UNIVERSITY OF CALIFORNIA

Los Angeles

Uncovering the Molecular Networks of Metabolic Diseases Using Systems Biology

A dissertation submitted in partial satisfaction of the requirements for the degree of Doctor of Philosophy in Molecular, Cellular, and Integrative Physiology

by

Le Shu

2018

ABSTRACT OF THE DISSERTATION

Uncovering the Molecular Networks of

Metabolic Diseases Using Systems Biology

by

Le Shu

Doctor of Philosophy in Molecular, Cellular, and Integrative Physiology

University of California, Los Angeles, 2018

Professor Xia Yang, Chair

The past few decades have seen dramatic increase in the prevalence of metabolic diseases (MetDs) including obesity, type 2 diabetes (T2D) and cardiovascular disease (CVD), imposing unprecedented burden on public health worldwide. MetDs stem from a complex interplay of multiple genes and cumulative exposure to environmental risk factors, yet the exact etiology remains elusive. To address this challenge, I embarked interdisciplinary systems biology studies encompassing the development of a multi-omics integration tool, elucidation of genetically perturbed tissue networks shared by T2D and CVD, and examination of environmentally perturbed gene networks by a prevalent endocrine disrupting chemical (EDC). First, I developed a multi-omics integration pipeline named Mergeomics, which consists of independent modules

that 1) leverage multi-omics association data to identify biological processes that are perturbed in disease, and 2) overlay the disease-associated processes onto molecular interaction networks to pinpoint hubs as potential key regulators. Unlike existing tools that are mostly dedicated to specific data type or settings, the Mergeomics pipeline accepts and integrates datasets across platforms, data types, and species. The performance of Mergeomics was demonstrated by both simulation and case studies that include genome-wide, epigenome-wide, and transcriptome-wide datasets of total cholesterol and fasting glucose. I then applied Mergeomics to identify the shared gene networks between CVD and T2D through a comprehensive integrative analysis driven by five multi-ethnic genome-wide association studies (GWAS) for CVD and T2D, expression quantitative trait loci (eQTLs), ENCODE, and tissue-specific gene network models from CVD and T2D relevant tissues. The shared networks captured both known and novel processes underlying CVD and T2D. I also predicted 15 key drivers for the shared gene networks and cross-validated the regulatory role of top key drivers using *in vitro* siRNA knockdown, *in vivo* gene knockout, and two Hybrid Mouse Diversity Panels each comprised of >100 strains. Lastly, I leveraged systems biology approaches to assess the target tissues, molecular pathways, and gene regulatory networks associated with a developmental exposure to the model EDC Bisphenol A (BPA). Prenatal BPA exposure was found to cause transcriptomic and methylomic alterations in the adipose, hypothalamus, and liver tissues in mouse offspring, with cross-tissue perturbations in lipid metabolism as well as tissue-specific alterations in histone subunits, glucose metabolism and extracellular matrix. Network modeling prioritized main molecular targets of BPA, including *Pparg*, *Hnf4a*, *Esr1*, and *Fasn*. Moreover, integrative analyses

identified the association of BPA molecular signatures with MetDs phenotypes in mouse and human. In summary, I presented the community a flexible and robust computational pipeline for multidimensional data integration, and offered mechanistic insights into the genetic and environmental underpinnings of MetDs by exploiting the power of systems biology through both computational and experimental approaches.

The dissertation of Le Shu is approved.

Aldons J Lusis

Matteo Pellegrini

Xinshu Xiao

Xia Yang, Committee Chair

University of California, Los Angeles

2018

DEDICATION

This dissertation is dedicated to Xiaoqing and my parents.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

AA	African Americans
BCAA	Branched chain amino acids
BN	Bayesian networks
BPA	Bisphenol A
CVD	Cardiovascular disease
DEG	Differentially expressed gene
DMC	Differentially methylated CpG
DOHAD	Developmental Origin of Adult Health and Disease
EA	European Americans
ECM	Extracellular matrix
eQTL	Expression quantitative trait loci
EWAS	Epigenome-wide association study
FDR	False discovery rate
FHS	Framingham Heart Study
GWAS	Genome-wide association study
HA	Hispanic Americans
HMDP	Hybrid Mouse Diversity Panel
IUGR	Intrauterine growth retardation
JHS	Jackson Heart Study

KD	Key driver
LD	Linkage disequilibrium
MetDs	Metabolic diseases
MetS	Metabolic syndrome
MSEA	Marker set enrichment analysis
SNP	Single nucleotide polymorphisms
T2D	Type 2 diabetes
TF	Transcription factor
TWAS	Transcriptome-wide association study
WHI	Women's Health Initiative
wKDA	Weighted key driver analysis

ACKNOWLEDGEMENT

My dissertation work wouldn't be possible without the guidance and support from my advisor, Xia Yang, who spared no effort in lighting the path of my doctoral training and future career. Thank you so much for training me into a better researcher, and for making the experience of my doctorate study so fruitful and meaningful that I will forever cherish and benefit from. And I would like to thank my committee for their expert advice on my professional growth. I am also grateful for the support from all my lab mates and fellow researchers and students at UCLA and external collaborative groups, with whom I worked closely during my graduate studies. In addition, I wish to express my gratitude to my parents, who made me what I am today. Finally, the past several years would never have been such a joyful ride for me without the company of Xiaoqing. Every moment with you has been so wonderful, and I am the luckiest to have you in my life.

Chapter 2 is a version of Shu L, Zhao Y, Kurt Z, Byars SG, Tukiainen T, Kettunen J, Orozco LD, Pellegrini M, Lusis AJ, Ripatti S, Zhang B, Inouye M, Makinen VP, Yang X. "Mergeomics: Multidimensional data integration to identify pathogenic perturbations to biological systems." BMC Genomics. 2016;17:874. The work was supported by American Heart Association Scientist Development Grant 13SDG17290032, Leducq Foundation, and NIH R01DK104363 to XY; American Heart Association Postdoctoral Fellowship 13POST17240095 to VPM; China Scholarship Council and UCLA Eureka Scholarship to LS; Australian NHMRC Grant 1062227 and 1061435, and Australian Heart Foundation Grant 100038 to MI, SB.

Chapter 3 is a version of Shu L, Chan KHK, Zhang G, Huan T, Kurt Z, Zhao Y, Codoni V, Tregouet DA, Cardiogenics C, Yang J, Wilson JG, Luo X, Levy D, Lusis AJ, Liu S, Yang X. "Shared genetic regulatory networks for cardiovascular disease and type 2 diabetes in multiple populations of diverse ethnicities in the united states." PLoS Genetics. 2017;13:e1007040. The study is partially supported by an AHA CVGPS pathway grant (SL, XY, DL, JGW, AJL, LS, KHKC, TH, ZK, YZ) and the Transatlantic Networks of Excellence Award (12CVD02) from Foundation Leducq (LS, ZK, YZ, VC, DAT, AJL, XY). LS is supported by UCLA Eureka Scholarship, Hyde Scholarship, Burroughs Wellcome Fund Inter-School Program in Metabolic Diseases Fellowship, and China Scholarship Council. XY is supported by NIDDK R01DK104363 and AHA 13SDG17290032. SL is supported by AHA and the NHLBI via the Women's Health Initiative (WHI), NIDDK R01DK103699, and the Brown University's China and Brazil Initiatives. Cardiogenics was supported by the European Project reference LSHM-CT-2006-037593. YZ is supported by AHA Postdoctoral Fellowship 16POST31160044. ZK is supported by AHA Postdoctoral Fellowship 17POST33670739. The Framingham Heart Study is funded by National Institutes of Health contract N01-HC-25195. The Jackson Heart Study is supported by contracts HHSN268201300046C, HHSN268201300047C, HHSN268201300048C, HHSN268201300049C, HHSN268201300050C from the National Heart, Lung, and Blood Institute and the National Institute on Minority Health and Health Disparities.

The work in **Chapter 4** is supported by UCLA Dissertation Year Fellowship, Eureka Scholarship, Hyde Scholarship, Burroughs Wellcome Fund Inter-School Program in Metabolic

Diseases Fellowship, and China Scholarship Council for LS. XY is supported by NIH DK104363 and NS103088, and Leducq Foundation.

VITA

EDUCATION

2018	PhD Candidate, Molecular, Cellular, and Integrative Physiology University of California, Los Angeles
2013	Bachelor of Medicine, Preventive Medicine Fudan University, Shanghai, China

RESEARCH EXPERIENCE

2013 – Present	PhD trainee Department of Integrative Biology and Physiology University of California, Los Angeles
2013 – 2015	Student researcher Department of Epidemiology Fudan University School of Public Health

PEER-REVIEWED PUBLICATIONS (*, co-first author)

Shu L*, Chan KHK*, Zhang G, Huan T, Kurt Z, Zhao Y, Codoni V, et al. Shared genetic regulatory networks for cardiovascular disease and type 2 diabetes in multiple populations of diverse ethnicities in the united states. ***PLoS genetics***. 2017;13:e1007040

Shu L, Zhao Y, Kurt Z, Byars SG, Tukiainen T, Kettunen J, Orozco LD, et al. Mergeomics: Multidimensional data integration to identify pathogenic perturbations to biological systems. ***BMC genomics***. 2016;17:874

Shu L, Blencowe M, Yang X. Translating GWAS findings to novel therapeutic targets for coronary artery disease. ***Frontiers in cardiovascular medicine***. In press.

Arneson D*, **Shu L***, Tsai B, Barrere-Cain R, Sun C, Yang X. Multidimensional integrative genomics approaches to dissecting cardiovascular disease. ***Frontiers in cardiovascular medicine***. 2017;4:8

Axelsson A, Mahdi T, Nenonen H, Singh T, Hänzelmann S, Wendt A, Bagge A, Reinbothe T, Millstein J, Yang X, Zhang B, Gusmao E, **Shu L**, et al. Sox5 regulates beta-cell phenotype and is reduced in type 2 diabetes. ***Nature communications***. 2017;8:15652

Chen Y, **Shu L**, Qiu Z, Lee DY, Settle SJ, Que Hee S, Telesca D, Yang X, Allard P. Exposure to the bpa-substitute bisphenol s causes unique alterations of germline function. ***PLoS genetics***. 2016;12:e1006223

Arneson D, Bhattacharya A, **Shu L**, Makinen VP, Yang X. Mergeomics: A web server for identifying pathological pathways, networks, and key regulators via multidimensional data integration. ***BMC genomics***. 2016;17:722

BOOK CHAPTERS

Shu L, Arneson D, Yang X. Bioinformatics Principles for Deciphering Cardiovascular Diseases. In: Vasan R., Sawyer, D.(eds.) ***The Encyclopedia of Cardiovascular Research and Medicine***, vol.[1], pp. 273-292. 2018. Oxford: Elsevier.

SELECTED MANUSCRIPTS IN REVISION (*, co-first author)

Shu L, Meng Q, Tsai B, Diamente G, Chen Y, Mikhail A, Luk H, Ritz B, Allard P, Yang X. Prenatal bisphenol A exposure in mice induces multi-tissue multi-omics disruptions linking to cardiometabolic disorders.

Zhou M*, Dong W*, Wu C*, **Shu L***, Shao J, Wynn R, Liu Y, et al. Targeting Catabolic Defect of Branched-Chain Amino Acid to Treat Insulin Resistance.

Shao J*, Zhou M*, **Shu L***, Yu J, Liu Y, Wang J, Wang J, et al. A leucine catabolic intermediate governs adipogenesis via activating mTORC1 and polyamine synthesis from methionine.

SELECTED CONFERENCE PRESENTATIONS

Shu L, Zhao Y, The GENESIS Consortium, Lusis AJ, Hao K, Quertermous T, Knowles JW, Yang X. Identification of Genetic Regulatory Networks for Insulin Resistance in Multiple Populations of Diverse Ethnicities. Poster presentation at: ***Basic Cardiovascular Sciences Scientific Sessions***. 2017.

Shu L, Meng Q, Tsai B, Yang X. Prenatal Exposure to Bisphenol A Perturbs Gene Networks and Key Regulators of Metabolic Disorders. Oral presentation at: ***Obesity Society Annual Meeting***. 2015.

Shu L, Chan K, Liu S, Yang X. Integrative Genomics Analyses Identifies Shared Regulatory Networks for Cardiovascular Disease and Type 2 Diabetes Mellitus in American Women of Caucasian, African and Hispanic Ethnicities. Poster presentation at: ***American Heart Association Scientific Session***. 2015.

Chapter 1. Introduction

MetDs such as type 2 diabetes (T2D) and cardiovascular diseases (CVD) are among the top 10 leading causes of death, according to World Health Organization and Center of Disease Control and Prevention. The recent growing MetDs epidemic has caused the number of diabetic and pre-diabetic persons in the U.S. to reach over 40% of the population [1] and the prevalence of CVD to be as high as 11.3% [2]. Although the exact etiology of MetDs remains elusive, mounting scientific evidence supports the role of genetic and environmental factors as well as the complicated interplay between the two [3]. A comprehensive understanding of the molecular mechanisms underlying the actions of genetic and environmental factors is, therefore, critical for our understanding of the molecular basis of MetDs and for the development of effective therapeutic and preventive strategies to reduce the prevalence of MetDs.

It has become increasingly recognized that the tightly regulated coordination among genes through tissue-specific networks underlies higher level physiological processes, and the elucidation of interactions among genes has led to significant insights into biological processes and disease etiology [4-6]. Accompanied by the advent of high-throughput technologies, a large number of gene networks have been constructed based on different algorithms like coexpression networks or Bayesian networks [4, 7]. Building upon these resources, previous studies have shown that the disruption of the specific part of these networks, termed “subnetworks”, by genetic and environmental risk factors, could confer risks toward MetDs [6, 8-13]. Therefore, rather than focusing on the investigation of isolated genes or linear pathways, it is more appealing to elucidate the gene-gene interactions to dissect the molecular mechanisms of

complex diseases like MetDs. More importantly, gene networks could be seamlessly integrated into a larger and more comprehensive systems genetics framework that takes advantage of vast amount of omics data available [14]. The value of omics data is usually limited when individual types of data are used in isolation. For example, although genome-wide association study (GWAS) provides unbiased information about the genetic basis of diseases, it lacks 1) sufficient power to adequately explain heritability and gene-by-environment interaction and 2) mechanistic connection between the risk loci and downstream events. By integrating different layers of biological information, scientists are better enabled to bridge the gap between genetic predisposition and observed phenotypes and investigate molecular interactions in a clinically relevant context. Indeed, systems genetics has already proven to be effective in revealing novel biological process and gene interactions underlying complex diseases [15-18]. Moreover, systems level information can be further modeled into tissue-specific network topology, which will offer an unbiased holistic view of how individual molecular events coordinate their effects and interact within and between different tissues during pathogenesis. Different from conventional candidate gene-driven target selection, the network-driven approach makes it possible to pinpointing key regulatory hubs, or key drivers (KDs) of the disease related subnetworks. These key drivers shall have the potential to normalize the entire gene network spectrum and serve as novel therapeutic targets [19-21]**.** Despite these progresses, well-defined high-throughput systems biology tools that effectively convert large-scale biological data and network resources into meaningful outputs are still lacking. Additionally, applications of systems biology approaches to understand the genetic and environmental mechanisms of disease pathogenesis remain limited.

