual inference from MR with literature evidence, we used glycine and coronary heart disease as an example to explore the semantic evidence connecting the metabolite and disease. The causal associations between SNP effects on glycine against the SNP effects on the coronary heart disease are shown in Figure

2a. Genetic predisposition to higher glycine levels are associated with lower risk of coronary heart disease. Besides, Figure 2b displays the semantic-triples connections between glycine and coronary heart disease after searching for enriched overlapping terms. A total of 73 overlapping terms were identified, including homocysteine [33-35], ethanol [36,37], and TNF protein [38,39]. In the case of homocysteine, "homocysteine – PREDISPOSES – Coronary Arteriosclerosis" is the most enriched semantic triples on the outcome side (p-value: 4.38×10^{-120}), whereas "Glycine – INTERACTS_WITH – homocysteine" has a p-value of 8.3×10^{-6}. Therefore, we could hypothesize that the protective effect of glycine on coronary heart disease may be due to the interactions with homocysteine, providing this as a potential mechanism of interest.

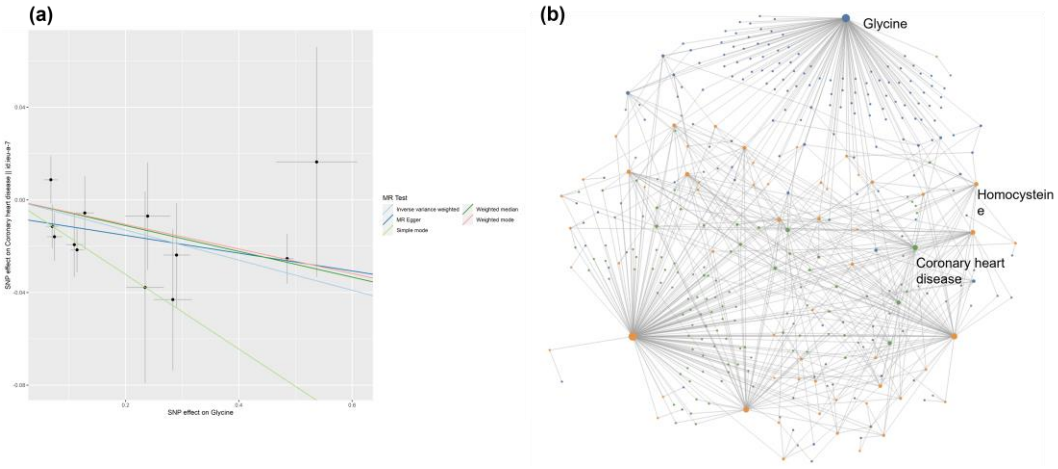

Figure 2 Triangulation of MR and literature evidence on the effects of glycine on coronary heart disease case study. (a) A scatter plot, showing the relationships between SNP effects on glycine against the SNP effects on coronary heart disease, with slope indicating the causal association; (b) a network of semantic triples (subject-predicate-object) from literature evidence between "glycine" and "coronary heart disease". Each node represents either an exposure subject (blue), an outcome object (green) or an overlapping enriched element (orange), where the object of a triple from the exposure query overlaps with a subject of a triple from the outcome query. Each edge is a "predicate" connecting two semantic elements.

Comparison with Other Tools

Table 1 compares mGWAS-Explorer 2.0 with its pervious version and several other web-based tools, including EpiGraphDB [21], The Molecular Human [40], and MR-Base [12]. EpiGraphDB is a graph database and an analytical platform containing comprehensive epidemiological and biomedical relationships, including pre-computed MR causal estimates, drugs, pathways, literatures evidence, ontology information, etc. The Molecular Human focuses on providing a comprehensive characterization of the molecular interactions using the integrated multi-omics data from 18 different platforms. MR-Base is an integrated platform that automates the two-sample MR analysis with a web interface, API, and R package, which incorporates a database of complete GWAS summary statistics. In comparison, mGWAS-Explorer 2.0 supports both linking multi-omics with diseases and performing MR analysis to identify metabolites with causal impacts on the diseases in the context of mGWAS.

Table 1. Comparison of the main features of mGWAS-Explorer (version 1.0-2.0) with other web-based tools. Symbols used for feature evaluations with ' √ ' for present, '−' for absent, and '+' for a more quantitative assessment (more '+' symbols indicate better support).

Tool Name	mGWAS-Explorer 2.0	mGWAS-Explorer 1.0	EpiGraphDB	The Molecular Human	MR-Base
Data input and processing					
Metabolite	√	√	√	√	√
SNP	√	√	√	√	–
Gene	√	√	√	√	–
MR exposure	√	–	√	–	√
MR outcome	√	–	√	–	√
Output					
Data table	√	√	√	√	√
Interactive network	+++	+++	++	++	–
Forest plot	√	–	–	–	√
Scatter plot	√	–	–	–	√
Funnel plot	√	–	–	–	√
Functions					
Mendelian randomization	√	–	*√	–	√
Exposure (metabolite)	**4238 metabolic traits, 65 studies	–	123 metabolic traits, 1 study	–	123 metabolic traits, 1 study
Enrichment analysis	√	√	–	–	–
Semantic triples evidence	√	–	√	–	–

URL links:

- EpiGraphDB: https://www.epigraphdb.org/ (accessed on 23 January 2023).
- The Molecular Human: http://comics.metabolomix.com/ (accessed on 23 January 2023).
- MR-Base: http://www.mrbase.org/ (accessed 23 January 2023)

* EpiGraphDB contains pre-computed MR causal estimates.

** Metabolic trait number includes both compounds and metabolic features, as well as metabolite ratios based on mGWAS-Explorer 1.0 when the effect size and standard error are available in the summary statistics.

Discussion

Systematic causal inference between modifiable risk factors and complex traits is a central challenge in the field of human genetics [41-43]. We have developed a platform that integrates

published GWAS summary statistics with analytical methods and visualization, with a particular focus on understanding the relationships between genetic variants, metabolites and diseases. The results from statistical Mendelian randomization analysis only support, but do not prove, causal relationships, however, it can guide interventional research when a randomized controlled trial may not be feasible [44]. Our first case study investigated the causal role of arachidonic acid (AA) on Crohn's diseases (CD) using MR. Arachidonic acid belongs to omega-6 polyunsaturated fatty acids and free AA enhances and modulate type 2 immune response, which are crucial for resistance to allergens and parasites [45,46]. In our analysis, negative causal effect of AA on CD is consistent with previous studies where CD patients has lower levels of AA [47]. However, much more studies need to be done to completely understand the fundamental mechanisms. The second case study highlights the protective causal role of glycine on coronary heart disease (CHD), which agrees with the findings from the MR study by Wittemans et al. [48]. In the semantic triples analysis, "Glycine – INTERACTS_WITH – homocysteine" and "homocysteine – PREDISPOSES – Coronary Arteriosclerosis" presents an example that how possible mechanisms could be drawn by mining literature data after MR analysis. A high homocysteine level is strongly associated with the prevalence of CHD (p-value: 4.38×10^{-120}). The role of homocysteine on CHD is explained by its negative effects vascular endothelium and smooth muscle cells [34]. On the other hand, it was reported that intracellular concentrations of homocysteine was lowered after 24 hours of co-incubation with glycine [35], although the mechanism of how glycine lowers the homocysteine concentrations is not clear.

Materials and Methods

Knowledgebase update and creation

(a) The data source for the mGWAS summary statistics can be found in the publication of version 1.0 of mGWAS-Explorer [6]. For complete GWAS summary data of the disease outcome, IEU OpenGWAS database was used by querying the Application Programming Interface (API) service of the database [17]. IEU OpenGWAS database contains manually curated collection of complete GWAS summary datasets. (b) eQTL from 49 tissues and pQTL data from blood are obtained from Genotype-Tissue Expression (GTEx) project and QTLbase database [49,50].

Methods for MR analysis

The statistical methods for pre-processing and MR analysis are based on the TwoSampleMR and MRInstruments R packages [12]. The pre-processing procedure facilitates the acquisition of independent instrumental variables by performing linkage disequilibrium (LD) clumping. In cases where the queried single nucleotide polymorphism (SNP) is absent in the outcome genome-wide association study (GWAS), we identify a proxy SNP in LD with the input SNP, utilizing the 1000 Genomes Project data as a reference. A crucial aspect of the analysis is harmonizing exposure and outcome data to guarantee that the effects of the SNP on exposure and outcome are associated with the same allele. Consequently, we provide three harmonization options for researchers to choose from, facilitating accurate analysis: i) assume all alleles are on the forward strand; ii) infer the forward strand alleles based on allele frequency; iii) adjust the strand for non-palindromic SNPs while excluding all palindromic SNPs.

Our approach incorporates 18 distinct MR methods, enabling users to compare results across different analytical techniques. In addition, we offer support for heterogeneity and horizontal pleiotropy testing. For the heterogeneity test, we implement Cochran's Q test, while the horizontal pleiotropy test is conducted using Egger regression. These tests contribute to a

comprehensive understanding of the potential biases within the MR analysis and promote robust and reliable results.

Semantic triples

The semantic triples are queried by using the API of the MELODI Presto [24], which uses a carefully-curated literature dataset from SemMedDB (96) and a high-performance architecture (i.e., Elasticsearch). MELODI Presto facilitates a rapid and efficient approach for systematically profiling semantic triples originating from the literature. This method enables the exploration of enriched literature data corresponding to specific search terms and the identification of potential intermediate disease mechanisms among term lists. The Semantic MEDLINE Database (SemMedDB) serves as a repository for semantic predications, including subject-predicate-object triples.

R package

mGWASR is written in the R programming language. The package is hosted on GitHub, which is build upon the R functions from the web server. To guarantee that the identical results will be generated, the R package and the web server have been thoroughly tested. mGWASR complies with the R package quality requirements and it includes detailed vignettes for step-by-step analysis.

Conclusions

In conclusion, our development of mGWAS-Explorer 2.0 represents a significant advancement in enabling researchers to investigate potential causal associations between

metabolites and disease phenotypes. With the anticipated increase in accessibility of published GWAS summary data, we foresee a tremendous expansion in the scope and impact of its applications, paving the way for novel discoveries and improved understanding of disease etiology.

References

1. Johnson, C.H.; Ivanisevic, J.; Siuzdak, G. Metabolomics: beyond biomarkers and towards mechanisms. Nature reviews Molecular cell biology 2016, 17, 451.

2. Lotta, L.A.; Pietzner, M.; Stewart, I.D.; Wittemans, L.B.L.; Li, C.; Bonelli, R.; Raffler, J.; Biggs, E.K.; Oliver-Williams, C.; Auyeung, V.P.W.; et al. A cross-platform approach identifies genetic regulators of human metabolism and health. Nature Genetics 2021, 53, 54-64, doi:10.1038/s41588-020-00751-5.

3. Surendran, P.; Stewart, I.D.; Au Yeung, V.P.W.; Pietzner, M.; Raffler, J.; Wörheide, M.A.; Li, C.; Smith, R.F.; Wittemans, L.B.L.; Bomba, L.; et al. Rare and common genetic determinants of metabolic individuality and their effects on human health. Nat Med 2022, 28, 2321-2332, doi:10.1038/s41591-022-02046-0.

4. Shin, S.-Y.; Fauman, E.B.; Petersen, A.-K.; Krumsiek, J.; Santos, R.; Huang, J.; Arnold, M.; Erte, I.; Forgetta, V.; Yang, T.-P. An atlas of genetic influences on human blood metabolites. Nature genetics 2014, 46, 543.

5. Gieger, C.; Geistlinger, L.; Altmaier, E.; De Angelis, M.H.; Kronenberg, F.; Meitinger, T.; Mewes, H.-W.; Wichmann, H.-E.; Weinberger, K.M.; Adamski, J. Genetics meets metabolomics: a genome-wide association study of metabolite profiles in human serum. PLoS genetics 2008, 4, e1000282.

6. Chang, L.; Zhou, G.; Ou, H.; Xia, J. mGWAS-Explorer: Linking SNPs, Genes, Metabolites, and Diseases for Functional Insights. Metabolites 2022, 12, doi:10.3390/metabo12060526.

7. Lord, J.; Jermy, B.; Green, R.; Wong, A.; Xu, J.; Legido-Quigley, C.; Dobson, R.; Richards, M.; Proitsi, P. Mendelian randomization identifies blood metabolites previously linked to midlife cognition as causal candidates in Alzheimer's disease. Proceedings of the National Academy of Sciences 2021, 118, e2009808118, doi:doi:10.1073/pnas.2009808118.

8. Qin, Y.; Méric, G.; Long, T.; Watrous, J.D.; Burgess, S.; Havulinna, A.S.; Ritchie, S.C.; Brożyńska, M.; Jousilahti, P.; Perola, M.; et al. Genome-wide association and Mendelian randomization analysis prioritizes bioactive metabolites with putative causal effects on common diseases. medRxiv 2020, 2020.2008.2001.20166413, doi:10.1101/2020.08.01.20166413.

9. Tanha, H.M.; Sathyanarayanan, A.; Nyholt, D.R. Genetic overlap and causality between blood metabolites and migraine. The American Journal of Human Genetics 2021, 108, 2086-2098, doi:https://doi.org/10.1016/j.ajhg.2021.09.011.

10. Smith, G.D.; Ebrahim, S. 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? Int J Epidemiol 2003, 32, 1-22, doi:10.1093/ije/dyg070.

11. de Leeuw, C.; Savage, J.; Bucur, I.G.; Heskes, T.; Posthuma, D. Understanding the assumptions underlying Mendelian randomization. Eur J Hum Genet 2022, 30, 653-660, doi:10.1038/s41431-022-01038-5.

12. Hemani, G.; Zheng, J.; Elsworth, B.; Wade, K.H.; Haberland, V.; Baird, D.; Laurin, C.; Burgess, S.; Bowden, J.; Langdon, R.; et al. The MR-Base platform supports systematic causal inference across the human phenome. eLife 2018, 7, doi:10.7554/eLife.34408.

13. Bowden, J.; Davey Smith, G.; Haycock, P.C.; Burgess, S. Consistent Estimation in Mendelian Randomization with Some Invalid Instruments Using a Weighted Median Estimator. Genet Epidemiol 2016, 40, 304-314, doi:10.1002/gepi.21965.

14. Bowden, J.; Davey Smith, G.; Burgess, S. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. International Journal of Epidemiology 2015, 44, 512-525, doi:10.1093/ije/dyv080.

15. Verbanck, M.; Chen, C.-Y.; Neale, B.; Do, R. Detection of widespread horizontal pleiotropy in causal relationships inferred from Mendelian randomization between complex traits and diseases. Nature Genetics 2018, 50, 693-698, doi:10.1038/s41588-018-0099-7.

16. Burgess, S.; Scott, R.A.; Timpson, N.J.; Davey Smith, G.; Thompson, S.G.; Consortium, E.-I. Using published data in Mendelian randomization: a blueprint for efficient identification of causal risk factors. European Journal of Epidemiology 2015, 30, 543-552, doi:10.1007/s10654-015-0011-z.

17. Elsworth, B.; Lyon, M.; Alexander, T.; Liu, Y.; Matthews, P.; Hallett, J.; Bates, P.; Palmer, T.; Haberland, V.; Smith, G.D.; et al. The MRC IEU OpenGWAS data infrastructure. bioRxiv 2020, 2020.2008.2010.244293, doi:10.1101/2020.08.10.244293.

18. Porcu, E.; Sjaarda, J.; Lepik, K.; Carmeli, C.; Darrous, L.; Sulc, J.; Mounier, N.; Kutalik, Z. Causal Inference Methods to Integrate Omics and Complex Traits. Cold Spring Harb Perspect Med 2021, 11, doi:10.1101/cshperspect.a040493.

19. Ference, B.A.; Yoo, W.; Alesh, I.; Mahajan, N.; Mirowska, K.K.; Mewada, A.; Kahn, J.; Afonso, L.; Williams, K.A., Sr.; Flack, J.M. Effect of long-term exposure to lower low-density lipoprotein cholesterol beginning early in life on the risk of coronary heart disease: a Mendelian randomization analysis. J Am Coll Cardiol 2012, 60, 2631-2639, doi:10.1016/j.jacc.2012.09.017.

20. Holmes, M.V.; Simon, T.; Exeter, H.J.; Folkersen, L.; Asselbergs, F.W.; Guardiola, M.; Cooper, J.A.; Palmen, J.; Hubacek, J.A.; Carruthers, K.F.; et al. Secretory phospholipase A(2)-IIA and cardiovascular disease: a mendelian randomization study. J Am Coll Cardiol 2013, 62, 1966-1976, doi:10.1016/j.jacc.2013.06.044.

21. Liu, Y.; Elsworth, B.; Erola, P.; Haberland, V.; Hemani, G.; Lyon, M.; Zheng, J.; Lloyd, O.; Vabistsevits, M.; Gaunt, T.R. EpiGraphDB: a database and data mining platform for health data science. Bioinformatics 2021, 37, 1304-1311, doi:10.1093/bioinformatics/btaa961.

22. Lawlor, D.A.; Tilling, K.; Davey Smith, G. Triangulation in aetiological epidemiology. International Journal of Epidemiology 2017, 45, 1866-1886, doi:10.1093/ije/dyw314.

23. Kilicoglu, H.; Shin, D.; Fiszman, M.; Rosemblat, G.; Rindflesch, T.C. SemMedDB: a PubMed-scale repository of biomedical semantic predications. Bioinformatics 2012, 28, 3158-3160, doi:10.1093/bioinformatics/bts591.

24. Elsworth, B.; Gaunt, T.R. MELODI Presto: a fast and agile tool to explore semantic triples derived from biomedical literature. Bioinformatics 2021, 37, 583-585, doi:10.1093/bioinformatics/btaa726.

25. Yin, X.; Bose, D.; Kwon, A.; Hanks, S.C.; Jackson, A.U.; Stringham, H.M.; Welch, R.; Oravilahti, A.; Fernandes Silva, L.; Locke, A.E.; et al. Integrating transcriptomics, metabolomics, and GWAS helps reveal molecular mechanisms for metabolite levels and disease risk. The American Journal of Human Genetics 2022, 109, 1727-1741, doi:https://doi.org/10.1016/j.ajhg.2022.08.007.

26. Pietzner, M.; Wheeler, E.; Carrasco-Zanini, J.; Cortes, A.; Koprulu, M.; Wörheide, M.A.; Oerton, E.; Cook, J.; Stewart, I.D.; Kerrison, N.D.; et al. Mapping the proteo-genomic convergence of human diseases. Science 2021, 374, eabj1541, doi:10.1126/science.abj1541.

27. Võsa, U.; Claringbould, A.; Westra, H.-J.; Bonder, M.J.; Deelen, P.; Zeng, B.; Kirsten, H.; Saha, A.; Kreuzhuber, R.; Yazar, S.; et al. Large-scale cis- and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. Nature Genetics 2021, 53, 1300-1310, doi:10.1038/s41588-021-00913-z.

28. Ye, Y.; Zhang, Z.; Liu, Y.; Diao, L.; Han, L. A Multi-Omics Perspective of Quantitative Trait Loci in Precision Medicine. Trends Genet 2020, 36, 318-336, doi:10.1016/j.tig.2020.01.009.

29. Jung, T.; Jung, Y.; Moon, M.K.; Kwon, O.; Hwang, G.S.; Park, T. Integrative Pathway Analysis of SNP and Metabolite Data Using a Hierarchical Structural Component Model. Front Genet 2022, 13, 814412, doi:10.3389/fgene.2022.814412.

30. Chang, H.Y.; Colby, S.M.; Du, X.; Gomez, J.D.; Helf, M.J.; Kechris, K.; Kirkpatrick, C.R.; Li, S.; Patti, G.J.; Renslow, R.S.; et al. A Practical Guide to Metabolomics Software Development. Anal Chem 2021, 93, 1912-1923, doi:10.1021/acs.analchem.0c03581.

31. Chu, X.; Jaeger, M.; Beumer, J.; Bakker, O.B.; Aguirre-Gamboa, R.; Oosting, M.; Smeekens, S.P.; Moorlag, S.; Mourits, V.P.; Koeken, V.A.C.M.; et al. Integration of metabolomics, genomics, and immune phenotypes reveals the causal roles of metabolites in disease. Genome Biology 2021, 22, 198, doi:10.1186/s13059-021-02413-z.