Aiming to address the aforementioned gaps, my work is centered on 1) the development of "Mergeomics", a generic pipeline that integrates genetic associations, tissue-specific functional genomics resources, canonical pathways and gene-gene interaction networks, 2) applying Mergeomics to a large-scale systems biology analysis on MetDs using available GWAS data on both CVD and T2D in conjunction with a multitude of independent genomic datasets, to identify crucial shared subnetworks and key regulators for CVD/T2D, 3) investigating the health impact of bisphenol A, a recognized environmental endocrine disruptor, by evaluating the perturbation at transcriptome and methylome following prenatal BPA treatment, exploring their potential regulatory relationship, and further analyzing the molecular signatures using network modeling to identify key network regulators mediating the effect of BPA challenge.

The methodological details and performance evaluation of Mergeomics is described in **Chapter 2.** Unlike existing tools that are mostly dedicated to specific data type or settings, the Mergeomics pipeline accepts and integrates datasets across platforms, data types, and species. I optimized and evaluated the performance of Mergeomics using simulation and multiple independent datasets, and benchmarked the results against alternative methods. I also demonstrated the versatility of Mergeomics in two case studies that include genome-wide, epigenome-wide, and transcriptome-wide datasets from human and mouse studies of total cholesterol and fasting glucose. In both cases, the Mergeomics pipeline provided statistical and contextual evidence to prioritize further investigations in the wet lab. Moreover, the pipeline is disseminated to the scientific community as an open source R package as well as a user-friendly webserver to enable broad usage.

In **Chapter 3**, I applied Mergeomics to a multitude of multi-omics datasets to unravel the shared gene networks and their key regulatory genes for both CVD and T2D. The datasets included five multi-ethnic genome-wide association studies (GWAS) for CVD and T2D, expression quantitative trait loci (eQTLs), ENCODE, and tissue-specific gene network models (both co-expression and graphical models) from CVD and T2D relevant tissues. I identified pathways regulating the metabolism of lipids, glucose, and branched-chain amino acids, along with those governing oxidation, extracellular matrix, immune response, and neuronal system as shared pathogenic processes for both diseases. Further, I uncovered 15 key drivers including *HMGCR, CAV1, IGF1* and *PCOLCE,* whose network neighbors collectively account for approximately 35% of known GWAS hits for CVD and 22% for T2D. Finally, I cross-validated the regulatory role of the top key drivers using *in vitro* siRNA knockdown, *in vivo* gene knockout, and two Hybrid Mouse Diversity Panels each comprised of >100 strains. Findings from this in-depth assessment of genetic and functional data from multiple human cohorts provide strong support that common sets of tissue-specific molecular networks drive the pathogenesis of both CVD and T2D across ethnicities and help prioritize new therapeutic avenues for both CVD and T2D.

In **Chapter 4**, I leveraged systems biology approaches to systemically assess the target tissues, molecular pathways, and gene regulatory networks associated with a developmental exposure to BPA. Prenatal BPA exposure led to tissue-specific transcriptomic and methylomic alterations in the adipose, hypothalamus, and liver tissues. I identified cross-tissue perturbations in lipid metabolism and oxidative phosphorylation pathways as well as tissue-specific alterations in histone subunits and glucose metabolism in adipose tissue, extracellular matrix in hypothalamus, and PPAR signaling in liver. Network analyses prioritized main molecular targets of BPA

effects, including *Cyp51*, *Pparg*, *Hnf4a*, estrogen receptors, *Fasn*, and *Acss2*. Lastly, the BPA molecular signatures were found to be significantly associated with cardiometabolic phenotypes or diseases in both the mouse model and human studies. The multi-omics systems biology investigation provides strong evidence that BPA perturbs diverse molecular networks in central and peripheral tissues, and offers insights into the tissue sensitivity and molecular targets that link BPA to cardiometabolic disorders.

In summary, the bioinformatics tool and the systems-level mechanistic information derived from my work is critical to elucidate the gene regulatory circuits and their response upon genetic and environmental challenges that determine metabolic reprograming and disease susceptibility. The molecular insights obtained open new avenues for future in-depth experimental investigation, and provide promising candidates as novel drug targets or tissue-specific biomarkers for translational research in clinical settings.

Chapter 2. Mergeomics: Multidimensional data integration to identify pathogenic perturbations to biological systems

2.1 Introduction

Most non-communicable diseases stem from a complex interplay between multiple genes, transcripts, proteins, metabolites and cumulative exposure to environmental risk factors [22]. In recent years, the advance of omics technologies has greatly enhanced our ability to measure the patterns of molecular entities and interactions at genome-scale. Public data repositories such as dbGaP for population-based genetic datasets [23] and Gene Expression Omnibus and ArrayExpress for gene expression and epigenomics datasets [24, 25] are continuously expanding with new experiments, and data acquisition projects such as ENCODE and GTEx are generating multidimensional datasets on the regulatory processes that link DNA variation with intermediate molecular traits and, ultimately, physiological or pathophysiological phenotypes [26-28]. Genome-wide association studies (GWAS), transcriptome-wise association studies (TWAS), epigenome-wide association studies (EWAS), and metabolome- and proteome-wide association studies have become commonplace in modern biomedical research. Therefore, data integration and interpretation has emerged as a new bottleneck on the road to discovery.

The combination of multiple omics studies is appealing, since a single genomic dataset is unlikely to provide deep mechanistic insight. Instead of one obvious candidate, most omic-wide studies produce a pattern of univariate statistical signals without a clear indication of what would be a suitable target for interventions [29, 30]. However, by integrating different types of data, converging patterns usually emerge and the search space for possible mechanisms can be greatly

reduced. For instance, simultaneous measurement of DNA and RNA (genetics of gene expression) allows investigators to see if a particular genetic variant affects the downstream expression of a gene [31, 32], and functional data such as transcription factor binding, epigenetic modification, or protein regulation from the ENCODE project [33, 34] can be used to further focus on the most promising candidates.

Multi-dimensional data integration has been previously addressed by pathway-based tools such as MAGENTA [35], iGSEA4GWAS [36], SSEA [15], and other network-based methods such as WGCNA [37], postgwas [38], dmGWAS and EW_dmGWAS [39], DAPPLE [40], NetWAS [41], and MetaOmics [42] have been developed to identify the biological processes (e.g. pathways) and specific genes or molecules that may be involved in pathogenesis. However, the available methods are typically tailored for a particular application area (e.g. human genetics with gene expression, protein-protein interactions, or metabolomics) and may not be suitable for cross-comparison of results across diverse data types. In addition, the majority of the network tools start from a limited set of known top loci or genes, but it may be necessary to include the complete genome-wide patterns of signals for maximum sensitivity. Commercial tools such as Ingenuity (http://www.ingenuity.com) have been available for pathway and network analysis of different types of omics data such as gene expression and genetic data. However, these tools are not open source, hence limiting the accessibility by individual users and lacking the availability of detailed underlying algorithms and proprietary information. Additionally, the commercial tools usually do not provide the flexibility to incorporate different types of networks and pathways. For example, Ingenuity networks are primarily based on gene-gene relationships derived from literature rather than data-driven, tissue-specific network patterns the users may

wish to use. For these reasons, we developed Mergeomics, an open source software to deliver flexible multi-omics integration, to identify pathways and model molecular networks of diseases, and to pinpoint promising targets for further experiments in a streamlined, generic and high-throughput manner.

In this report, we describe the main features of Mergeomics, and present simulations and case studies to demonstrate its performance. Mergeomics is the first publicly available implementation of a proven integrative methodology [18]. It employs two broad areas of analysis: Marker Set Enrichment Analysis (MSEA) identifies disease-associated biological processes via integration of omics-disease association and functional genomics data, and Key Driver Analysis (wKDA) determines the key drivers that are suitable for targeted interventions to these processes. Here, we introduce new algorithms (hierarchical permutations and adaptive test statistics) and new concepts (co-hubs and weighted key drivers) to improve the applicability and performance over previous applications. We also report a case study on circulating cholesterol that shows how multiple human studies can be combined, and another case study on glucose regulation that demonstrates analysis across data types (genome, transcriptome, and epigenome) and species (human and mouse). The source code for Mergeomics is available in Bioconductor (https://www.bioconductor.org/packages/devel/bioc/html/Mergeomics.html).

2.2 Results and discussions

Overview of Mergeomics

Figure 2.1 shows the information flow within the Mergeomics pipeline. The Marker Set Enrichment Analysis (MSEA), combines disease association data (e.g., GWAS, EWAS, TWAS) of molecular markers (e.g., genetic, epigenetic, and transcript variants), functional genomics data from projects such as GTEx and ENCODE, and pre-defined sets of connected genes. The output from MSEA is a ranked list of gene sets. We collectively denote these gene sets – which can be metabolic and signalling pathways, co-expression modules or gene signatures – as 'disease-associated gene sets'. When multiple datasets of the same data type or different data types are available for a given disease or phenotype, the meta-MSEA component that is based on the same principles as MSEA but performs meta-analysis at the gene set level can be utilized. The Weighted Key Driver Analysis (wKDA, on the left in **Figure 2.1** and detailed in **Figure 2.2**) was developed to identify local hubs in a gene network whose neighbours are enriched for genes in the disease-associated gene sets. Henceforth these hubs are referred to as key drivers.

Marker Set Enrichment Analysis

Rationale and design: MSEA is based on the idea that a collection of multiple associations is likely to contain true causal variants even if causality cannot be reliably established by univariate analysis. For instance, if multiple genes in a pathway are implicated, then the pathway as a whole is likely to be causal even if some of the gene signals were false positives. The primary inputs for MSEA include 1) marker to disease association statistics, where markers can be SNPs from

GWAS, genes or transcripts from microarrays or RNA sequencing, epigenetic markers from DNA methylation profiling, metabolites from metabolomics, or proteins from proteomics; 2) assignment of markers to their functional downstream target and 3) sets of functional units of genes that co-operate or interact to perform a biological function or process.

MSEA starts with the conversion of a gene set representing a functional unit into a marker set. The corresponding disease association value for each marker is then collected for analysis. In most cases, the association P-values are used. If there are a large number of small P-values in the marker set compared to what can be expected by chance, we conclude that the gene set we have started from is enriched for disease associations (technical details in Methods). Each step in MSEA is fully customizable: it allows 1) association studies of different types or species; 2) different methods of marker-gene assignments, including expression quantitative trait loci (eQTL), transcription factor binding or sequence-proximity to regulatory or coding sequences; 3) filtering based on user-supplied confounding dependencies such as linkage disequilibrium between genetic markers; and it also utilizes 4) a non-parametric test statistic with multiple user-definable quantile thresholds to automatically adapt to a diverse range of association study datasets with different sample sizes and statistical power. For added applicability, MSEA runs a hierarchical gene-based permutation strategy to estimate null distributions that adjusts for shared markers between genes and gene size.

Parameter optimization: To test the performance and identify optimal parameters of MSEA, we performed simulation tests based on three cholesterol GWAS of varying sample sizes (a Finnish study of 8,330 individuals [43], the Framingham Heart Study with 7,572 participants [44], and

GLGC with 100,184 people [45]) and a set of known causal lipid homeostasis genes from the Reactome pathway R-HSA-556833, "Metabolism of lipids and lipoproteins". We resampled genes from this pathway into positive control signals of different magnitudes, and generated negative control signals from the gene pool excluding known cholesterol genes. This procedure was repeated 100 times, and performance was evaluated as sensitivity, specificity and positive likelihood ratio, as described in Methods.

We identified two important parameters, the percentage of top markers included and the threshold for confounding marker dependencies, that affect the performance of MSEA (**Table S2.1**). The signal to noise ratio typically improved when genetic loci with relatively stronger associations (e.g., top 50% markers) rather than the full association sets were used (**Figure 2.3**). This confirms previous findings for complex traits that variants in the extremely weak association spectrum add noise and contribute little to disease biology [46]. For GWAS, linkage disequilibrium is a source of confounding marker dependencies. MSEA was less sensitive to LD thresholds for better powered studies such as the GLGC GWAS but smaller studies benefited from less stringent LD cutoffs, presumably due to improved statistical power (**Figure 2.3**). Overall, we chose to use the top 50% of GWAS loci, and an LD cutoff of $r^2 < 0.5$ as the default setting for GWAS studies. Of note, differences due to datasets were typically larger than those due to parameters (**Figure 2.4**) or variations in marker to gene assignment criteria (**Table S2.1**).

Performance comparison with previous methods: We compared the performance of MSEA to MAGENTA [35] and i-GSEA4GWAS [36], two widely used implementations of gene enrichment analysis [47]. Compared to these methods, MSEA differs in test statistics,

confounder adjustment, and flexibility in data input. The same simulated positive and negative control pathways that were used for calibrating MSEA were also used to compare the three different methods (**Figure 2.5)**. The results are similar across all three total cholesterol GWAS: i-GSEA4GWAS lacked specificity and MAGENTA lacked sensitivity, whereas MSEA provided the best balance and receiver operator characteristics. The results remain robust when different false discovery rate (FDR) cutoffs were used (**Figure 2.6)**. Notably, the superior performance of MSEA over the other two established methods is more obvious when the GWAS involved smaller sample size (the Finnish and Framingham studies compared to GLGC) or heterogeneous population (the Framingham study compared to the more homogenous Finnish cohort), making MSEA useful for all types of studies including the underpowered ones. As the above performance comparison based on simulated positive and negative controls may give an unfair advantage to MSEA due to optimized calibration towards the positive controls, we performed additional tests with 1,346 canonical pathways curated by Reactome [48], BioCarta (http://cgap.nci.nih.gov/Pathways/BioCarta_Pathways) and KEGG [49] (**Table 2.1**). Consistent with results from the simulation approach, MSEA captured the largest number of true positive signals (calculated as the number of overlapped significant pathways among all three GWAS, minus the expected number of overlapped pathways from random gene sets).

Meta-MSEA: gene-set level meta-analysis of multiple association studies

Rationale and design: For a disease phenotype, it is typical that multiple association studies of either the same data type (e.g., multiple GWAS) or different data types (e.g., an EWAS plus a TWAS) are available. Aggregating multiple studies of the same disease is an appealing strategy

to increase signal-to-noise ratio, but marker-level integration is usually complicated by technical challenges. Therefore, we developed Meta-MSEA, which performs gene set-level meta-analysis of multiple association studies to avoid the need to match data platforms, species, or ethnicity, an advantage not present in previous methods.

Performance evaluation: Mergeomics was specifically designed to produce output that is suitable for gene set-level meta-analysis (detailed in Methods). In particular, the reported P-values from permutation analysis are always greater than zero, and can be converted to Z-scores by using the inverse Gaussian density function. To demonstrate the practical benefits of Meta-MSEA, we applied Meta-MSEA to the three cholesterol GWAS used in the calibration analysis, and then compared the results with those from the marker-level meta-analysis of the GWAS (denoted as meta-GWAS). While retaining the same-level of specificity, Meta-MSEA showed better sensitivity, positive likelihood ratio and larger area under the ROC curve (**Figure 2.7**). These results suggest that the gene set-level meta-analysis is more powerful than the traditional marker-centric approach to meta-analysis when investigating perturbations to biological processes.

Weighted key driver analysis (wKDA) to detect disease regulators

Rationale and design: The MSEA and meta-MSEA components of Mergeomics identify pathways or co-regulated gene sets that are perturbed in a disease. However, these analyses do not provide information on the detailed interactions between disease genes or help choose which gene to pursue in further mechanistic studies. To provide the answers, the key driver analysis (KDA) was previously developed to detect important hub genes, or key drivers, whose network

neighbourhoods are over-represented with disease associated genes [11, 18]. The key driver concept is based on the projection of the disease-associated gene sets onto a network model of gene regulation that represents molecular interactions in the full system (**Figure 2.2a**). However, the original KDA ignored the edge weight information generated by most network inference algorithms. As edge weight typically represents association strength or reliability of the connection between genes, this data carries valuable topological information. Here, we introduce wKDA, a new algorithm that takes into account edge weights to increase accuracy (**Figure 2.2**). Briefly, the edge weights are encoded as local node strengths in the neighbourhood of a hub, and then aggregated to estimate an effective membership score for a disease-associated pathway (technical details in **Methods**). This approach serves two purposes: firstly, the key driver scores can be quickly recalculated after permuting the node labels thus enabling the empirical estimation of the null distribution and, secondly, the key driver score takes the local connectivity into account when evaluating the impact of a node membership. wKDA starts by searching a network for candidate hub genes and ignores genes with few connections. It then collects the neighbouring genes for each candidate hub, and estimates the contribution of the disease-associated genes within the neighbourhood of the hub. If the contribution is stronger than what would be expected by chance, we conclude that the hub is a key driver of the disease-associated gene sets (**Figure 2.2a**).

If a subnetwork of genes has multiple highly interconnected genes at the center, it is critical to consider them collectively due to the inherent topological redundancy. For practical purposes, we developed the co-hub concept for wKDA (**Figure 2.2b**) by selecting one of the central genes as the independent hub, and the rest as co-hubs. The rationale is two-fold: first, the statistical power

is increased by only considering the independent hubs when adjusting P-values, as they also capture the signals from their respective co-hubs. Second, the co-hub concept is a useful qualitative measure when selecting the most promising subnetworks and key drivers for experimental validation. For instance, if a key driver has co-hubs with known functions, these can give clues as to the role of poorly understood genes. On the other hand, if a key driver is to be perturbed in an experiment, it may be important to incorporate the co-hubs as integral parts of the experimental design.

Performance evaluation of wKDA and comparison with KDA: To evaluate the performance of wKDA in comparison to the unweighted KDA, we used the reproducibility of KDs of a given gene set mapped to independently constructed gene networks as the performance measure. We first set up three disease-associated gene sets as the test gene sets. These included two lipid subnetworks (denoted as Lipid I and Lipid II) derived from a previous study [18] and the R-HSA-556833 (Metabolism of lipids and lipoproteins) pathway from Reactome. To identify KDs of these test gene sets, we also set up four gene regulatory networks of two tissues (2 independent networks per tissue). The gene-gene interaction network models were probabilistic Bayesian gene regulatory networks constructed from multiple adipose and liver datasets (**Table S2.2**). We organized these networks into two independent weighted adipose networks and two independent weighted liver networks using non-overlapping datasets, where edge weight represents the estimated reliability of an edge between genes.

We used the Jaccard overlap index of the identified KD genes between the two independent networks of the same tissue to assess the prediction accuracy of wKDA and KDA (detailed in

Methods). The higher the proportion of KDs replicated between independent networks using a method, the higher the Jaccard overlap index and the higher the reliability and performance of the method. As shown in **Figure 2.8**, the new wKDA outperformed the unweighted KDA for all three test gene sets against independent networks in both tissues. To test the sensitivity of the key driver approach, we also partially randomized the adipose and liver networks as a model of topological noise. As expected, when some of the edges were randomly rewired, the number of consistent key drivers between two independent networks of the same tissue declined, and when all edges were rewired, no consistent key drivers were detected (**Figure 2.8**). Notably, wKDA was able to detect consistent signals even when half the network was rewired, thus demonstrating the inherent robustness of the wKDA concept compared to the unweighted version. Importantly, because wKDA was specifically designed for weighted networks whereas the unweighted KDA mainly focuses on the network topology without considering weight information, key drivers with high-weight (i.e., high reliability) edges between subnetwork genes were preferred by wKDA. This difference likely explains the better reproducibility of wKDA compared to the unweighted KDA.

Case study 1: Application of Mergeomics to multiple cholesterol datasets from different cohorts

In the first case study, we applied the entire Mergeomics pipeline (MSEA, Meta-MSEA, wKDA) to integrate multiple association studies of the same data type with functional genomics and gene networks. We utilized the three cholesterol GWAS from the Finnish, Framingham and GLGC studies described in previous sections, and performed MSEA on individual studies followed by

Meta-MSEA across studies. **Table 2.2** lists the top pathways from Meta-MSEA, and the full results are available in **Table S2.3**. Meta-MSEA yielded more significant P-values than those obtained from the pathway analysis of conventional SNP-level meta-GWAS, which was consistent with the simulated signals in the calibration tests. Importantly, when we only included the Finnish and Framingham studies, the two smaller GWAS in Meta-MSEA, the signals for the top pathways were comparable to GLGC, which has 6 times larger samples size than Finnish and Framingham combined.

We observed 82 significant pathways with a Meta-MSEA P-value < 0.05. The top hits included major lipoprotein and lipid transport pathways and the receptors that mediate lipid transfer to and from lipoprotein particles. Interestingly, we also found 'Cytosolic tRNA aminoacylation' and 'PPAR-alpha activates gene expression', suggesting that these transcriptional regulatory processes are intrinsically intertwined with the traditional concepts of enzyme-driven metabolic pathways in cholesterol biosynthesis and transport. Because of the presence of overlaps in gene memberships between certain curated pathways, we merged the 82 significant pathways into 43 non-overlapping gene "subnetworks" at a Jaccard index cutoff of 20%, and performed a second run of Meta-MSEA using these merged subnetworks to retrieve the top six subnetworks (**Table S2.4**). The strongest signal was observed for Subnetwork 1 ($P < 10^{-16}$) that contained genes encoding key apolipoproteins and lipid transport proteins (such as *LDLR*, *CETP* and *PLTP*). Subnetwork 2 ($P < 10^{-8}$) included genes related to lipid biosynthesis and catabolism (including the statin target *HMGCR*), oxidoreductive enzymes, metalloproteins and mitochondria. Subnetwork 3 represents a biologically intriguing connection between circulating cholesterol and the immune system: it contained proteins that are involved in the transport of fatty acids and

lipids in blood (Albumin and apolipoproteins A1, B, A and L1), collagen genes, and the immunoglobulin family. Subnetwork 4 mainly contained the ATP-binding cassette family of transmembrane transporters responsible for lipid and cholesterol transfer across cell membranes. Subnetwork 5 included genes for metabolizing retinoid, an important mediator of cholesterol transport and Subnetwork 6 may reflect the connection between transcriptional regulation with fatty acid metabolism.

Next, we investigated if specific genes could be the key drivers for the aforementioned processes. We applied wKDA to overlay the six cholesterol-associated subnetworks onto gene regulatory networks in liver and adipose tissue. The top five key drivers and their co-hubs are listed in **Table 2.3** with examples of visualization in **Figure 2.9,** and the full results are available in **Table S2.5**. The top adipose key driver for Subnetwork 2 was *ACADVL* (very long chain acyl-CoA dehydrogenase), which catalyzes the first step in mitochondrial beta-oxidation (**Figure 2.9a**). Notably, the two co-hubs for *ACADVL* (*PPARA* and *CIDEA*) are also highly relevant genes for maintaining lipid homeostasis: *PPARA* is one of the master regulators of lipid metabolism with clinically approved class of drugs (fibrates) already in use; *CIDEA* has been linked to apoptosis, and mouse knock-outs have demonstrated significant effects on the metabolic rate and lipolysis [50]. In the liver (**Figure 2.9b**), the top key driver of Subnetwork 2 was *FASN* (fatty acid synthase), which was a key driver in adipose tissue as well. The second top key driver *SQLE* (squalene epoxidase) and its co-hubs *FDFT1, IDI1, MSMO1, NSDHL, HMGCS1* and *ALDOC* either catalyze or regulate cholesterol biosynthesis. *HMGCR,* although not listed as top five key drivers, was a highly significant key driver ($P < 10^{-14}$). Subnetwork 2 and Subnetwork 6 shared multiple common key drivers in the adipose network (**Figure 2.9a**).

These included *ACO2* (aconitase 2), an enzyme that catalyzes citrate to isocitrate in mitochondria, and *ACADVL* and its co-hubs. Perturbations to most of the top key drivers, including *ACADVL, FASN, SCD, ACO2, COL1A2, POSTN, EHHADH, DHCR7, HSD17B7, GC, AQP8, INSIG1,* cause abnormal cholesterol and lipid homeostasis according to the Mouse Gene Informatics database and the International Mouse Phenotyping Consortium [51, 52]. In summary, both literature and experiments support the fundamental role of the key drivers in regulating cholesterol metabolism.

Case study 2: Application of Mergeomics to glucose datasets of various data types and species

The second case study demonstrates the integrated use of human and mouse resources with diverse data types, and it provides insights into the genes involved in glucose metabolism. The human data came from a GWAS of fasting glucose that included 46,186 non-diabetic participants [53]. The mouse data came from the Hybrid Mouse Diversity Panel (HMDP), and comprised a GWAS [54], TWAS [54] and EWAS [55] of glucose. The HMDP datasets were derived from genotyping, gene expression profiling, epigenome profiling and clinical phenotyping of one hundred mouse strains.

The Meta-MSEA approach was applied to all the human and mouse association studies. The top hits captured important glucose homeostasis pathways including glycolysis/gluconeogenesis, beta-cell regulation, incretin homeostasis, adipocytokine signalling and glucose transport (**Table 2.4, full results in Table S2.6**). The results also implicated mechanistic connections between lipid metabolism and glucose level based on the findings of carbohydrate-responsive element-binding protein (ChREBP), steroid biosynthesis and lipid transport. Moreover, we highlighted

alpha-linolenic acids, an essential fatty acid, whose role in glucose control and metabolic health is under active investigation [56]. When comparing the pathway signals across datasets, it is noticeable that the mouse studies yield relatively weaker association strength. This could be partly explained by the tissue-specificity of HMDP data, as gene expression, methylation and eQTLs used in our analysis were all from the liver tissue, which could have missed pathways in non-hepatic tissues. Despite the weak power of the mouse datasets, 8 out the 13 top pathways demonstrated stronger significance across studies than in the human GWAS alone (**Table 2.4**). These results (and the earlier example of circulating cholesterol) demonstrate how Mergeomics was able to identify important biological signals that are subtle in any isolated omics dataset, but consistent across multiple data types and species.

2.3 Conclusions

The explosion of genomics data has shifted the technical challenge from data acquisition to data analysis and interpretation. To respond to the challenge, we developed Mergeomics, a generic pipeline that helps to leverage combined statistical patterns of univariate associations of diverse data types and molecular networks to identify important pathways and key drivers in biological systems. We demonstrated how to use Mergeomics in multi-omics projects with human and animal datasets, and also tested the technical robustness with simulated examples. Through the case studies of cholesterol and glucose regulation, we found that gene networks orchestrated by existing drug targets (such as PPARA and HMGCR) and less known genes (such as ACADVL and collagen genes) potentially regulate circulating cholesterol level, and that both known and novel biological processes likely participate in the genetic and epigenetic regulation of glucose

levels. This evidence supports the biological relevance of Mergeomics output. With simulated and real data we demonstrated the robustness of the algorithms in a wide variety of settings and how Mergeomics outperformed other popular tools. Importantly, the inputs to Mergeomics are fully customizable, and accommodate any source dataset that can be represented by i) univariate associations, ii) hierarchical relationships between markers, genes or gene-sets or iii) weighted (gene) networks. Therefore, Mergeomics can guide hypothesis generation across a wide variety of applications.