32. de Lange, K.M.; Moutsianas, L.; Lee, J.C.; Lamb, C.A.; Luo, Y.; Kennedy, N.A.; Jostins, L.; Rice, D.L.; Gutierrez-Achury, J.; Ji, S.G.; et al. Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease. Nat Genet 2017, 49, 256-261, doi:10.1038/ng.3760.

33. McCarty, M.F.; O'Keefe, J.H.; DiNicolantonio, J.J. Interleukin-1beta may act on hepatocytes to boost plasma homocysteine - The increased cardiovascular risk associated with

elevated homocysteine may be mediated by this cytokine. Med Hypotheses 2017, 102, 78-81, doi:10.1016/j.mehy.2017.03.022.

34. Feng, L.; Nian, S.; Zhang, S.; Xu, W.; Zhang, X.; Ye, D.; Zheng, L. The associations between serum biomarkers and stenosis of the coronary arteries. Oncotarget 2016, 7, 39231-39240, doi:10.18632/oncotarget.9645.

35. Sim, W.C.; Han, I.; Lee, W.; Choi, Y.J.; Lee, K.Y.; Kim, D.G.; Jung, S.H.; Oh, S.H.; Lee, B.H. Inhibition of homocysteine-induced endoplasmic reticulum stress and endothelial cell damage by l-serine and glycine. Toxicol In Vitro 2016, 34, 138-145, doi:10.1016/j.tiv.2016.04.004.

36. Movva, R.; Figueredo, V.M. Alcohol and the heart: to abstain or not to abstain? Int J Cardiol 2013, 164, 267-276, doi:10.1016/j.ijcard.2012.01.030.

37. Gallegos, S.; Muñoz, B.; Araya, A.; Aguayo, L.G. High ethanol sensitive glycine receptors regulate firing in D1 medium spiny neurons in the nucleus accumbens. Neuropharmacology 2019, 160, 107773, doi:10.1016/j.neuropharm.2019.107773.

38. Grira, N.; Lahidheb, D.; Lamine, O.; Ayoub, M.; Wassaifi, S.; Aouni, Z.; Fehri, W.; Mazigh, C. The Association of IL-6, TNFα and CRP Gene Polymorphisms with Coronary Artery Disease in a Tunisian Population: A Case-Control study. Biochem Genet 2021, 59, 751-766, doi:10.1007/s10528-021-10035-0.

39. Liu, Y.; Wang, X.; Wu, H.; Chen, S.; Zhu, H.; Zhang, J.; Hou, Y.; Hu, C.A.; Zhang, G. Glycine enhances muscle protein mass associated with maintaining Akt-mTOR-FOXO1 signaling and suppressing TLR4 and NOD2 signaling in piglets challenged with LPS. Am J Physiol Regul Integr Comp Physiol 2016, 311, R365-373, doi:10.1152/ajpregu.00043.2016.

40. Halama, A.; Zaghlool, S.; Thareja, G.; Kader, S.; Muftha, W.A.; Mook-Kanamori, M.; Sarwath, H.; Mohamoud, Y.A.; Ameling, S.; Baković, M.P.; et al. The Molecular Human – A

Roadmap of Molecular Interactions Linking Multiomics Networks with Disease Endpoints. medRxiv 2022, 2022.2010.2031.22281758, doi:10.1101/2022.10.31.22281758.

41. Pingault, J.-B.; O'Reilly, P.F.; Schoeler, T.; Ploubidis, G.B.; Rijsdijk, F.; Dudbridge, F. Using genetic data to strengthen causal inference in observational research. Nature Reviews Genetics 2018, 19, 566-580, doi:10.1038/s41576-018-0020-3.

42. Pingault, J.B.; Richmond, R.; Davey Smith, G. Causal Inference with Genetic Data: Past, Present, and Future. Cold Spring Harb Perspect Med 2022, 12, doi:10.1101/cshperspect.a041271.

43. Zheng, J.; Haberland, V.; Baird, D.; Walker, V.; Haycock, P.C.; Hurle, M.R.; Gutteridge, A.; Erola, P.; Liu, Y.; Luo, S.; et al. Phenome-wide Mendelian randomization mapping the influence of the plasma proteome on complex diseases. Nature Genetics 2020, 52, 1122-1131, doi:10.1038/s41588-020-0682-6.

44. Davies, N.M.; Holmes, M.V.; Davey Smith, G. Reading Mendelian randomisation studies: a guide, glossary, and checklist for clinicians. Bmj 2018, 362, k601, doi:10.1136/bmj.k601.

45. Tallima, H.; El Ridi, R. Arachidonic acid: Physiological roles and potential health benefits - A review. Journal of advanced research 2018, 11, 33-41, doi:10.1016/j.jare.2017.11.004.

46. Wang, B.; Wu, L.; Chen, J.; Dong, L.; Chen, C.; Wen, Z.; Hu, J.; Fleming, I.; Wang, D.W. Metabolism pathways of arachidonic acids: mechanisms and potential therapeutic targets. Signal Transduction and Targeted Therapy 2021, 6, 94, doi:10.1038/s41392-020-00443-w.

47. Trebble, T.M.; Arden, N.K.; Wootton, S.A.; Mullee, M.A.; Calder, P.C.; Burdge, G.C.; Fine, D.R.; Stroud, M.A. Peripheral blood mononuclear cell fatty acid composition and inflammatory mediator production in adult Crohn's disease. Clin Nutr 2004, 23, 647-655, doi:10.1016/j.clnu.2003.10.017.

48. Wittemans, L.B.L.; Lotta, L.A.; Oliver-Williams, C.; Stewart, I.D.; Surendran, P.; Karthikeyan, S.; Day, F.R.; Koulman, A.; Imamura, F.; Zeng, L.; et al. Assessing the causal association of glycine with risk of cardio-metabolic diseases. Nature Communications 2019, 10, 1060, doi:10.1038/s41467-019-08936-1.

49. The Genotype-Tissue Expression (GTEx) project. Nat Genet 2013, 45, 580-585, doi:10.1038/ng.2653.

50. Zheng, Z.; Huang, D.; Wang, J.; Zhao, K.; Zhou, Y.; Guo, Z.; Zhai, S.; Xu, H.; Cui, H.; Yao, H.; et al. QTLbase: an integrative resource for quantitative trait loci across multiple human molecular phenotypes. Nucleic Acids Research 2019, 48, D983-D991, doi:10.1093/nar/gkz888.

Chapter 5: General discussion

Major technological advancements have greatly accelerated large-scale collections of genetic and phenotypic data. Despite the success in identifying genetic risk loci with diseases and molecular phenotypes (i.e., omics), it has been challenging to translate genetic associations and regulatory interactions to functional insights for understanding mechanisms of diseases and development of therapeutic targets. The main goal of my thesis was to bridge the gap from genetic associations, regulatory interactions to biological interpretation in order to better comprehend the results. Specifically, I focused on metabolome genome-wide association studies (mGWAS) in **Chapter 2, 4** and miRNA regulatory networks in **Chapter 3**. To achieve my goal, I have developed several easy-to-use web-based and R package software tools, including mGWAS-Explorer (www.mgwas.ca) (99), miRNet (www.mirnet.ca) version 2.0 (100,101), miRNetR (https://github.com/xia-lab/miRNetR), mGWAS-Explorer version 2.0 and mGWASR (https://github.com/xia-lab/mGWASR).

Interpreting mGWAS results to generate mechanistic insights requires navigating a range of databases related to genomics and metabolomics as well as advanced visualization tools, which can be a daunting task and may cause errors in combining different data types. To provide an integrated platform for functional interpretation of mGWAS, I developed a user-friendly web application, mGWAS-Explorer version 1.0 by leveraging a variety of high-quality databases and advanced visual analytics technologies (**Chapter 2**). mGWAS-Explorer allows users to (1) upload a list of SNPs, metabolites or genes; (2) browse or search significant statistical associations between SNP and metabolites from published mGWAS summary statistics. Numerous functional annotation databases are included in the knowledgebase, such as HaploReg (102), VEP (103), DisGeNET (104) for genetic annotation; HMDB (105), KEGG (106), Recon3D (107), TCDB (108)

for metabolite annotation; and STRING (109), InnateDB (110) for protein-protein interactions. Additionally, several well-known databases are used for enrichment analysis, including KEGG, GO (111), Reactome (112), Orphanet (113), DisGeNET (104), DrugMatrix (114), and DSigDB (115). mGWAS-Explorer prioritizes genes based on positional mapping, which maps SNPs to genes based on physical distances. Notably, it is suggested that the nearest gene mapping method is a useful indicator of true positive genes for mQTLs (116) and we have used this method in the gene module. Meanwhile, LD proxy search is also supported where users can search for variants that are in linkage disequilibrium with other variants. In addition to SNP annotation, mGWAS-Explorer also annotates mapped genes with functional information to explore the group behavior of gene function across a collection of gene, such as enrichment of genes in molecular pathways. Besides SNP annotation, mGWAS-Explorer supports metabolite to gene mapping based on enzyme or transporter information, and metabolite to disease mapping for metabolite annotation. The popular metabolite-set enrichment analysis is also provided to facilitate identifying biologically meaningful patterns of metabolite function across a set of metabolites (117). Apart from deep annotation, mGWAS-Explorer includes a powerful network visual analytics system enabling interactive exploration and topological analysis. The rationale behind network-based approach is that network is an intuitive model to dissect mQTLs, which is characterized by polygenicity and pleiotropy (45). In a network, nodes represent genetic or phenotypic entities and edges represent associations between them. Such a network can help in the analysis of genetic associations across metabolic and disease phenotypes and the identification of potential shared genetic influences. A variety of features are available to facilitate the visual exploration and customization of the network. For instance, degree and between centrality can be used to identify key nodes in the network. The utility of mGWAS-Explorer was illustrated in our COVID-19 and

type 2 diabetes case studies (99). By integrating SNP-metabolite, SNP-disease, SNP-gene mapping and protein-protein interactions, I identified shared genetic signals between metabolites (i.e., citric acid and fibrinogen A-α peptides) and COVID-19 at *ABO* locus, as well as *ENPEP* and *FUT2* as potential candidate gene. This has illustrated the power of leveraging prior knowledge with statistical associations in revealing biologically meaningful signals. In the second case study, mGWAS-Explorer not only confirmed the genetic associations between T2D and citrulline metabolites, but also identified shared signals between mQTL, T2D and COVID-19. The results suggested potential comorbidity between T2D and COVID-19, which has been reported in multiple other studies (118-120).

In **Chapter 3**, I aimed to improve the understanding of the miRNA regulatory networks by developing version 2.0 of miRNet, which is a web-based visual analytics platform for miRNA-centric gene regulatory analysis. The rationale for focusing on miRNA related regulatory interactions is that miRNAs are highly powerful genetic regulators, as shown by the fact that a single miRNA can control a wide range of target genes to drive entire cellular pathways (121). Due to this characteristic, miRNAs are now considered to be very intriguing therapeutic tools to recover cell functions that have been disrupted as a result of a disease phenotype (121,122). To better understand miRNA regulatory interactions with target genes and other important players (e.g., TFs and lncRNAs), I significantly improved the knowledgebase for miRNA-target interactions by adding new modules for TFs and miRSNPs, as well as extending the miRNA-lncRNA interactions to incorporate all other important ncRNAs, such as circRNA, pseudogene, and sncRNA based on starBase (123). Besides interactions mapping, miRNet 2.0 provides a comprehensive support for functional enrichment analysis, including nine annotation libraries (four target gene set libraries and six miRNA set libraries), and two enrichment algorithms

(hypergeometric tests and empirical sampling). In addition to the functional enrichment analysis, miRNet 2.0 supports flexible user input, accepting a list of miRNAs, miR-SNPs, genes, transcription factors, small molecules, ncRNAs, diseases, epigenetic modifiers, any of their combinations or a data table from microarray, RNAseq or RT-qPCR experiments. The workflow is as follows. Users can start the analysis by uploading various data types or by selecting queries from built-in databases. The input will be mapped to the knowledgebase to generate tables and networks. Subsequently, users can easily explore the network in the visualization page and conduct topological or functional analysis to identify key nodes or modules. miRNet 2.0 also supports filtering dense networks based on degree or betweenness centrality, shortest path, to reduce the network based on a list, or to calculate minimum network. To improve transparency and reproducibility, I have developed the underlying R package (https://github.com/xia-lab/miRNetR). As illustrated in our multiple sclerosis case study, miRNet 2.0 can easily pinpoint key regulators by visually inspecting the networks, where hub nodes (e.g., miRNAs) with high degree value demonstrate that they have high numbers of target genes. Notably, our tool has been used by multiple institutions to investigate disease mechanism related to miRNAs, for instance to study spaceflight-associated miRNA signature by NASA gene lab, and to understand the role of miR-2392 in driving SARS-CoV-2 infection (124,125).

In **Chapter 4**, I aimed to gain causal insights into the metabolites on diseases by developing mGWAS-Explorer version 2.0 with new modules and features. I implemented the two-sample Mendelian Randomization analysis framework, which leverages the summary statistics of mGWAS and GWAS with disease phenotypes to infer causal relationships between metabolites and diseases. mGWAS-Explorer 2.0 allow users to obtain SNPs (i.e., instrumental variables) that are significant for the metabolites (i.e., exposure) and extract the instrument SNPs from the disease

GWAS (i.e., outcome). The SNP-metabolite association data is based on the curated significant mGWAS summary statistics and the disease outcome data is obtained by querying the API of IEU OpenGWAS database (126), which contain complete summary statistics. Various functions are included to preprocess the data, such as performing LD pruning to get independent SNPs, searching for LD proxies, and harmonizing the effects between exposure and outcome. mGWAS-Explorer 2.0 supports commonly used Mendelian randomization methods and related sensitivity analysis. Mendelian randomization is a powerful method used to estimate causal relationships between genetic variants, exposures, and outcomes in observational data. It relies on three main assumptions for the results to be valid: i) Relevance assumption: the genetic variant(s) used as an instrumental variable (IV) must be robustly associated with the exposure of interest. ii) Independence assumption: the genetic variant(s) must be independent of any confounders that influence both the exposure and outcome. iii) Exclusion restriction assumption: the genetic variant(s) must only affect the outcome through the exposure and not through alternative pathways. Violations of these assumptions can lead to biased causal estimates. To minimize these violations, both algorithmic and database protections can be implemented. Algorithmic protections include several strategies: (i) multiple IVs: leveraging multiple independent genetic variants as IVs can help mitigate the effects of individual genetic variants violating the MR assumptions. Techniques such as weighted median or MR-Egger regression can account for potential violations and yield more reliable causal estimates. (ii) Sensitivity analyses: Conducting various sensitivity analyses can aid in identifying and evaluating the impact of potential MR assumption violations. For instance, MR-Egger regression can detect whether directional horizontal pleiotropy is driving the results of an MR analysis. Database protections involve the following measures: (i) large and well-characterized cohorts: utilizing data from sizable, well-characterized cohorts with high-quality

genotyping and phenotyping information can minimize measurement errors and enhance the precision of causal estimates. This can help reduce violations of the relevance and independence assumptions. (ii) Harmonization of data: ensuring consistency in exposure and outcome definitions across the studies included in the MR analysis is crucial. Data harmonization can help to diminish potential biases stemming from varying definitions or data collection methods. (iii) Ancestry and population stratification: accounting for population stratification, or the presence of systematic differences in allele frequencies between subpopulations, can help minimize biases introduced by the violation of the independence assumption. This can be achieved by either limiting the analysis to a homogeneous population or adjusting for population structure using techniques such as principal components analysis. In summary, using both algorithmic and database protections can minimize violations of the Mendelian randomization assumptions, ultimately leading to more reliable and valid causal estimates.

In addition, multiple visualization of results is available to facilitate the interpretation. Apart from two-sample MR, mGWAS-Explorer 2.0 also allow users to triangulate evidence from literature sources. The idea behind triangulation is that if the findings from different approaches all direct to the same conclusion, we are more confident in the results (93). This is particular important in causal inference in observation studies. As randomized controlled trial (RCT) is costly and may not be feasible, causal inference from observational studies could prioritize the targets for RCT or to assist in public health policy making process (94). A single source of evidence could have its own bias due to study design or measurement errors. Therefore, triangulated evidence could minimize the bias and strengthens the reliability of the results. However, triangulation can be influenced by biases that stem from the literature or previous experiments. For example: (i) publication bias: studies with significant findings are more likely to be published, leading to a bias

in the available literature. This could result in an overemphasis on certain diseases. (ii) Limited available data: if previous studies have focused on certain populations or ethnic groups, triangulation efforts may be biased due to the lack of data from other populations. These biases can be mitigated to some extent by using larger and more diverse samples, conducting unbiased hypothesis-free research. While triangulation is not without its limitations, it remains a valuable approach for identifying and validating causal estimates and understanding their biological mechanisms.

Limitations

The application of network-based multi-omics integration offers substantial potential for elucidating complex biological processes and enhancing our understanding of the genetics of metabolism and regulatory interactions. This approach includes the combination and analysis of data from multiple sources of biological information (e.g., genomics, transcriptomics, proteomics, and metabolomics) utilizing network-based models. Despite its immense promise, certain limitations persist in network-based multi-omics integration, which warrant further examination: (i) Data heterogeneity: multi-omics data arise from different sources and platforms, characterized by differing data types and scales. Consequently, integrating this data proves challenging, necessitating pre-processing and normalization procedures to ensure compatibility and comparability. (ii) Incomplete data: owing to the rapid evolution of high-throughput techniques generating multi-omics data, datasets can be incomplete or lack of comprehensive coverage of biological systems. This missing data may yield biased or restricted insights. (iii) Computational complexity: network-based methodologies frequently demand computationally demanding algorithms, which present substantial processing power and memory requirements, particularly

when handling large-scale multi-omics datasets. (iv) Noise and biases: multi-omics data accuracy and reproducibility can be adversely impacted by experimental and technical noise, as well as platform- and batch-specific biases. It can be difficult to apply the right statistical methods to account for these factors. Notwithstanding these limitations, network-based multi-omics integration has demonstrated potential in yielding valuable insights into complex biological systems and revolutionizing our understanding of human health and disease. Continuous improvements in computational algorithms, data processing, and experimental methods will assist overcome these obstacles and advance the field.

The present thesis acknowledges the indispensable role of databases in constructing networks. However, it is imperative to recognize certain limitations that may arise in specific contexts, which could potentially impact the conclusions drawn and the development of bioinformatics applications. These limitations can be categorized into three primary areas: data quality, data bias, and accessibility. Firstly, data quality remains a significant concern, as databases may contain outdated or inaccurate information. The presence of such incorrect data could ultimately result in biased conclusions, thereby undermining the validity of the research. Secondly, data bias poses a challenge, as databases may predominantly feature data from well-studied diseases, tissues, or populations, leading to an inadvertent bias in the results obtained. This skewed representation may inadvertently prioritize certain areas of study over others, further limiting the scope of potential insights. The third limitation pertains to accessibility, as some databases may either be difficult to access or necessitate permission for utilization. This restriction can constrain the resources available for the development of open-source bioinformatics tools, thereby impacting the breadth of knowledge that can be incorporated into these applications. To mitigate the potential adverse effects of these limitations, researchers are encouraged to adopt a rigorous approach when

sourcing, cleaning, and verifying data from databases. Using multiple data sources could further reduce the risk of data bias, as a more comprehensive and representative dataset can be compiled. By addressing these limitations and adopting rigorous strategies for data acquisition and validation, the bioinformatics and genetics community can continue to develop more robust, unbiased, and accessible tools and approaches, thus driving progress and innovation in the field.