We acknowledge the following limitations of Mergenomics. First, the current pipeline only takes disease association strength and static information but not directionality and temporal information into consideration, which fortunately covers the majority of available genomics data, but may limit the detection of additional biological signals. Second, although genetic information, when available, can help infer causal relationships, the bioinformatics analyses from Mergeomics mainly serve to generate testable biological hypotheses rather than directly implicating causality. Therefore, the causal roles of the key driver genes, pathways, and networks inferred by Mergeomics require explicit experimental validation. Despite the limitations, Mergeomics provides the scientific community with the first open source implementation of a methodology that has a proven track record of successful biomedical applications. Future development of Mergeomics will focus on addressing the limitations and improve its flexibility and performance by incorporating directional information, dynamic time-course data, and prediction of potential therapeutic agents.

2.4 Methods

Market set enrichment analysis

The default setting of MSEA takes as input 1) summary statistics from global marker association studies (e.g., GWAS, EWAS, TWAS), 2) measurement of relatedness or dependency between genomic markers, 3) mapping between markers and genes, and 4) functionally defined gene sets (e.g., biological pathways or co-regulated genes). For GWAS, SNPs are first filtered based on the linkage disequilibrium (LD) structure to select for only SNPs that are relatively independent given an LD threshold [18] (See **Appendix)**. For other types of association studies, correlations between co-localized markers may be used. For a given gene set, gene members are first mapped to markers based on the mapping file and then the disease association p values of the corresponding markers are extracted to test for enrichment of association signals based on the following null hypothesis:

Given the set of all distinct markers from a set of N genes, these markers contain an equal proportion of positive association study findings when compared to all the distinct markers from a set of N random genes

We only focus on distinct markers to reduce the effect of shared markers among gene families that are both close in the genome and belong to the same pathway (and presumably have overlapping functionality). Furthermore, our software has a feature that merges genes with shared markers before analysis to further reduce artefacts from shared markers. MSEA uniquely adopts a hierarchical gene-based permutation which estimates the expected distribution of the

test statistic under the null hypothesis by randomly shuffling the gene labels while retaining the assignment of mapped markers to genes, preserving the hierarchical marker-gene-pathway cascade (**Figure 2.10**). As an alternative option, the marker labels can also be shuffled to form the null distribution. Both options are offered in the R package.

To avoid assessing enrichment based on any pre-defined association study p-value threshold (e.g., $p<0.05$) which can mean different association strengths in studies of varying sample size and power, we developed a test statistic with multiple quantile thresholds to automatically adapt to any dataset:

$$\chi = \sum_{i=1}^{n} \frac{O_i - E_i}{\sqrt{E_i} + \kappa}$$

In the formula, *n* denotes the number of quantile points, *O* and *E* denote the observed and expected counts of positive findings (i.e. signals above the quantile point), and $\kappa = 1$ is a stability parameter to reduce artefacts from low expected counts for small gene sets. The frequency of permuted signals that exceed the observation is determined as the enrichment P-value. For highly significant signals where the frequency-based value is zero (i.e. no permuted signal exceeds the observation), we fit a parametric model to the simulated null distribution to approximate the corresponding Z-score (See **Appendix**). For meta-MSEA of multiple association studies, pathway enrichment Z-scores from each dataset are first estimated with MSEA. The meta P value is then estimated by integrating individual Z-scores using the Stouffer's method [57].

Weighted key driver analysis

wKDA utilizes both the network topology information and the edge weight information of a molecular network when available (illustrated in **Figure 2.11**). In wKDA, a network is first screened for suitable hub genes whose degree (number of genes connected to the hub) is in the top 25% of all network nodes (**Figure 2.11**, middle box on the left). We further classify these genes as either independent hubs or co-hubs, where a co-hub is defined as a gene that shares a large proportion of its neighbours with an independent hub (Details in **Figure 2.11**). Once the hubs and co-hubs have been defined, the disease-associated gene sets that were discovered by MSEA or meta-MSEA are overlaid onto the molecular network to see if a particular part of the network is enriched for the disease genes. First, the edges that connect a hub to its neighbours are simplified into node strengths (strength = sum of adjacent edge weights) within the neighbourhood (Plots B-D in **Figure 2.11**), except for the hub itself. For example, the top-most node in Plot C has three edges that connect it with the other neighbours with weights that add up to 7 in Plot D. By definition, the hub at the center will have a high strength which will skew the results, so we use the average strength over the neighbourhood for the hub itself. The reduction of the hub neighbourhood into locally defined node strengths improves the speed of the algorithm and makes it easier to define an enrichment statistic that takes into account the local interconnectivity. In particular, the weighting of the statistic with the node strengths emphasizes signals that involve locally important genes over isolated peripheral nodes. In Plot D of **Figure 2.11**, the overlap between the hub neighbourhood and a hypothetical disease-associated gene set is indicated by the circles around the top three nodes. The sum of the strengths of the disease genes is 15, which represents 57% of the total sum of 26.4 in the neighbourhood (pie chart in Plot D). The final enrichment score is estimated as described below.

The null hypothesis for the enrichment of disease genes within a subnetwork can be expressed as

Weighted key driver H_0 : Given the set of nodes adjacent to a key driver, and with each node having a local strength as estimated by the mutual connectivity, the ratio of the sum of strengths of disease genes to the total sum of strengths of all genes in the key driver subnetwork is equal to the ratio for a randomly selected gene set that matches the number of disease genes.

The test statistic for the wKDA is analogous to the one used for MSEA

$$\chi = \frac{O - E}{\sqrt{E} - \kappa}$$

except that the values *O* and *E* represent the observed and expected ratios of disease genes in a hub neighbourhood. In particular,

$$E = \frac{N_k N_p}{N}$$

is estimated based on the hub degree N_k, disease gene set size N_p and the order of the full network *N*, with the implicit assumption that the weight distribution is isotropic across the network.

Statistical significance of the disease-enriched hubs, henceforth key drivers, is estimated by permuting the gene labels in the network and estimating the P-value based on the simulated null distribution. To control for multiple testing, we perform adjustments in two tiers. First, the P-values for a single subnetwork are multiplied by the number of independent hubs (Bonferroni adjustment). All hubs with adjusted P > 1 are discarded. For random data, the truncated results will be uniformly distributed between 0 and 1, and hence they can be treated as regular P-values.

In the second stage, all the P-values for the subnetworks are pooled and the final FDRs are estimated by the Benjamini-Hochberg method [58].

MSEA performance evaluation

MSEA can be reconfigured depending on the type of dataset and study design. We identified several parameters that could affect the performance of the pipeline such as marker filtering by including top disease/trait associated markers based on a percentage cutoff, marker dependency or relatedness (such as LD) cutoff for pruning redundant markers, and the mapping between genes and markers. Here we focus on the marker filtering percentage and marker dependency cutoff as they represent the two key technical challenges. Of note, the mapping between genes and markers can be defined empirically [31, 33], but we used a chromosomal distance-based approach for testing to make Mergeomics consistent with most of the existing pathway enrichment tools. In fact, for GWAS, the assignment of SNPs to their target genes based on their chromosomal location is the commonly adopted approach in other methods, whereas Mergeomics allows users to apply any available assignment method, including the data from tissue-specific eQTL studies and ENCODE.

To evaluate the performance of MSEA in independent datasets, we collected GWAS summary data for circulating cholesterol from 7,572 individuals in the Framingham Study [44], 8,330 Finnish individuals [43], and 100,184 participants from the Global Lipid Genetics Consortium [45]. Cholesterol metabolism and transport is one of the most studied and understood areas of human biology, which makes cholesterol GWAS [45] an informative dataset for method assessment. The Framingham and Finnish studies are completely independent. The GLGC

dataset is the largest meta study for cholesterol traits and contains the two smaller studies, but the total overlap was less than 10% between the datasets. All participants were predominantly Caucasian descent, and we used the corresponding LD data from HapMap [59] and 1000 genomes project [60] in our analyses to remove redundant SNPs in LD. To determine a suitable combination of parameters and to compare performance of different methods, we simulated true positives and true negatives. True positive signals related to cholesterol and lipid metabolism were collected from the Reactome pathway R-HSA-556833, "the metabolism of lipids and lipoproteins". These genes were grouped into 300 positive control pathways, including 100 with size 25, 100 with size 100, 100 with size 250, respectively. Simultaneously, 300 negative control pathways with the same size distribution as the positive control pathways were generated by randomly selecting genes from the non-cholesterol gene pool consists of 8633 genes from the pathway database of Reactome [48], BioCarta (http://cgap.nci.nih.gov/Pathways/BioCarta_Pathways) and KEGG [49]. These manually generated control pathways were combined with 1,346 original canonical pathways for benchmarking.

The control and canonical pathways were analysed by MSEA and two widely-used existing tools MAGENTA and i-GSEA4GWAS. The latter two tools estimate the genetic associations for each gene, and then test if the aggregate gene score for a pathway is higher than expected. MAGENTA identifies the peak disease-associated SNP for each gene, and then adjusts the statistical significance of the peak SNP according to the size of the gene, LD and other potential confounders to produce the gene score. i-GSEA4GWAS uses a similar approach where a gene is considered significant if it contains any of the top 5% SNPs, and the pathway score is estimated

by comparing the observed ratio of significant genes within the pathway against the expected ratio in the full set of genes that were covered by the GWAS. The performance was evaluated as sensitivity (number of positive control pathways at FDR < 25% divided by total number of positive control pathways), specificity (number of negative control pathways at FDR < 25% divided by total number of negative control pathways) and the likelihood to pick up true positive pathways (Positive Likelihood Ratio), calculated as sensitivity / (1 − specificity).

Integrated analysis of diverse data types across species was tested in the second case study. The datasets included a human GWAS for fasting glucose on 46,186 non diabetic subjects [53], and mouse GWAS, EWAS and TWAS for fasting glucose from the HMDP, which consists of 100 different mouse strains [54, 55]. The epigenome and transcriptome data were generated from the liver tissues from the mouse strains on standard chow diet, and mouse liver eQTLs were used in gene-SNP assignment for the HMDP GWAS data for consistency. For EWAS data, DNA methylation sites were mapped to adjacent genes based on a chromosomal distance of 50kb. All other MSEA parameters were the same as those applied in the cholesterol analysis (see the descriptions of the case studies for more information).

wKDA performance evaluation

We assessed the performance of wKDA based on the robustness of the key driver signals in independent gene networks. We collected Bayesian networks that were constructed from published genomic studies where both DNA and RNA were extracted from adipose and liver tissue samples [61, 62]. The collection of Bayesian networks was split into two independent sets of weighted adipose networks and weighted liver networks from non-overlapping datasets

(**Table S2.2**). Edge weights were quantified based on the consistency of the edge between datasets. Using these networks and three test gene sets related to lipid metabolism as inputs, we ran wKDA to identify liver and adipose key drivers of the lipid gene sets. To benchmark the wKDA performance, we compared wKDA with the previously developed unweighted KDA [11, 18]. The prediction accuracy of wKDA and KDA was evaluated by the Jaccard overlap index between the top key driver genes from the two independent networks of each tissue, which represents the proportion of KDs that can be replicated between independent networks. Jaccard overlap index measures the overlap between two KD sets X and Y each containing lists of KD genes, and is calculated based on the following formula: $\mathrm{overlap}(X, Y) = \frac{|X \cap Y|}{\min(|X|,|Y|)}$. The higher the overlap or replication rates of KDs detected between two independent network using a KDA method, the higher the Jaccard overlap index and the higher the performance of the corresponding method.

2.5 Tables

Table 2.1 Performance comparison of MSEA, MAGENTA and i-GSEA4GWAS across three GWAS datasets using real gene sets.

Method	Marker Incl[a]	LD cutoff[b]	Finnish[c]	Framingham[d]	GLGC[e]	Obs[f]	Exp[g]	True Pos[h]	False Pos[i]	Alpha[j]
MSEA	50%	50%	84	61	77	13.00	0.22	12.78	0.13	0.046
MAGENTA	n/a	n/a	114	81	144	6.00	0.73	5.27	0.70	0.080
i-GSEA4GWAS	n/a	n/a	426	503	32	17.00	3.80	13.20	16.31	0.230

[a] Proportion of top markers included

[b] Linkage disequilibrium r^2 cutoff

[c] Number of pathways with $P < 0.05$ from Finnish cohorts

[d] Number of pathways with $P < 0.05$ from the Framingham Study

[e] Number of pathways with $P < 0.05$ from the Global Lipid Genetics Consortium

[f] Observed overlap between the top pathways from the three datasets

[g] Expected overlap for size-matched randomly picked pathways for the three datasets

[h] Estimated number of true positive signals

[i] Estimated number of false positives based on the alpha level (see next)

[j] Estimated actual significance threshold based on the average number of signals across each dataset, and the number of consistent signals

Table 2.2 Top 15 pathways associated with cholesterol levels out of 1,346 canonical pathways tested in three GWAS datasets.

Pathway	MSEA			Meta-MSEA		Meta-GWAS
	Finnish (n=8330)	Framingham (n=7572)	GLGC (n=100184)	Without GLGC	With GLGC	
Lipid digestion, mobilization, and transport	4.16	5.46	6.15	8.67	13.76	5.00
Lipoprotein metabolism	4.67	4.82	5.94	8.59	13.49	5.41
Chylomicron-mediated lipid transport	4.88	4.87	4.72	8.85	12.61	5.03
Metabolism of lipids and lipoproteins	3.15	1.71	6.15	4.00	8.53	3.56
Cytosolic tRNA aminoacylation	3.58	2.09	1.92	4.77	5.86	2.70
Binding and Uptake of Ligands by Scavenger Receptors	1.88	2.29	3.36	3.46	5.86	2.92
Scavenging by Class A Receptors	1.83	2.22	3.22	3.33	5.62	3.47
Metabolism	1.83	1.48	3.94	2.65	5.36	2.98
PPARA Activates Gene Expression	1.66	2.22	2.83	3.17	5.13	1.33
Retinoid metabolism and transport	1.01	2.75	3.04	2.84	4.94	1.42
Regulation of Lipid Metabolism by Peroxisome proliferator-activated receptor alpha (PPARalpha)	1.32	2.02	2.79	2.64	4.52	1.60
Fatty acid, triacylglycerol, and ketone body metabolism	1.48	1.65	2.49	2.49	4.13	1.56
Clathrin derived vesicle budding	1.91	1.27	2.36	2.50	4.05	1.30
Diseases associated with visual transduction	1.41	1.89	2.18	2.62	4.03	2.34
ABC transporters	1.77	0.89	3.16	1.97	4.01	2.75

*The results are listed as $-\log_{10}$P-values, and the full table is available in Additional file 3. MSEA was run with top 50% of markers and LD cutoff $r^2 < 50\%$. The column 'Meta-GWAS' was estimated according inverse-variance meta-analysis of the cohort specific P-values. The Bonferroni-adjusted 5% significance level for 1,346 independent tests is at $-\log_{10}P = 4.43$.