Rationale for technical implementations

The choice of technical implementations was primarily driven by their suitability for specific tasks, widespread usage, and compatibility with other used technologies. (1) R programming language: the R language and environment, known for its statistical computing and graphical capabilities, is widely utilized for data analysis and statistical modeling. In the context of mGWAS-Explorer and miRNet 2.0, R serves as the computational backbone for data processing and analysis. (2) JavaServer Faces (JSF) technology: JSF, a Java-based web application framework, streamlines the development of user interfaces for web applications by providing a set of reusable UI components. JSF was chosen for its ability to facilitate the construction of the web interface. (3) PrimeFaces: As a UI component library designed for JavaServer Faces, PrimeFaces offers a comprehensive set of rich, ready-to-use components that can be seamlessly integrated into JSF applications. PrimeFaces was selected to develop the user interface elements in the web application. (4) SQLite: SQLite, a lightweight and serverless relational database management system (RDBMS), is widely employed for small to medium-sized applications due to its self-contained nature. In my project, SQLite serves as the storage solution for the integrated data. (5) jQuery: jQuery, a renowned JavaScript library, simplifies client-side HTML scripting by providing extensive functionality for DOM manipulation, event handling, and animation. jQuery was

adopted for general-purpose scripting in the web application. (6) Sigma.js: Sigma.js is a JavaScript library specifically designed for the display and manipulation of graph data on web pages, making it highly suitable for visualizing large networks. In the context of my project, Sigma.js is used for network display and interactions. (7) iGraph: iGraph, a collection of network analysis tools with implementations in multiple programming languages, including R, enables the creation, analysis, and visualization of networks. iGraph is utilized for network analysis and layout in the web application. (8) ECharts-GL: ECharts-GL, an extension pack of Apache ECharts, enhances the powerful charting and visualization capabilities of ECharts by introducing WebGL support for advanced visualization techniques such as 3D plots. In mGWAS-Explorer, ECharts-GL is used to generate the 3D Manhattan plots.

Chapter 6: Conclusions and future directions

Computational platforms developed in Chapter 2, 3 and 4 have significantly accelerated genetic research discovery by reducing the barriers in data pre-processing, analysis, and visualization in mGWAS and miRNA regulatory network research. Several future directions can be suggested to continue this work.

Mendelian randomization to integrate other omics and complex traits

While my focus was on metabolome-wide Mendelian randomization, integrating other omics data with disease outcomes could pinpoint molecular traits causally related to disease development at different levels, such as transcriptome, methylome and proteome (127). For instance, a recent study identified the causal effects of 65 proteins on 52 disease-related phenotypes by performing phenome-wide Mendelian randomization and colocalization analysis on proteins (36). The importance of this approach in finding and prioritizing prospective therapeutic targets was demonstrated by an analysis of data from historical drug development programs that revealed target-indication pairs with MR and colocalization support were more likely to be approved.

Other causal inference methods in genetics

Despite successfully applications of MR in causal inference, this method requires strong assumptions in horizontal pleiotropy, where the genetic instruments must influence the disease outcome only through the exposure viable (90). However, the assumptions may not be met in real-case scenario. Bayesian network (BN) is an emerging approach to infer the relationships in directed acyclic graph and is not restricted to the assumptions by MR. The causal inference is achieved by identifying conditional dependencies while testing multiple traits (128). A recent study identified potential causal links using the data from genotyped population-scale biobanks by using the

Bayesian network. However, more work needs to be done to test the performance and applicability to other molecular GWAS data.

Chapter 7: Master reference list

1. Kellis, M., Wold, B., Snyder, M.P., Bernstein, B.E., Kundaje, A., Marinov, G.K., Ward, L.D., Birney, E., Crawford, G.E. and Dekker, J. (2014) Defining functional DNA elements in the human genome. *Proceedings of the National Academy of Sciences*, **111**, 6131-6138.
2. Sinnott-Armstrong, N., Naqvi, S., Rivas, M. and Pritchard, J.K. (2021) GWAS of three molecular traits highlights core genes and pathways alongside a highly polygenic background. *eLife*, **10**, e58615.
3. Ye, Y., Zhang, Z., Liu, Y., Diao, L. and Han, L. (2020) A Multi-Omics Perspective of Quantitative Trait Loci in Precision Medicine. *Trends Genet*, **36**, 318-336.
4. Kastenmüller, G., Raffler, J., Gieger, C. and Suhre, K. (2015) Genetics of human metabolism: an update. *Human molecular genetics*, **24**, R93-R101.
5. Wang, Z., Gerstein, M. and Snyder, M. (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet*, **10**, 57-63.
6. Krueger, F., Kreck, B., Franke, A. and Andrews, S.R. (2012) DNA methylome analysis using short bisulfite sequencing data. *Nature methods*, **9**, 145-151.
7. Milne, T.A., Zhao, K. and Hess, J.L. (2009) Chromatin immunoprecipitation (ChIP) for analysis of histone modifications and chromatin-associated proteins. *Methods Mol Biol*, **538**, 409-423.
8. Jovel, J., Patterson, J., Wang, W., Hotte, N., O'Keefe, S., Mitchel, T., Perry, T., Kao, D., Mason, A.L., Madsen, K.L. *et al.* (2016) Characterization of the Gut Microbiome Using 16S or Shotgun Metagenomics. *Front Microbiol*, **7**, 459.
9. Griffiths, W.J. and Wang, Y. (2009) Mass spectrometry: from proteomics to metabolomics and lipidomics. *Chem Soc Rev*, **38**, 1882-1896.

10. Zhu, Z., Zhang, F., Hu, H., Bakshi, A., Robinson, M.R., Powell, J.E., Montgomery, G.W., Goddard, M.E., Wray, N.R., Visscher, P.M. *et al.* (2016) Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nature Genetics*, **48**, 481-487.

11. Boyle, E.A., Li, Y.I. and Pritchard, J.K. (2017) An Expanded View of Complex Traits: From Polygenic to Omnigenic. *Cell*, **169**, 1177-1186.

12. Petretto, E., Mangion, J., Dickens, N.J., Cook, S.A., Kumaran, M.K., Lu, H., Fischer, J., Maatz, H., Kren, V., Pravenec, M. *et al.* (2006) Heritability and tissue specificity of expression quantitative trait loci. *PLoS Genet*, **2**, e172.

13. Westra, H.J., Arends, D., Esko, T., Peters, M.J., Schurmann, C., Schramm, K., Kettunen, J., Yaghootkar, H., Fairfax, B.P., Andiappan, A.K. *et al.* (2015) Cell Specific eQTL Analysis without Sorting Cells. *PLoS Genet*, **11**, e1005223.

14. Guelfi, S., Botia, J.A., Thom, M., Ramasamy, A., Perona, M., Stanyer, L., Martinian, L., Trabzuni, D., Smith, C., Walker, R. *et al.* (2019) Transcriptomic and genetic analyses reveal potential causal drivers for intractable partial epilepsy. *Brain*, **142**, 1616-1630.

15. Jaffe, A.E., Straub, R.E., Shin, J.H., Tao, R., Gao, Y., Collado-Torres, L., Kam-Thong, T., Xi, H.S., Quan, J., Chen, Q. *et al.* (2018) Developmental and genetic regulation of the human cortex transcriptome illuminate schizophrenia pathogenesis. *Nature Neuroscience*, **21**, 1117-1125.

16. Ram, R., Mehta, M., Nguyen, Q.T., Larma, I., Boehm, B.O., Pociot, F., Concannon, P. and Morahan, G. (2016) Systematic evaluation of genes and genetic variants associated with type 1 diabetes susceptibility. *Journal of Immunology*, **196**, 3043-3053.

17. Yao, C., Joehanes, R., Johnson, A.D., Huan, T., Liu, C., Freedman, J.E., Munson, P.J., Hill, D.E., Vidal, M. and Levy, D. (2017) Dynamic Role of trans Regulation of Gene Expression in Relation to Complex Traits. *American journal of human genetics*, **100**, 571-580.

18. Westra, H.-J., Peters, M.J., Esko, T., Yaghootkar, H., Schurmann, C., Kettunen, J., Christiansen, M.W., Fairfax, B.P., Schramm, K., Powell, J.E. *et al.* (2013) Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nature Genetics*, **45**, 1238-1243.

19. Li, J., Xue, Y., Amin, M.T., Yang, Y., Yang, J., Zhang, W., Yang, W., Niu, X., Zhang, H.-Y. and Gong, J. (2019) ncRNA-eQTL: a database to systematically evaluate the effects of SNPs on non-coding RNA expression across cancer types. *Nucleic Acids Research*, **48**, D956-D963.

20. Cammaerts, S., Strazisar, M., De Rijk, P. and Del Favero, J. (2015) Genetic variants in microRNA genes: impact on microRNA expression, function, and disease. *Frontiers in genetics*, **6**, 186.

21. Huan, T., Rong, J., Liu, C., Zhang, X., Tanriverdi, K., Joehanes, R., Chen, B.H., Murabito, J.M., Yao, C. and Courchesne, P. (2015) Genome-wide identification of microRNA expression quantitative trait loci. *Nature communications*, **6**, 6601.

22. Liu, Z., Ran, Y., Tao, C., Li, S., Chen, J. and Yang, E. (2019) Detection of circular RNA expression and related quantitative trait loci in the human dorsolateral prefrontal cortex. *Genome biology*, **20**, 1-16.

23. Li, S. and Han, L. (2019) Circular RNAs as promising biomarkers in cancer: detection, function, and beyond. *Genome medicine*, **11**, 1-3.

24. Chen, S., Huang, V., Xu, X., Livingstone, J., Soares, F., Jeon, J., Zeng, Y., Hua, J.T., Petricca, J. and Guo, H. (2019) Widespread and functional RNA circularization in localized prostate cancer. *Cell*, **176**, 831-843. e822.

25. Vo, J.N., Cieslik, M., Zhang, Y., Shukla, S., Xiao, L., Zhang, Y., Wu, Y.-M., Dhanasekaran, S.M., Engelke, C.G. and Cao, X. (2019) The landscape of circular RNA in cancer. *Cell*, **176**, 869-881. e813.

26. Ruan, H., Xiang, Y., Ko, J., Li, S., Jing, Y., Zhu, X., Ye, Y., Zhang, Z., Mills, T. and Feng, J. (2019) Comprehensive characterization of circular RNAs in~ 1000 human cancer cell lines. *Genome Medicine*, **11**, 1-14.

27. Oliva, M., Demanelis, K., Lu, Y., Chernoff, M., Jasmine, F., Ahsan, H., Kibriya, M.G., Chen, L.S. and Pierce, B.L. (2023) DNA methylation QTL mapping across diverse human tissues provides molecular links between genetic variation and complex traits. *Nature Genetics*, **55**, 112-122.

28. McVicker, G., van de Geijn, B., Degner, J.F., Cain, C.E., Banovich, N.E., Raj, A., Lewellen, N., Myrthil, M., Gilad, Y. and Pritchard, J.K. (2013) Identification of genetic variants that affect histone modifications in human cells. *Science*, **342**, 747-749.

29. Garieri, M., Delaneau, O., Santoni, F., Fish, R.J., Mull, D., Carninci, P., Dermitzakis, E.T., Antonarakis, S.E. and Fort, A. (2017) The effect of genetic variation on promoter usage and enhancer activity. *Nature Communications*, **8**, 1358.

30. Chen, L., Ge, B., Casale, F.P., Vasquez, L., Kwan, T., Garrido-Martín, D., Watt, S., Yan, Y., Kundu, K., Ecker, S. *et al.* (2016) Genetic Drivers of Epigenetic and Transcriptional Variation in Human Immune Cells. *Cell*, **167**, 1398-1414.e1324.

31. Zhang, D., Cheng, L., Badner, J.A., Chen, C., Chen, Q., Luo, W., Craig, D.W., Redman, M., Gershon, E.S. and Liu, C. (2010) Genetic control of individual differences in gene-specific methylation in human brain. *American journal of human genetics*, **86**, 411-419.

32. Almli, L.M., Stevens, J.S., Smith, A.K., Kilaru, V., Meng, Q., Flory, J., Abu-Amara, D., Hammamieh, R., Yang, R., Mercer, K.B. *et al.* (2015) A genome-wide identified risk variant for PTSD is a methylation quantitative trait locus and confers decreased cortical activation to fearful faces. *Am J Med Genet B Neuropsychiatr Genet*, **168b**, 327-336.

33. Alasoo, K., Rodrigues, J., Danesh, J., Freitag, D.F., Paul, D.S. and Gaffney, D.J. (2019) Genetic effects on promoter usage are highly context-specific and contribute to complex traits. *eLife*, **8**.

34. Suhre, K., McCarthy, M.I. and Schwenk, J.M. (2021) Genetics meets proteomics: perspectives for large population-based studies. *Nature Reviews Genetics*, **22**, 19-37.

35. Pietzner, M., Wheeler, E., Carrasco-Zanini, J., Cortes, A., Koprulu, M., Wörheide, M.A., Oerton, E., Cook, J., Stewart, I.D., Kerrison, N.D. *et al.* (2021) Mapping the proteo-genomic convergence of human diseases. *Science*, **374**, eabj1541.

36. Zheng, J., Haberland, V., Baird, D., Walker, V., Haycock, P.C., Hurle, M.R., Gutteridge, A., Erola, P., Liu, Y., Luo, S. *et al.* (2020) Phenome-wide Mendelian randomization mapping the influence of the plasma proteome on complex diseases. *Nature Genetics*, **52**, 1122-1131.

37. Patti, G., Yanes, O. and Siuzdak, G. (2012) Metabolomics: the apogee of the omics trilogy: Innovation. *Nat Rev Mol Cell Biol*, **13**, 263-269.

38. Johnson, C.H., Ivanisevic, J. and Siuzdak, G. (2016) Metabolomics: beyond biomarkers and towards mechanisms. *Nature reviews Molecular cell biology*, **17**, 451.

39. Gieger, C., Geistlinger, L., Altmaier, E., De Angelis, M.H., Kronenberg, F., Meitinger, T., Mewes, H.-W., Wichmann, H.-E., Weinberger, K.M. and Adamski, J. (2008) Genetics meets metabolomics: a genome-wide association study of metabolite profiles in human serum. *PLoS genetics*, **4**, e1000282.

40. Suhre, K., Shin, S.-Y., Petersen, A.-K., Mohney, R.P., Meredith, D., Wägele, B., Altmaier, E., Deloukas, P., Erdmann, J., Grundberg, E. *et al.* (2011) Human metabolic individuality in biomedical and pharmaceutical research. *Nature*, **477**, 54-60.

41. Lotta, L.A., Pietzner, M., Stewart, I.D., Wittemans, L.B.L., Li, C., Bonelli, R., Raffler, J., Biggs, E.K., Oliver-Williams, C., Auyeung, V.P.W. *et al.* (2021) A cross-platform approach identifies genetic regulators of human metabolism and health. *Nature Genetics*, **53**, 54-64.

42. Surendran, P., Stewart, I.D., Au Yeung, V.P.W., Pietzner, M., Raffler, J., Wörheide, M.A., Li, C., Smith, R.F., Wittemans, L.B.L., Bomba, L. *et al.* (2022) Rare and common genetic determinants of metabolic individuality and their effects on human health. *Nat Med*, **28**, 2321-2332.

43. Shin, S.-Y., Fauman, E.B., Petersen, A.-K., Krumsiek, J., Santos, R., Huang, J., Arnold, M., Erte, I., Forgetta, V. and Yang, T.-P. (2014) An atlas of genetic influences on human blood metabolites. *Nature genetics*, **46**, 543.

44. Long, T., Hicks, M., Yu, H.-C., Biggs, W.H., Kirkness, E.F., Menni, C., Zierer, J., Small, K.S., Mangino, M., Messier, H. *et al.* (2017) Whole-genome sequencing identifies common-to-rare variants associated with human blood metabolites. *Nature Genetics*, **49**, 568-578.

45. Gallois, A., Mefford, J., Ko, A., Vaysse, A., Julienne, H., Ala-Korpela, M., Laakso, M., Zaitlen, N., Pajukanta, P. and Aschard, H. (2019) A comprehensive study of metabolite genetics reveals strong pleiotropy and heterogeneity across time and context. *Nature Communications*, **10**, 4788.

46. Demirkan, A., Henneman, P., Verhoeven, A., Dharuri, H., Amin, N., van Klinken, J.B., Karssen, L.C., de Vries, B., Meissner, A., Göraler, S. *et al.* (2015) Insight in genome-wide association of metabolite quantitative traits by exome sequence analyses. *PLoS Genet*, **11**, e1004835.

47. Köttgen, A., Raffler, J., Sekula, P. and Kastenmüller, G. (2018) Genome-Wide Association Studies of Metabolite Concentrations (mGWAS): Relevance for Nephrology. *Semin Nephrol*, **38**, 151-174.

48. Wittemans, L.B.L., Lotta, L.A., Oliver-Williams, C., Stewart, I.D., Surendran, P., Karthikeyan, S., Day, F.R., Koulman, A., Imamura, F., Zeng, L. *et al.* (2019) Assessing the causal association of glycine with risk of cardio-metabolic diseases. *Nature Communications*, **10**, 1060.

49. Bonder, M.J., Kurilshikov, A., Tigchelaar, E.F., Mujagic, Z., Imhann, F., Vila, A.V., Deelen, P., Vatanen, T., Schirmer, M., Smeekens, S.P. *et al.* (2016) The effect of host genetics on the gut microbiome. *Nat Genet*, **48**, 1407-1412.

50. Knights, D., Silverberg, M.S., Weersma, R.K., Gevers, D., Dijkstra, G., Huang, H., Tyler, A.D., van Sommeren, S., Imhann, F., Stempak, J.M. *et al.* (2014) Complex host genetics influence the microbiome in inflammatory bowel disease. *Genome Med*, **6**, 107.

51. Davenport, E.R., Cusanovich, D.A., Michelini, K., Barreiro, L.B., Ober, C. and Gilad, Y. (2015) Genome-Wide Association Studies of the Human Gut Microbiota. *PLoS One*, **10**, e0140301.

52. Si, J., Lee, S., Park, J.M., Sung, J. and Ko, G.P. (2015) Genetic associations and shared environmental effects on the skin microbiome of Korean twins. *BMC Genomics*, **16**.

53. Jostins, L., Ripke, S., Weersma, R.K., Duerr, R.H., McGovern, D.P., Hui, K.Y., Lee, J.C., Philip Schumm, L., Sharma, Y., Anderson, C.A. *et al.* (2012) Host–microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature*, **491**, 119-124.

54. Blekhman, R., Goodrich, J.K., Huang, K., Sun, Q., Bukowski, R., Bell, J.T., Spector, T.D., Keinan, A., Ley, R.E., Gevers, D. *et al.* (2015) Host genetic variation impacts microbiome composition across human body sites. *Genome Biol*, **16**, 191.

55. Carter, H., Hofree, M. and Ideker, T. (2013) Genotype to phenotype via network analysis. *Curr Opin Genet Dev*, **23**, 611-621.

56. Wörheide, M.A., Krumsiek, J., Kastenmüller, G. and Arnold, M. (2021) Multi-omics integration in biomedical research – A metabolomics-centric review. *Analytica Chimica Acta*, **1141**, 144-162.

57. Zhu, X., Gerstein, M. and Snyder, M. (2007) Getting connected: analysis and principles of biological networks. *Genes & development*, **21**, 1010-1024.

58. Zhou, G., Li, S. and Xia, J. (2020) In Li, S. (ed.), *Computational Methods and Data Analysis for Metabolomics*. Springer US, New York, NY, pp. 469-487.

59. Xia, J., Gill, E.E. and Hancock, R.E.W. (2015) NetworkAnalyst for statistical, visual and network-based meta-analysis of gene expression data. *Nature Protocols*, **10**, 823-844.

60. Zhou, G., Pang, Z., Lu, Y., Ewald, J. and Xia, J. (2022) OmicsNet 2.0: a web-based platform for multi-omics integration and network visual analytics. *Nucleic Acids Res*, **50**, W527-533.

61. Sonawane, A.R., Weiss, S.T., Glass, K. and Sharma, A. (2019) Network Medicine in the age of biomedical big data. *Frontiers in Genetics*, **10**.

62. Bracken, C.P., Scott, H.S. and Goodall, G.J. (2016) A network-biology perspective of microRNA function and dysfunction in cancer. *Nature Reviews Genetics*, **17**, 719.

63. Barrat, A., Barthélemy, M., Pastor-Satorras, R. and Vespignani, A. (2004) The architecture of complex weighted networks. *Proc Natl Acad Sci U S A*, **101**, 3747-3752.