Table 2.3 Key drivers for cholesterol-associated gene subnetworks. Initially, canonical pathways were evaluated for the enrichment of genetic perturbations to circulating cholesterol.

Subnetworks	$-\log_{10}$ P	Functional annotation	Top adipose KDs			Top liver KDs		
			Key driver	$-\log_{10}$ P	Co-hubs	Key driver	$-\log_{10}$ P	Co-hubs
Subnetwork 1 Lipoprotein	16.0	Lipid transport; cholesterol metabolism; lipoprotein; blood plasma	-	-	-	*SPRY4*	9.5	*ABCG8*
						S100A10	4.5	-
Subnetwork 2 Lipid metabolism	8.1	Lipid metabolism; metalloprotein; oxidoreductase; endoplasmic reticulum	*ACADVL*	33.7	*PPARA, CIDEA*	*FASN*	49.0	*GPAM, ACLY*
			FASN	26.8	*ME1, ACSS2, ACLY, ELOVL6*	*SQLE*	37.4	*FDFT1, IDI1, MSMO1, NSDHL, HMGCS1, ALDOC*
			SCD	24.0	*DNMT3L*	*DHCR7*	26.9	*PMVK, MUM1, FDPS, LSS, RDH11, MVD*
			CCBL2	23.3	-	*HSD17B7*	23.9	-
			ACO2	23.0	*AV075202, GPD2, NDUFV1*	*MMT00007490*	18.8	*HMGCR, LSS, FDFT1, MVD, ACSL3*
Subnetwork 3 Immunoglobulin	6.1	Immunoglobulin V-set	*COL1A1*	12.4	-	*COL6A3*	21.4	-
			COL1A2	9.4	*COL3A1,COL2A1, MFAP2*	*VIM*	11.0	-
			OLFML3	8.8	-	*CCDC3*	10.4	*OLFML3*
			POSTN	8.3	*COL2A1*	*CXCR7*	9.9	-
			FN1	7.2	-	*FBLN2*	9.0	-
Subnetwork 4 ABC transport	5.0	ATP-binding cassette genes	-	-	-	*SPRY4*	12.0	*ABCG8*
						MMT00062095	4.3	-
						S100A10	3.2	-
Subnetwork 5	4.5	Retinoid	-	-	-	*GC*	11.2	*RBP4,APOH*

Retinoid metabolism		metabolism; Visual transduction				*TFPI2*	3.2	-
						AQP8	2.9	-
Subnetwork 6 Transcription	3.8	Transcription regulation; fatty acid metabolism; acyltransferase	*SLC2A5*	18.2	-	*PKLR*	23.6	*MMT00060232, ELOVL6*
			ACADVL	17.7	*PPARA, CIDEA*	*PNPLA5*	19.0	*ACLY, ACACA, PNPLA3*
			CPT2	15.9	-	*PGD*	12.2	-
			EHHADH	15.1	-	*FASN*	11.6	*GPAM,ACLY*
			ACO2	13.7	*AV075202, GPD2, NDUFV1*	*INSIG1*	10.7	-

* As these pathways overlap with each other, non-redundant "subnetworks" were constructed that represent the most shared core genes between overlapping pathways. To verify the association with cholester, enrichment was re-evaluated for the subnetworks (second column in the table). Statistical significance was estimated as described in Table 1. Functional annotations were determined with the DAVID Bioinformatics Tool [63]. Key drivers and co-hubs were determined with the wKDA module within Mergeomics. Bayesian networks from multiple mouse studies were combined to create weighted adipose and liver consensus networks [61, 62]. Gene symbols were translated to human when available.

Table 2.4 Pathways associated with fasting glucose across human and mouse association datasets.

Pathway	MSEA				Meta-MSEA	
	Human	Mouse	Mouse	Mouse		
	GWAS	GWAS	TWAS	EWAS	Value	FDR
Glycolysis / Gluconeogenesis	2.56	0.88	3.84	0.63	4.73	2.22%
Starch and sucrose metabolism	3.67	1.37	3.29	0.17	4.57	2.22%
Regulation And Function Of ChREBP in Liver	3.10	0.93	2.74	0.41	4.08	3.60%
Nuclear Receptors in Lipid Metabolism and Toxicity	5.58	0.48	1.99	0.35	4.00	3.60%
Regulation of gene expression in beta cells	4.16	1.75	1.48	0.19	3.97	3.60%
Type II diabetes mellitus	2.11	1.09	1.48	1.08	3.66	6.00%
Integration of energy metabolism	2.42	0.33	2.17	1.09	3.34	10.82%
Steroid biosynthesis	1.10	2.04	1.27	0.76	3.10	14.34%
alpha-linolenic acid (ALA) metabolism	3.40	0.32	1.72	0.69	3.09	14.34%
Incretin Synthesis, Secretion, and Inactivation	3.24	1.14	0.09	2.06	3.02	14.34%
Adipocytokine signaling pathway	2.55	0.41	1.13	1.26	2.94	14.34%
Chylomicron-mediated lipid transport	0.43	0.89	2.89	1.26	2.92	14.34%
Glucose transport	5.57	1.31	0.27	0.29	2.92	14.34%

***** The results are listed as $-\log_{10}$P-values, and the full table is available in Additional file 5. MSEA was run with top 50% of markers, and an LD cutoff $r^2 < 50\%$ was applied to the GWAS. The column 'Meta-GWAS' was estimated according inverse-variance meta-analysis of the dataset-specific P-values. For the human GWAS, SNPs were assigned to genes based on a 20kb window in the genome sequence. For the mouse GWAS, liver eQTL data were used for gene assignment. The Bonferroni-adjusted 5% significance level for 1,346 independent tests is at $-\log_{10}P = 4.43$.

2.6 Figures

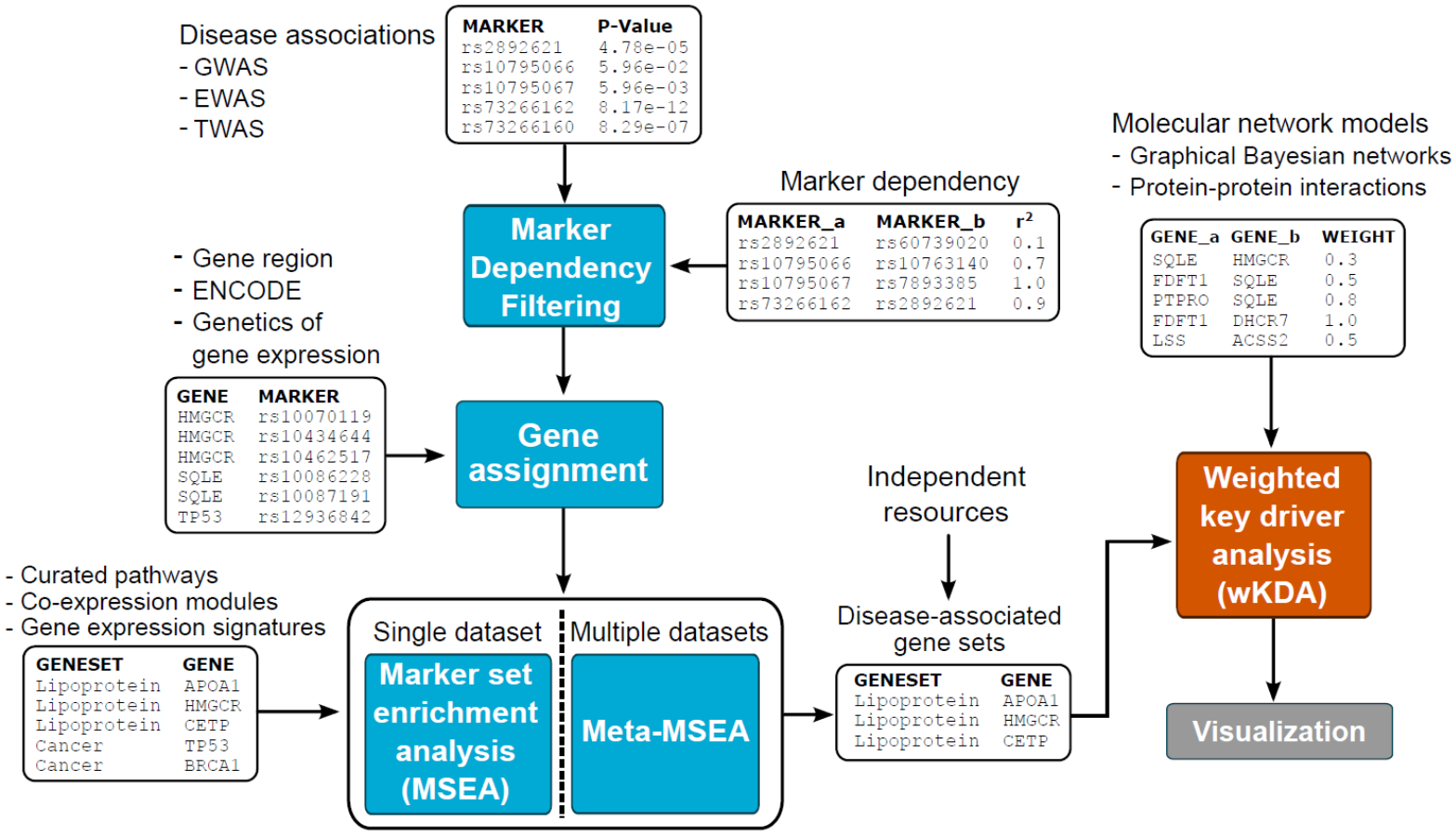


Figure 2.1 Main modules, data flow between them, and examples of data types that can be integrated by Mergeomics.

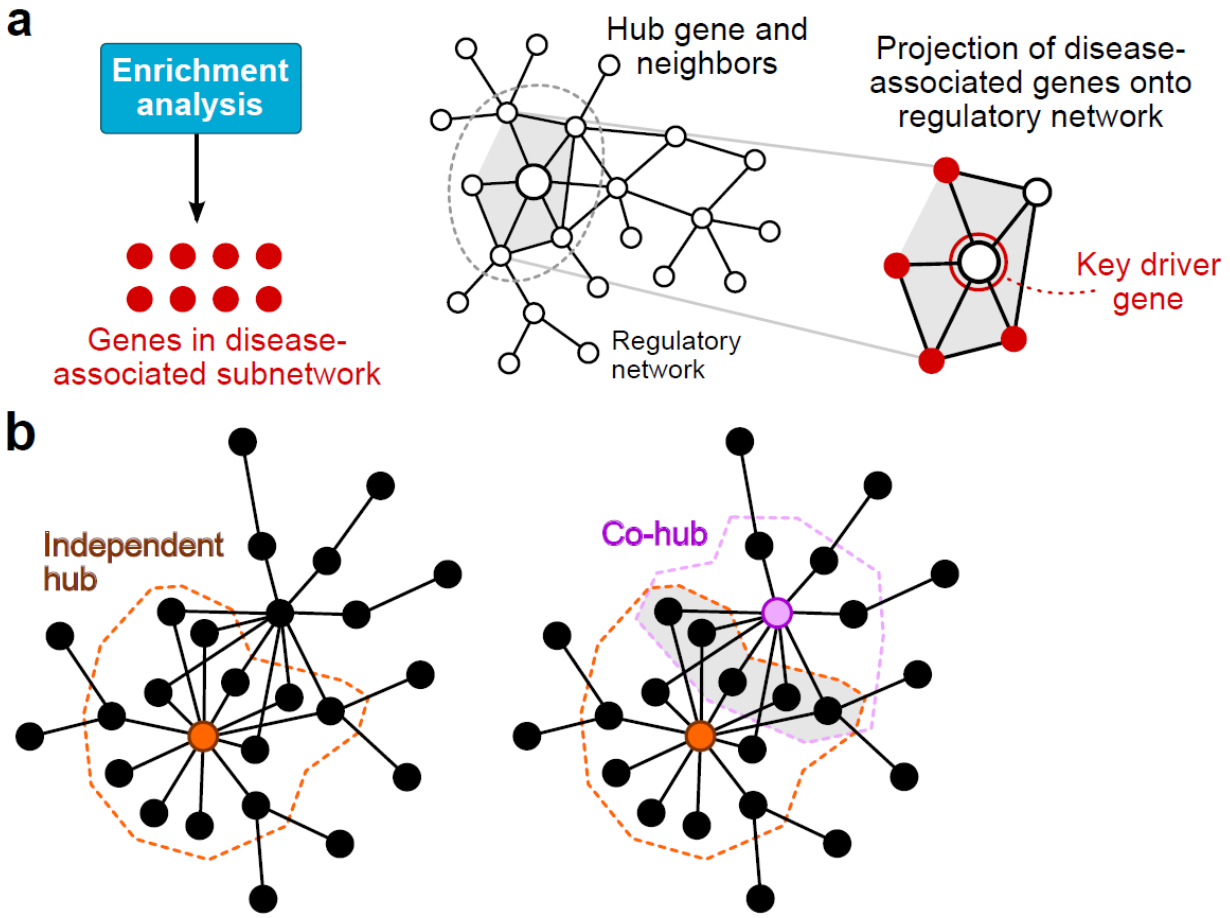


Figure 2.2 Schematic illustration of the concept of a key driver gene (**a**) and local hubs with overlapping neighborhoods (**b**).