64. Girvan, M. and Newman, M.E.J. (2002) Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, **99**, 7821-7826.

65. Akhmedov, M., Kedaigle, A., Chong, R.E., Montemanni, R., Bertoni, F., Fraenkel, E. and Kwee, I. (2017) PCSF: An R-package for network-based interpretation of high-throughput data. *PLoS Comput Biol*, **13**, e1005694.

66. Tuncbag, N., Gosline, S.J., Kedaigle, A., Soltis, A.R., Gitter, A. and Fraenkel, E. (2016) Network-Based Interpretation of Diverse High-Throughput Datasets through the Omics Integrator Software Package. *PLoS Comput Biol*, **12**, e1004879.

67. Paulsen, B., Velasco, S., Kedaigle, A.J., Pigoni, M., Quadrato, G., Deo, A.J., Adiconis, X., Uzquiano, A., Sartore, R., Yang, S.M. *et al.* (2022) Autism genes converge on asynchronous development of shared neuron classes. *Nature*, **602**, 268-273.

68. Fitzgerald, K.C., Smith, M.D., Kim, S., Sotirchos, E.S., Kornberg, M.D., Douglas, M., Nourbakhsh, B., Graves, J., Rattan, R., Poisson, L. *et al.* (2021) Multi-omic evaluation of

metabolic alterations in multiple sclerosis identifies shifts in aromatic amino acid metabolism. *Cell reports. Medicine*, **2**, 100424.

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

70. Pons, P. and Latapy, M. (2005), *Computer and Information Sciences-ISCIS 2005: 20th International Symposium, Istanbul, Turkey, October 26-28, 2005. Proceedings 20*. Springer, pp. 284-293.

71. Raghavan, U.N., Albert, R. and Kumara, S. (2007) Near linear time algorithm to detect community structures in large-scale networks. *Physical review E*, **76**, 036106.

72. Alon, U.J.N.R.G. (2007) Network motifs: theory and experimental approaches. **8**, 450-461.

73. Shen-Orr, S.S., Milo, R., Mangan, S. and Alon, U. (2002) Network motifs in the transcriptional regulation network of Escherichia coli. *Nat Genet*, **31**, 64-68.

74. Shalgi, R., Brosh, R., Oren, M., Pilpel, Y. and Rotter, V.J.A. (2009) Coupling transcriptional and post-transcriptional miRNA regulation in the control of cell fate. **1**, 762.

75. Bu, P., Wang, L., Chen, K.-Y., Srinivasan, T., Murthy, P.K.L., Tung, K.-L., Varanko, A.K., Chen, H.J., Ai, Y. and King, S.J.C.s.c. (2016) A miR-34a-Numb feedforward loop triggered by inflammation regulates asymmetric stem cell division in intestine and colon cancer. **18**, 189-202.

76. Tarca, A.L., Draghici, S., Khatri, P., Hassan, S.S., Mittal, P., Kim, J.S., Kim, C.J., Kusanovic, J.P. and Romero, R. (2009) A novel signaling pathway impact analysis. *Bioinformatics*, **25**, 75-82.

77. Ma, J., Shojaie, A. and Michailidis, G. (2016) Network-based pathway enrichment analysis with incomplete network information. *Bioinformatics*, **32**, 3165-3174.

78. Glaab, E., Baudot, A., Krasnogor, N., Schneider, R. and Valencia, A. (2012) EnrichNet: network-based gene set enrichment analysis. *Bioinformatics*, **28**, i451-i457.

79. Pavlopoulos, G.A., Malliarakis, D., Papanikolaou, N., Theodosiou, T., Enright, A.J. and Iliopoulos, I. (2015) Visualizing genome and systems biology: technologies, tools, implementation techniques and trends, past, present and future. *Gigascience*, **4**, 38.

80. Cui, W. (2019) Visual Analytics: A Comprehensive Overview. *IEEE Access*, **7**, 81555-81573.

81. Pingault, J.-B., O'Reilly, P.F., Schoeler, T., Ploubidis, G.B., Rijsdijk, F. and Dudbridge, F. (2018) Using genetic data to strengthen causal inference in observational research. *Nature Reviews Genetics*, **19**, 566-580.

82. Hariton, E. and Locascio, J.J. (2018) Randomised controlled trials - the gold standard for effectiveness research: Study design: randomised controlled trials. *Bjog*, **125**, 1716.

83. Davey Smith, G. and Hemani, G. (2014) Mendelian randomization: genetic anchors for causal inference in epidemiological studies. *Hum Mol Genet*, **23**, R89-98.

84. Davey Smith, G. and Ebrahim, S. (2005) What can mendelian randomisation tell us about modifiable behavioural and environmental exposures? *Bmj*, **330**, 1076-1079.

85. Smith, G.D. (2010) Mendelian randomization for strengthening causal inference in observational studies: application to gene× environment interactions. *Perspectives on Psychological Science*, **5**, 527-545.

86. Brion, M.-J.A., Benyamin, B., Visscher, P.M. and Smith, G.D. (2014) Beyond the single SNP: emerging developments in Mendelian randomization in the "Omics" era. *Current Epidemiology Reports*, **1**, 228-236.

87. Nitsch, D., Molokhia, M., Smeeth, L., DeStavola, B.L., Whittaker, J.C. and Leon, D.A. (2006) Limits to causal inference based on Mendelian randomization: a comparison with randomized controlled trials. *Am J Epidemiol*, **163**, 397-403.

88. Davies, N.M., Holmes, M.V. and Davey Smith, G. (2018) Reading Mendelian randomisation studies: a guide, glossary, and checklist for clinicians. *Bmj*, **362**, k601.

89. Pierce, B.L. and Burgess, S. (2013) Efficient design for Mendelian randomization studies: subsample and 2-sample instrumental variable estimators. *Am J Epidemiol*, **178**, 1177-1184.

90. Hemani, G., Zheng, J., Elsworth, B., Wade, K.H., Haberland, V., Baird, D., Laurin, C., Burgess, S., Bowden, J., Langdon, R. *et al.* (2018) The MR-Base platform supports systematic causal inference across the human phenome. *eLife*, **7**.

91. Li, C., Wu, A., Song, K., Gao, J., Huang, E., Bai, Y. and Liu, X. (2021) Identifying Putative Causal Links between MicroRNAs and Severe COVID-19 Using Mendelian Randomization. *Cells*, **10**.

92. Huang, C., Shi, M., Wu, H., Luk, A.O.Y., Chan, J.C.N. and Ma, R.C.W. (2022) Human Serum Metabolites as Potential Mediators from Type 2 Diabetes and Obesity to COVID-19 Severity and Susceptibility: Evidence from Mendelian Randomization Study. *Metabolites*, **12**.

93. Lawlor, D.A., Tilling, K. and Davey Smith, G. (2017) Triangulation in aetiological epidemiology. *International Journal of Epidemiology*, **45**, 1866-1886.

94. Hammerton, G. and Munafò, M.R. (2021) Causal inference with observational data: the need for triangulation of evidence. *Psychol Med*, **51**, 563-578.

95. Elsworth, B., Dawe, K., Vincent, E.E., Langdon, R., Lynch, B.M., Martin, R.M., Relton, C., Higgins, J.P. and Gaunt, T.R. (2018). Oxford University Press.

96. Kilicoglu, H., Shin, D., Fiszman, M., Rosemblat, G. and Rindflesch, T.C. (2012) SemMedDB: a PubMed-scale repository of biomedical semantic predications. *Bioinformatics*, **28**, 3158-3160.

97. Elsworth, B. and Gaunt, T.R. (2021) MELODI Presto: a fast and agile tool to explore semantic triples derived from biomedical literature. *Bioinformatics*, **37**, 583-585.

98. Liu, Y., Elsworth, B., Erola, P., Haberland, V., Hemani, G., Lyon, M., Zheng, J., Lloyd, O., Vabistsevits, M. and Gaunt, T.R. (2021) EpiGraphDB: a database and data mining platform for health data science. *Bioinformatics*, **37**, 1304-1311.

99. Chang, L., Zhou, G., Ou, H. and Xia, J. (2022) mGWAS-Explorer: Linking SNPs, Genes, Metabolites, and Diseases for Functional Insights. *Metabolites*, **12**.

100. Chang, L., Zhou, G., Soufan, O. and Xia, J. (2020) miRNet 2.0: network-based visual analytics for miRNA functional analysis and systems biology. *Nucleic Acids Res*, **48**, W244-w251.

101. Chang, L. and Xia, J. (2023) MicroRNA Regulatory Network Analysis Using miRNet 2.0. *Methods Mol Biol*, **2594**, 185-204.

102. Ward, L.D. and Kellis, M. (2015) HaploReg v4: systematic mining of putative causal variants, cell types, regulators and target genes for human complex traits and disease. *Nucleic acids research*, **44**, D877-D881.

103. McLaren, W., Gil, L., Hunt, S.E., Riat, H.S., Ritchie, G.R., Thormann, A., Flicek, P. and Cunningham, F. (2016) The ensembl variant effect predictor. *Genome biology*, **17**, 122.

104. Piñero, J., Ramírez-Anguita, J.M., Saüch-Pitarch, J., Ronzano, F., Centeno, E., Sanz, F. and Furlong, L.I. (2020) The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic Acids Res*, **48**, D845-d855.

105. Wishart, D.S., Guo, A., Oler, E., Wang, F., Anjum, A., Peters, H., Dizon, R., Sayeeda, Z., Tian, S., Lee, B.L. *et al.* (2022) HMDB 5.0: the Human Metabolome Database for 2022. *Nucleic Acids Res*, **50**, D622-d631.

106. Kanehisa, M. (2017) Enzyme annotation and metabolic reconstruction using KEGG. *Protein Function Prediction: Methods and Protocols*, 135-145.

107. Brunk, E., Sahoo, S., Zielinski, D.C., Altunkaya, A., Dräger, A., Mih, N., Gatto, F., Nilsson, A., Preciat Gonzalez, G.A., Aurich, M.K. *et al.* (2018) Recon3D enables a three-dimensional view of gene variation in human metabolism. *Nat Biotechnol*, **36**, 272-281.