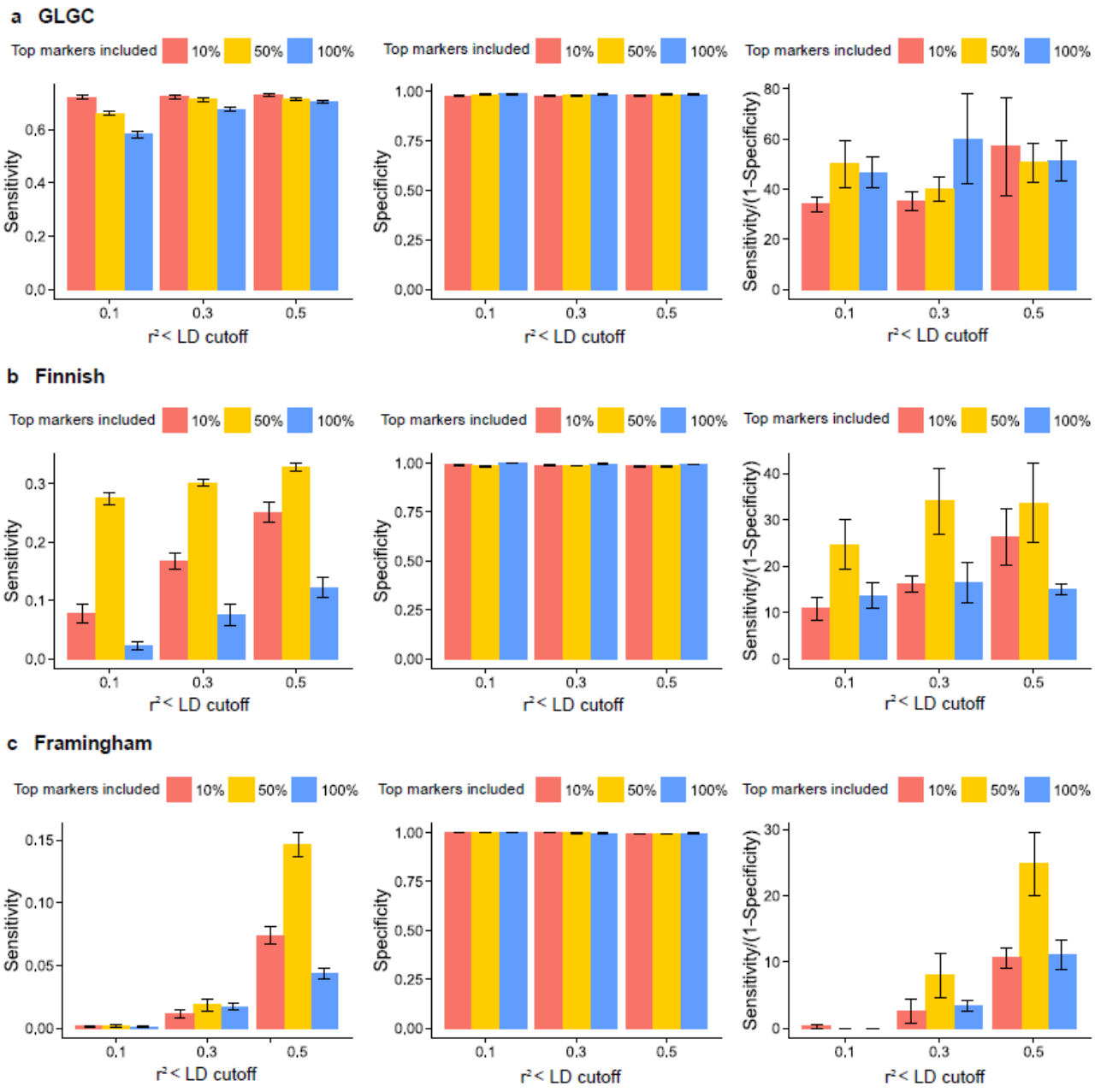


Figure 2.3 Sensitivity, specificity and positive likelihood ratios of MSEA in capturing simulated lipid genesets across combinations of GWAS datasets, top markers included and LD cutoffs.

a) Results from GLGC GWAS. **b**) Results from Finnish GWAS. **c**) Results from Framingham GWAS. 20kb window distance mapping is used for all analyses.

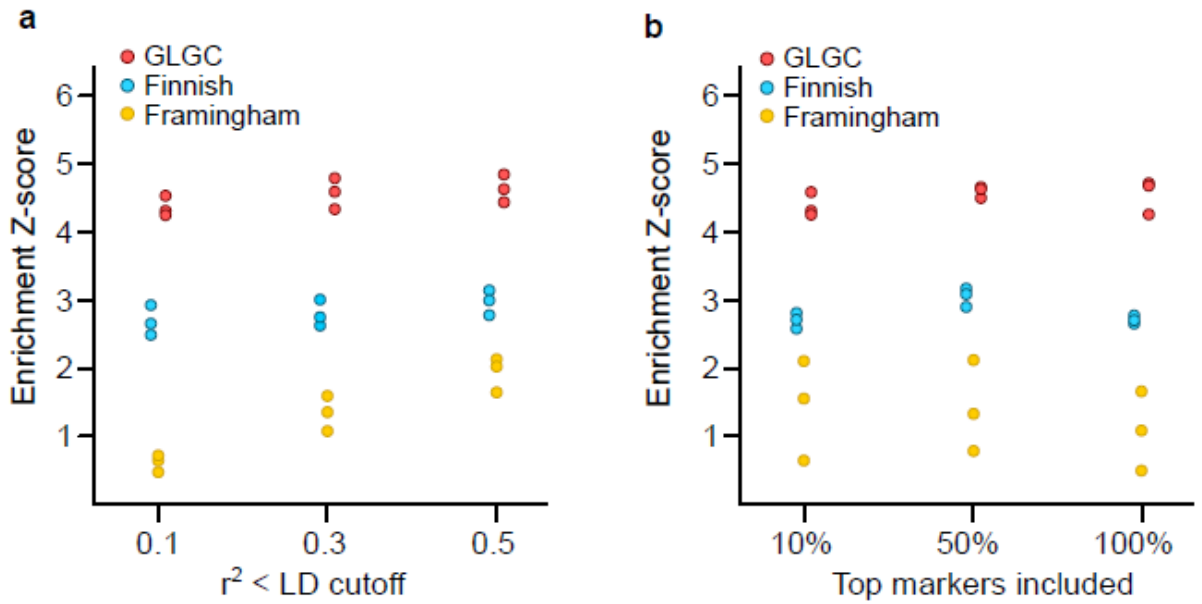


Figure 2.4 MSEA signals of lipid homeostasis genes for combinations of linkage disequilibrium pruning parameter and marker filtering (top markers included).

Results are from 20kb window distance mapping. Z-scores of dots with the same color in the LD cutoff plot (**a**) represent the values from different marker filtering settings, and vice versa in (**b**).

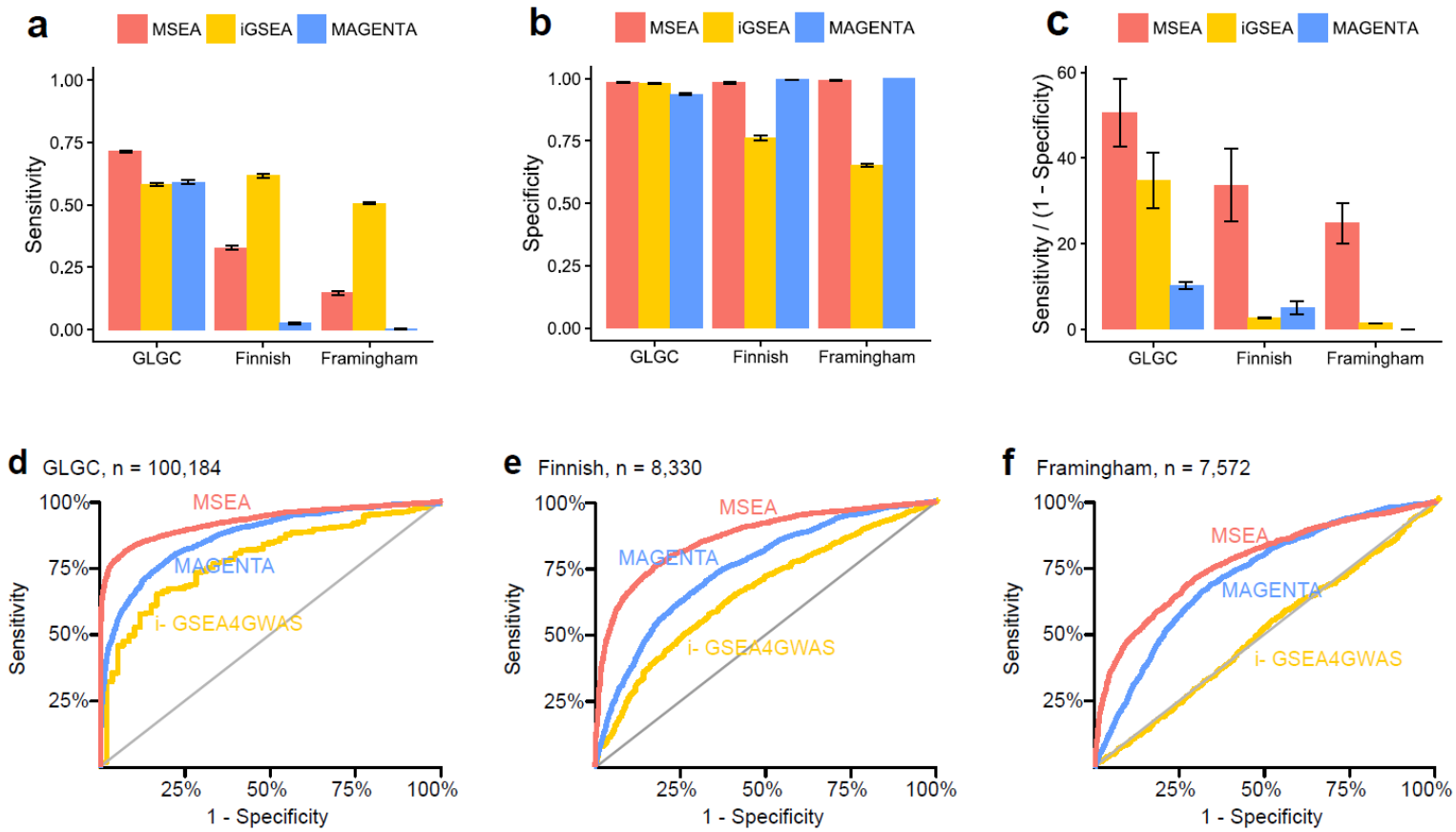


Figure 2.5 Comparison of three pathway enrichment methods across three GWAS.

Performance is evaluated by sensitivity (**a**), specificity (**b**), positive likelihood ratio (sensitivity/(1-specificity)) (**c**) and receiver operating characteristic curve (**d-f**). Sensitivity was defined as the proportion of positive control pathways detected at FDR < 25%. Specificity was defined as the proportion of negative controls rejected at FDR $\geqslant$ 25%. Error bars denote the standard error of simulation results.

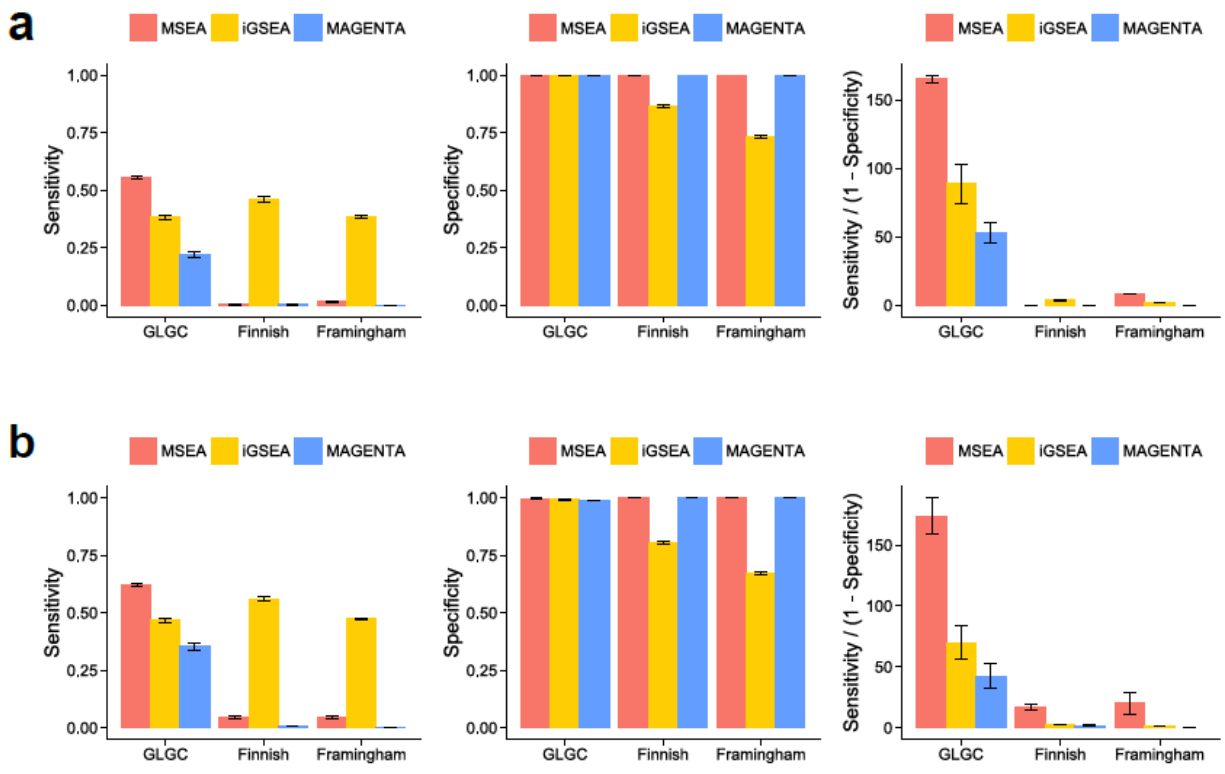


Figure 2.6 Performance comparison of MSEA, iGSEA and MAGENTA at FDR 5% (**a**) and FDR 10% (**b**).