108. Saier, M.H., Reddy, V.S., Moreno-Hagelsieb, G., Hendargo, K.J., Zhang, Y., Iddamsetty, V., Lam, K.J.K., Tian, N., Russum, S., Wang, J. *et al.* (2021) The Transporter Classification Database (TCDB): 2021 update. *Nucleic Acids Res*, **49**, D461-d467.

109. Szklarczyk, D., Gable, A.L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., Simonovic, M., Doncheva, N.T., Morris, J.H. and Bork, P. (2018) STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic acids research*, **47**, D607-D613.

110. Breuer, K., Foroushani, A.K., Laird, M.R., Chen, C., Sribnaia, A., Lo, R., Winsor, G.L., Hancock, R.E., Brinkman, F.S. and Lynn, D.J. (2013) InnateDB: systems biology of innate immunity and beyond--recent updates and continuing curation. *Nucleic Acids Res*, **41**, D1228-1233.

111. Consortium, G.O. (2012) Gene Ontology annotations and resources. *Nucleic acids research*, **41**, D530-D535.

112. Gillespie, M., Jassal, B., Stephan, R., Milacic, M., Rothfels, K., Senff-Ribeiro, A., Griss, J., Sevilla, C., Matthews, L., Gong, C. *et al.* (2022) The reactome pathway knowledgebase 2022. *Nucleic Acids Res*, **50**, D687-d692.

113. Nguengang Wakap, S., Lambert, D.M., Olry, A., Rodwell, C., Gueydan, C., Lanneau, V., Murphy, D., Le Cam, Y. and Rath, A. (2020) Estimating cumulative point prevalence of rare diseases: analysis of the Orphanet database. *Eur J Hum Genet*, **28**, 165-173.

114. Ganter, B., Snyder, R.D., Halbert, D.N. and Lee, M.D. (2006) Toxicogenomics in drug discovery and development: mechanistic analysis of compound/class-dependent effects using the DrugMatrix database. *Pharmacogenomics*, **7**, 1025-1044.

115. Yoo, M., Shin, J., Kim, J., Ryall, K.A., Lee, K., Lee, S., Jeon, M., Kang, J. and Tan, A.C. (2015) DSigDB: drug signatures database for gene set analysis. *Bioinformatics*, **31**, 3069-3071.

116. Stacey, D., Fauman, E.B., Ziemek, D., Sun, B.B., Harshfield, E.L., Wood, A.M., Butterworth, A.S., Suhre, K. and Paul, D.S. (2018) ProGeM: a framework for the prioritization of candidate causal genes at molecular quantitative trait loci. *Nucleic Acids Research*, **47**, e3-e3.

117. Xia, J. and Wishart, D.S. (2010) MSEA: a web-based tool to identify biologically meaningful patterns in quantitative metabolomic data. *Nucleic Acids Res*, **38**, W71-77.

118. Zhu, L., She, Z.G., Cheng, X., Qin, J.J., Zhang, X.J., Cai, J., Lei, F., Wang, H., Xie, J., Wang, W. *et al.* (2020) Association of Blood Glucose Control and Outcomes in Patients with COVID-19 and Pre-existing Type 2 Diabetes. *Cell Metab*, **31**, 1068-1077.e1063.

119. Metwally, A.A., Mehta, P., Johnson, B.S., Nagarjuna, A. and Snyder, M.P. (2021) COVID-19-Induced New-Onset Diabetes: Trends and Technologies. *Diabetes*, **70**, 2733-2744.

120. Rajpal, A., Rahimi, L. and Ismail-Beigi, F. (2020) Factors leading to high morbidity and mortality of COVID-19 in patients with type 2 diabetes. *Journal of diabetes*, **12**, 895-908.

121. Diener, C., Keller, A. and Meese, E. (2022) Emerging concepts of miRNA therapeutics: from cells to clinic. *Trends Genet*, **38**, 613-626.

122. Rupaimoole, R. and Slack, F.J. (2017) MicroRNA therapeutics: towards a new era for the management of cancer and other diseases. *Nat Rev Drug Discov*, **16**, 203-222.

123. Li, J.-H., Liu, S., Zhou, H., Qu, L.-H. and Yang, J.-H. (2014) starBase v2. 0: decoding miRNA-ceRNA, miRNA-ncRNA and protein–RNA interaction networks from large-scale CLIP-Seq data. *Nucleic acids research*, **42**, D92-D97.

124. Malkani, S., Chin, C.R., Cekanaviciute, E., Mortreux, M., Okinula, H., Tarbier, M., Schreurs, A.S., Shirazi-Fard, Y., Tahimic, C.G.T., Rodriguez, D.N. *et al.* (2020) Circulating miRNA Spaceflight Signature Reveals Targets for Countermeasure Development. *Cell Rep*, **33**, 108448.

125. McDonald, J.T., Enguita, F.J., Taylor, D., Griffin, R.J., Priebe, W., Emmett, M.R., Sajadi, M.M., Harris, A.D., Clement, J., Dybas, J.M. *et al.* (2021) Role of miR-2392 in driving SARS-CoV-2 infection. *Cell Rep*, **37**, 109839.

126. Elsworth, B., Lyon, M., Alexander, T., Liu, Y., Matthews, P., Hallett, J., Bates, P., Palmer, T., Haberland, V., Smith, G.D. *et al.* (2020) The MRC IEU OpenGWAS data infrastructure. *bioRxiv*, 2020.2008.2010.244293.

127. Porcu, E., Sjaarda, J., Lepik, K., Carmeli, C., Darrous, L., Sulc, J., Mounier, N. and Kutalik, Z. (2021) Causal Inference Methods to Integrate Omics and Complex Traits. *Cold Spring Harb Perspect Med*, **11**.

128. Amar, D., Sinnott-Armstrong, N., Ashley, E.A. and Rivas, M.A. (2021) Graphical analysis for phenome-wide causal discovery in genotyped population-scale biobanks. *Nature Communications*, **12**, 350.

Appendices

Copyright permissions

Chapter 2:

Chapter 3:

Supplementary tables

Supplementary Table 1. A summary of the databases and software mentioned in the thesis.

Database/Software Name	Versions	PubMed ID or URL	Notes
miRTarBase	v8.0	31647101	Experimentally validated
TarBase	v8.0	29156006	Experimentally validated
miRecords	v1.0	18996891	Experimentally validated
SM2miR	v1.0	23220571	Experimentally validated
Pharmaco-miR	v1.0	23376192	Experimentally validated Computationally Predicted
HMDD	v3.2	30364956	Experimentally validated
miR2Disease	v1.0	18927107	Curated
PhenomiR	v2.0	20089154	Curated
EpimiR	v1.0	24682734	Curated
starBase	v2.0	24297251	Experimentally validated
TransmiR	v2.0	30371815	Curated
ADmiRE	v1.0	30302893	Annotated
PolymiRTS	v3.0	24163105	Experimentally validated
SNP2TFBS	v1.0	27899579	Computationally Predicted
TSmiR	v1.0	24889152	Curated
IMOTA	v1.0	28977416	Experimentally validated
ExoCarta	v1.0	19810033	Experimentally validated
TAM	v2.0	29878154	Curated
Exo-miRExplorer	v1.0	28203233	Experimentally validated
DisGeNET (gene-disease)	v7.0	31680165	Curated, Animal models,

DisGeNET (variant-disease)	v7.0	31680165	Inferred and Literature Curated and Literature
KEGG	2021	28451977	Curated
TCDB	2021	33170213	Curated
Recon3D	2018	29457794	Curated
HMDB	v5.0	34986597	Curated
VEP	v1.0	27268795	Computationally Predicted
InnateDB	v1.0	23180781	Curated
STRING	v10.0	25352553	Curated
HuRI	2020	32296183	Curated
Gene Ontology	v1.0	30395331	Curated
Reactome	v1.0	34788843	Curated
Orphanet	v1.0	https://www.orpha.net/consor/cgi-bin/index.php	Curated
DrugMatrix	v1.0	https://ntp.niehs.nih.gov/data/drugmatrix/	Curated
DSigDB	v1.0	25990557	Curated
OpenGWAS	v1.0	https://gwas.mrcieu.ac.uk/	Curated
*MELODI Presto	v1.0	32810207	-
GTEx	v8.0	23715323	Experimental
QTLbase	v2.0	31598699	Curated
**Elasticsearch	v1.0	https://www.elastic.co/	-
***SemMedDB	v1.0	23044550	-

*MELODI Presto: a quicker and more agile method to identify overlapping enriched literature elements between any number of exposures and outcomes.

**Elasticsearch: Elasticsearch is an open-source, distributed, RESTful search and analytics engine built on top of Apache Lucene.

***SemMedDB: a PubMed-scale repository of biomedical semantic predications.

List of publications

- **Chang, L.**, & Xia, J. (2023). MicroRNA Regulatory Network Analysis Using miRNet 2.0. *Methods in molecular biology (Clifton, N.J.)*, 2594, 185–204. https://doi.org/10.1007/978-1-0716-2815-7_14
- Pang, Z., Zhou, G., Ewald, J., **Chang, L.**, Hacariz, O., Basu, N., & Xia, J. (2022). Using MetaboAnalyst 5.0 for LC-HRMS spectra processing, multi-omics integration and covariate adjustment of global metabolomics data. *Nature protocols*, 17(8), 1735–1761. https://doi.org/10.1038/s41596-022-00710-w
- **Chang, L.**, Zhou, G., Ou, H., & Xia, J. (2022). mGWAS-Explorer: Linking SNPs, Genes, Metabolites, and Diseases for Functional Insights. *Metabolites*, 12(6), 526. https://doi.org/10.3390/metabo12060526
- Pang, Z., Chong, J., Zhou, G., de Lima Morais, D. A., **Chang, L.**, Barrette, M., Gauthier, C., Jacques, P. É., Li, S., & Xia, J. (2021). MetaboAnalyst 5.0: narrowing the gap between raw spectra and functional insights. *Nucleic acids research*, 49(W1), W388–W396. https://doi.org/10.1093/nar/gkab382
- **Chang, L.**, Zhou, G., Soufan, O., & Xia, J. (2020). miRNet 2.0: network-based visual analytics for miRNA functional analysis and systems biology. *Nucleic acids research*, 48(W1), W244–W251. https://doi.org/10.1093/nar/gkaa467