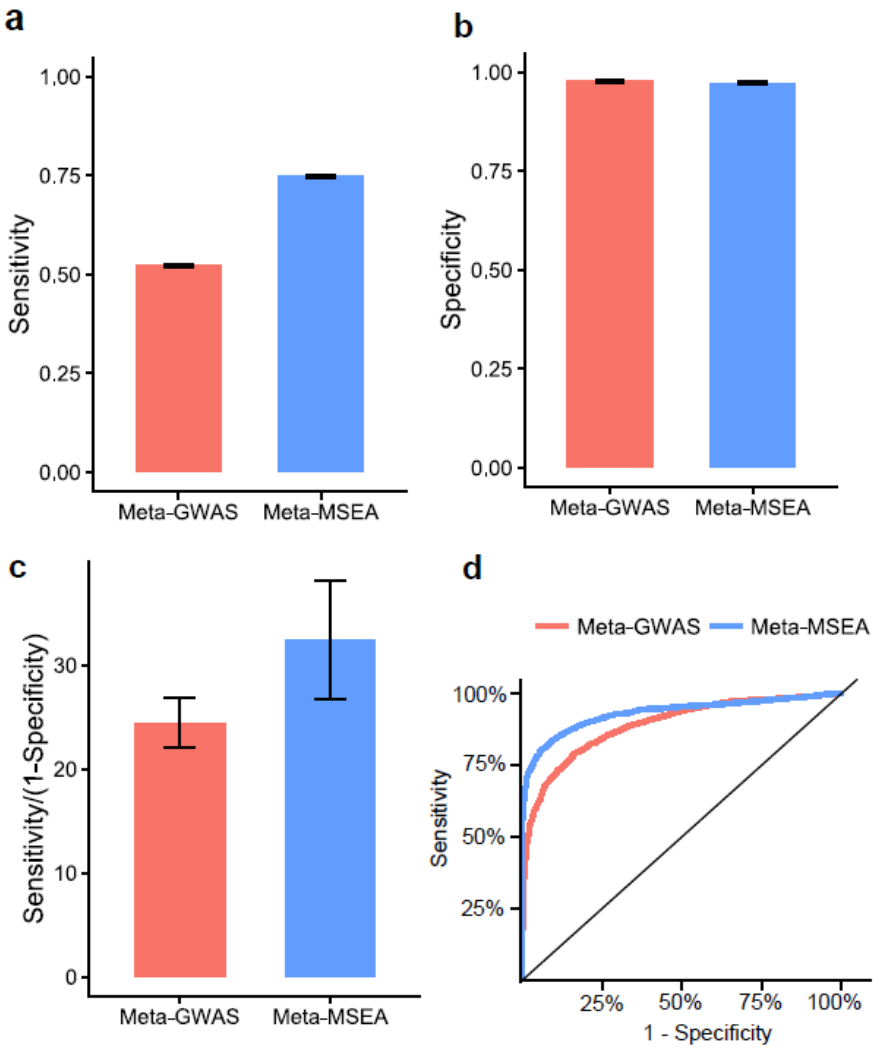


Figure 2.7 Comparison of performance of SNP-level meta-analysis and pathway-level meta-analysis using simulated gene-sets.

Results are produced in the same workflow as stated in **Table 2.2**. **a**) Sensitivity. **b**) Specificity. **c**) Positive likelihood ratio (Sensitivity/(1-Specificity)). **d**) Receiver operating characteristic curve. Error bars denote the standard error of simulation results.

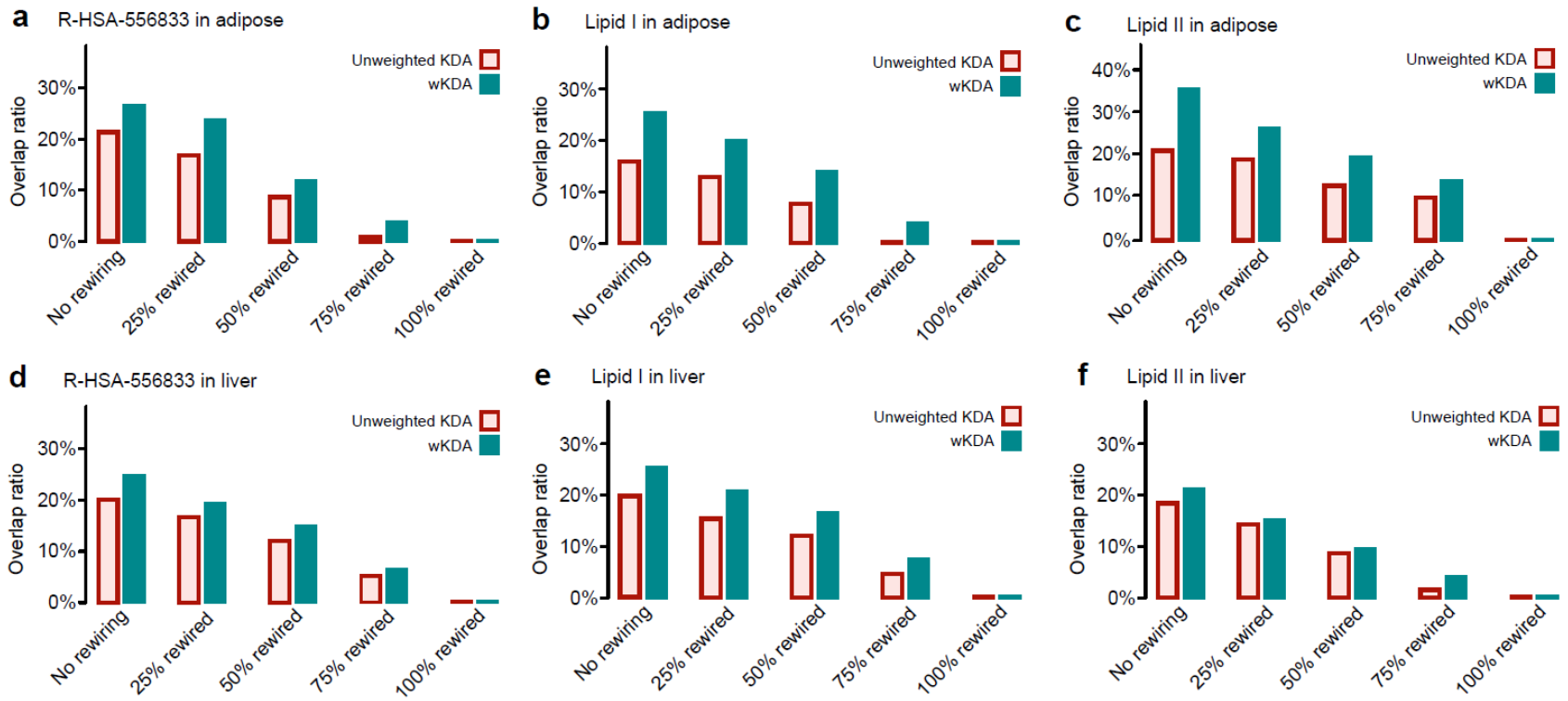


Figure 2.8 Performance comparison between wKDA and the unweighted key driver analysis.

Two empirical subnetworks (Lipid I & II) were obtained from a previous publication [18], and a canonical metabolism of lipids and lipoproteins pathway was obtained from the Reactome database (R-HSA-556833). The methods were tested by projecting the three functional subnetworks onto two independent adipose networks (**a-c**) and two independent liver regulatory networks (**d-f**). The adipose and liver networks were constructed from a collection of Bayesian tissue-specific network models (**Table S2.2**). Overlap between the tissue-specific key driver signals across two independent regulatory networks was defined according to the Jaccard index. Overlap ratio was calculated for both original networks and networks with 25%, 50%, 75% or 100% rewiring of edges.

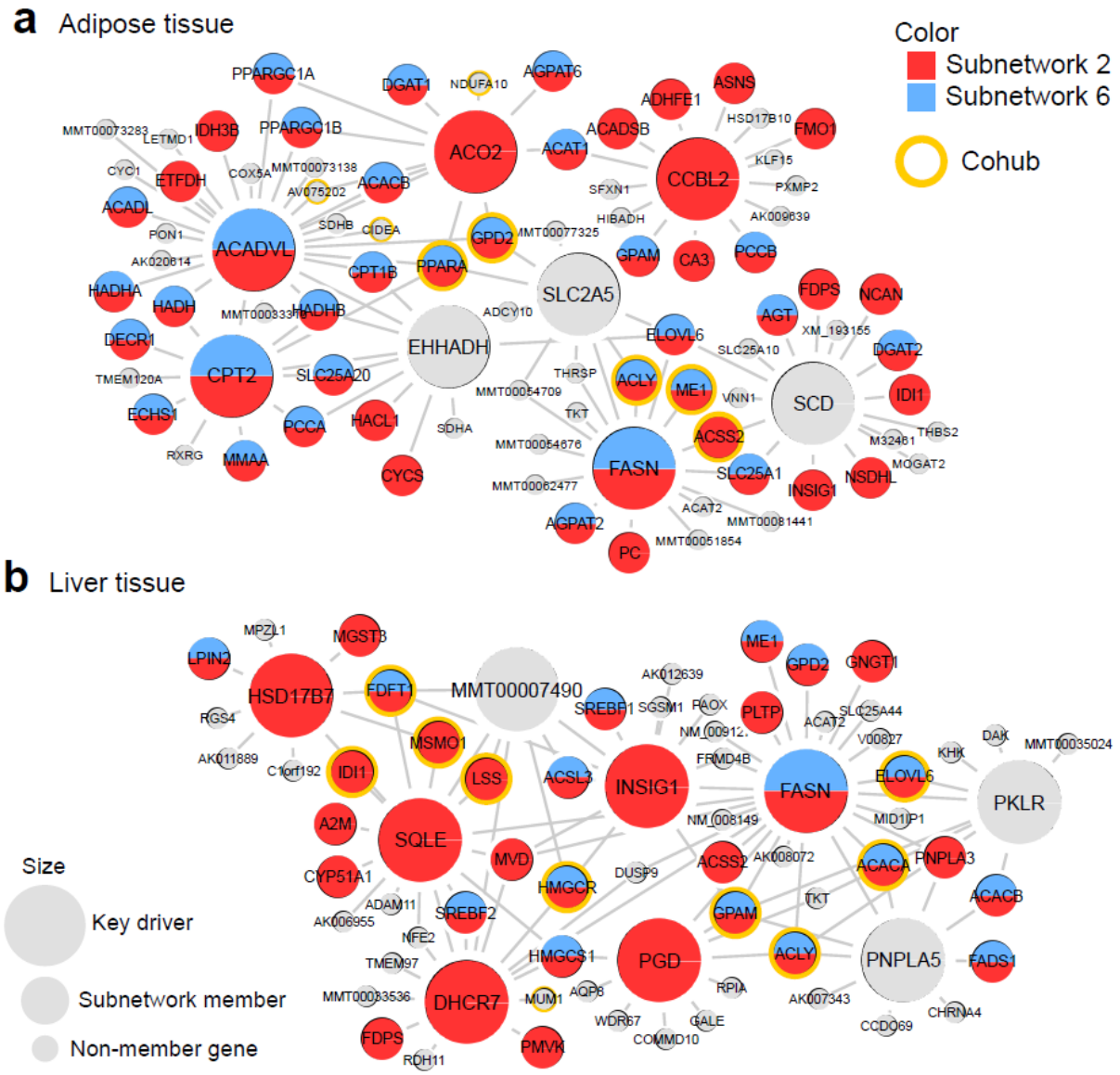


Figure 2.9 Visualization of adipose (**a**) and liver (**b**) networks around top key drivers that were identified for cholesterol-associated subnetworks.

Top key drivers (nodes with the largest size) are selected as the top five independent key regulatory genes (genes whose neighbourhood has less than 25% overlap with the neighbourhood of other independent hubs) for subnetwork 2 and subnetwork 6. Subnetwork member genes are denoted as medium size nodes and non-member genes as small size nodes. Top co-hubs (co-hubs with FDR < 10^{-10} in wKDA) are highlighted by yellow circles. Only edges that were supported by at least two studies are drawn.

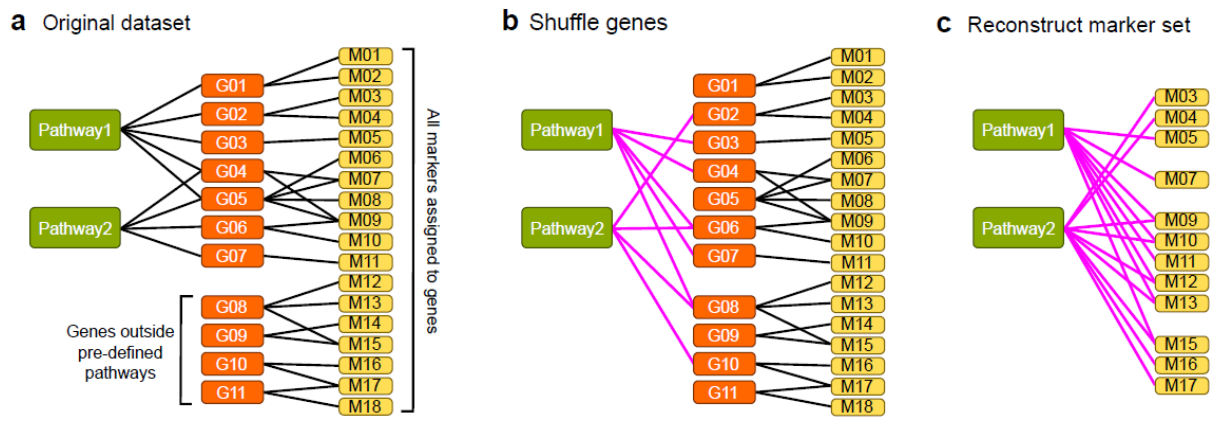


Figure 2.10 Schematic illustration of the hierarchical structure in genetic datasets (**a**) and the randomization procedure that was used in the gene-permuted MSEA (**b**-**c**).

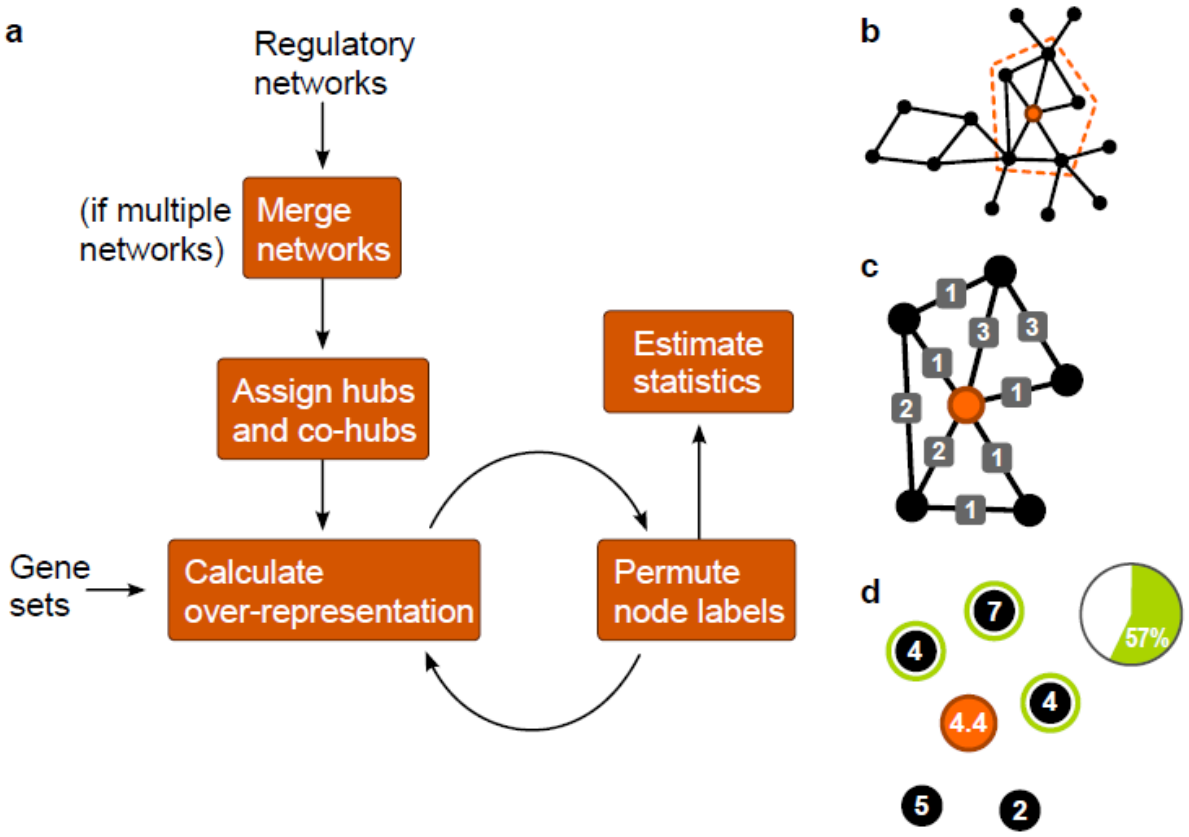


Figure 2.11 Schematic illustration of the weighted key driver analysis (**a**) and the key driver enrichment statistic (**b-d**).

2.7 Appendix

Linkage disequilibrium filtering

The LD filtering algorithm comprises two phases: first, the SNPs are sorted by the disease associations (GWAS data) to prioritize the most important loci. The second phase starts by the selection of the top SNP. The next highest ranked SNP is then compared against the first one to see if LD is acceptable. If not, the SNP is discarded. The same step is then repeated for the remaining list while always checking if the next SNP is in LD with any of the already accepted ones. Subsequently, all accepted SNPs are guaranteed not to be in LD with each other above a designated LD threshold.

Definition of co-hubs in wKDA

The candidate independent hubs are first sorted according to the node degree, from low to high. This is to ensure that we capture local structures rather than one master hub that covers the majority of the network (e.g. housekeeping genes would make poor drug targets due to global side-effects). Next, the sorted hubs are tested one by one for neighbourhood overlaps with the already accepted hubs. If sufficient overlap (as defined under section “definition of overlap between two gene sets” below, default value is 33%) is detected, the current hub is assigned as a co-hub for the previously accepted overlapping hub.

Adaptive Gaussian approximation for estimating P-values in MSEA and wKDA

The exact shape of the null distribution is dependent on the size of the gene set and on the mapping between the genes and the markers (MSEA) or on the size and topology of the gene network (wKDA). To estimate the P-value from these various permutation approaches, we created a generic algorithm for a parametric approximation using the Gaussian function. In the range where a direct frequency-based P-value is accurate (i.e. with 10,000 permutations it is possible to accurately estimate P-values above ~0.001), we found that the Gaussian approximation was highly concordant. For $P < 0.001$, we found that the Gaussian model produced biologically plausible rankings of statistical significance. We tested other models, but found that the potential benefit from using more long-tailed distributions was outweighed by the difficulties in applying them in practice. For instance, the t-distribution was more conservative than the Gaussian estimate, but assigning an appropriate degree of freedom was problematic given the diverse nature of the null hypotheses.

Let X denote the series of simulated test statistics (as defined in the previous section) from the permutation analysis. Then the transformation algorithm can be expressed as

1) $\alpha = \min(X_0)$, $X_1 = X_0 - \alpha$

2) $\beta = \text{median}(X_1)$, $X_2 = X_1 / \beta$

3) $X_3 = \log(\gamma X_2 + 1)$

4) $\mu = \text{mean}(X_3)$, $\sigma = \text{sd}(X_3)$

5) Evaluate how well X_3 approximates $N(\mu, \sigma)$

6) If necessary, try a different γ and go back to Step 3.

The parameters from Steps 1-4 can be saved and reapplied to new data, which makes it possible to determine the transformation exclusively based on simulated statistics, and then apply it to the observed test statistic to yield the parametric enrichment score

$$\mathrm{Z}_N = \frac{\log(\gamma(X-\alpha)\beta^{-1}+1)-\mu}{\sigma}$$

The rationale for Gaussian approximation is based on the attractive analytical properties of Gaussian distributions. Nevertheless, if the approximation is inaccurate, the results can be biased and lead to erroneous conclusion. In particular, any dependencies between markers tend to elongate the tails of the "true" distribution when using marker permutations for the MSEA. For this reason, we also report the raw frequency of false positive findings from the permutation analysis for each gene set.

Chapter 3. Shared Genetic Regulatory Networks for Cardiovascular Disease and Type 2 Diabetes in Multiple Populations of Diverse Ethnicities in the United States

3.1 Introduction

Cardiovascular disease (CVD) and type 2 diabetes (T2D) are two leading causes of death in the United States [64]. Patients with T2D are at two to six times higher risk of developing CVD compared to those without T2D [65], indicating the importance of targeting common pathogenic pathways to improve the prevention, diagnosis, and treatment for these two diseases. While decades of work has revealed dyslipidemia, dysglycemia, inflammation, and hemodynamic disturbances as common pathophysiological intermediates for both CVD and T2D [66-68], very few studies have directly investigated the genomic architectures shared by the two diseases. While genetic factors are known to play a fundamental role in the pathogenesis of both CVD and T2D [69], a direct comparison of the top risk variants between these diseases has revealed few overlapping loci in genome-wide association studies (GWAS) from multiple large consortia. Aside from the speculation that the strongest genetic risks may be disease-specific, the agnostic approach requiring the application of strict statistical adjustment for multiple comparisons also increases false negative rate because of the lack of "genome-wide significance".

To meet these challenges, we and others have previously shown that hidden disease mechanisms can be unraveled through the assessment of the combined activities of genetic loci with weak to moderate effects on disease susceptibility by integrating GWAS with functional genomics and regulatory gene networks [17, 18, 70-72]. Importantly, such high-level integration approaches are able to overcome substantial heterogeneity between independent datasets and extract robust

biological signals across molecular layers, tissue types, and even species [70, 73-75]. This advantage is mainly conferred by the aggregation of genetic signals from individual studies onto a comparable ground – molecular pathways and gene networks, before conducting meta-analysis across studies [75, 76]. In other words, even if the genetic variants and linkage architecture can be different between studies, the biological processes implicated are more reproducible and comparable across studies [77]. In the current investigation, we employed a systematic data-driven approach that leveraged multi-dimensional omics datasets including GWAS, tissue-specific expression quantitative trait loci (eQTLs), ENCODE, and tissue-specific gene networks (**Figure 3.1**). GWAS datasets were from three well-characterized and high-quality prospective cohorts of African Americans (AA), European Americans (EA), and Hispanic Americans (HA) - the national Women's Health Initiative (WHI) [70], the Framingham Heart Study (FHS) [78], and the Jackson Heart Study (JHS) [79]. To maximize the reproducibility of our findings across different populations, we also incorporated meta-analyses of CVD and T2D genetics from CARDIoGRAMplusC4D [80] and DIAGRAM [81]. Further, we comprehensively curated functional genomics and gene networks derived from 25 tissue or cell types relevant to CVD and T2D. A streamlined integration of these rich data sources using our Mergeomics pipeline [75, 76] enabled the identification of shared pathways, gene subnetworks, and key regulators for both CVD and T2D across cohorts and ethnicities. Finally, we validated the subnetworks using adipocyte and knockout mouse models, and confirmed their associations with cardiometabolic traits in the Hybrid Mouse Diversity Panel (HMDP) comprised of >100 mouse strains [54, 82, 83].

3.2 Results

Identification of Co-expression Modules Genetically Associated with CVD and T2D across Cohorts

We first investigated whether genetic risk variants of CVD and T2D from GWAS of each cohort/ethnicity were aggregated in a functionally coherent manner by integrating GWAS with tissue-specific eQTLs or ENCODE information and gene co-expression networks that define functional units of genes (**Figure 3.1A**). Briefly, co-expression networks were constructed from an array of transcriptomic datasets of various tissues relevant to CVD and T2D (details in **Methods**). These modules were mainly used to define sets of functionally related genes in a data-driven manner. Genes within the co-expression modules (a module captures functionally related genes) were mapped to single nucleotide polymorphisms (SNPs) that most likely regulate gene functions via tissue-specific eQTLs or ENCODE information. SNPs were filtered by linkage disequilibrium (LD) and then a chi-square like statistic was used to assess whether a co-expression module shows enrichment of potential functional disease SNPs compared to random chance using the marker set enrichment analysis (MSEA) implemented in our Mergeomics pipeline (details in **Methods**) [75]. Subsequently, meta-analyses across individual MSEA results at the co-expression module level were conducted using the Meta-MSEA function in Mergeomics to retrieve robust signals across studies. Among the 2,672 co-expression modules tested, 131 were found to be significant as defined by false discovery rate (FDR) < 5% in Meta-MSEA across studies (**Table 3.1, Table S3.1**). Moreover, the majority of the disease relevant tissues or cell types included in our analysis yielded informative signal, supporting the systemic

pathogenic perturbations known for CVD and T2D (**Figure 3.2**). Of the significant modules identified, 79 were associated with CVD and 54 with T2D. Two modules were associated with both diseases, with one enriched for "carbohydrate metabolism" genes and the other over-represented with "other glycan degradation; known T2D genes" (**Figure 3.3A, Table S3.1**). Examination of these two shared modules showed that the genetic signals driving the module significance were largely different between CVD and T2D, with 14.8% lead SNPs overlapping for the carbohydrate metabolism module and 5.8% lead SNPs overlapping for the glycan degradation module between diseases. These results indicate that the GWAS signals for the two diseases in each module do not necessarily overlap, but the CVD and T2D genes are likely functionally connected since they are co-expressed in the same modules and annotated with coherent functions. Additionally, the majority of the CVD modules and T2D modules were identified in more than one ethnic group based on MSEA analysis of individual studies, supporting consistency across ethnicities (**Figure 3.3B)**.

Shared Biological Processes among the CVD/T2D-associated Co-expression Modules

Apart from the two directly overlapping modules, between the CVD- and T2D-associated modules there were many overlapping genes, indicating additional shared functions that contribute to both diseases (**Figure 3.4**). Upon annotating the disease-associated modules using functional categories curated in Kyoto Encyclopedia of Genes and Genomes (KEGG) and Reactome while correcting for the overlaps between pathways (method details in **Appendix**; **Figure 3.5**; **Table S3.2**)**,** we found significant functional overlaps between the CVD and T2D modules (overlap p = 3.1e-15 by Fisher's exact test, **Figure 3.3C**). We further ranked all the

enriched functional categories by the number of CVD/T2D modules that were annotated with each functional term (**Figure 3.6**), which showed a wide spectrum of biological processes shared by both CVD and T2D across ethnicities and cohorts. Of the top ranked processes for the significant co-expression modules identified, we observed well-established pathogenic processes such as lipid and fatty acid metabolism [84], glucose metabolism [85], oxidation [86], and cytokine signaling [87]. Pathways previously implicated mainly for T2D such as beta-cell function were also found to be shared for both CVD and T2D. Interestingly, our completely data-driven approach also identified extracellular matrix (ECM) and branched chain amino acids (BCAA) metabolism as top functional categories whose roles in the development of cardiometabolic disorders have only been implicated in recent experimental work [88-90]. Furthermore, our analysis also revealed under-appreciated processes involving the neuronal system and transport of small molecules.

Identification and Prioritization of Key Drivers (KDs) and Subnetworks for the CVD/T2D-associated Modules

The coexpression networks used above mainly served to capture coexpression patterns between genes and to define data-driven gene sets that contain functionally related genes, but they do not carry detailed topology information on gene-gene regulatory relationships. To dissect the gene-gene interactions within and between the 131 disease-associated modules, and to identify key perturbation points shared for both CVD and T2D modules, we used the GIANT networks [41] and Bayesian networks (BNs) from 25 CVD and T2D relevant tissue and cell types, which provide detailed topological information on gene-gene regulatory relationships necessary for the

wKDA analysis. The BNs used in our study were generated using similar sets of mouse and human gene expression datasets as used for the co-expression networks, but additionally incorporated genetic data to model causal gene regulatory networks, whereas the GIANT networks were derived based on independent gene expression datasets and protein interaction information. We included both types of gene regulatory networks to increase the coverage of functional connections between genes and only considered KDs identified in both to enhance the robustness of KD prediction.

Specifically, all genes in each of the co-expression modules genetically associated with CVD or T2D as identified in our study were mapped onto the GIANT and BN graphical networks to identify KDs using the weighted key driver analysis (wKDA) implemented in Mergeomics [75], where KDs were defined as genes whose local network neighborhoods demonstrate significant enrichment of genes from disease-associated modules (details in **Methods**; concept depicted in **Figure 3.7**). Of note, wKDA gives higher weight to network edges that are consistent across network models constructed from independent studies, therefore alleviating potential bias caused by dataset heterogeneity. We identified 226 KDs that were consistently captured in Bayesian and GIANT network at Bonferroni-corrected p-value < 0.05 **(Figure 3.1B)**, among which 162 were KDs for both CVD and T2D associated modules. Bonferroni-correction was used here to focus on the strongest KDs for prioritization purposes. To further prioritize these 162 shared KDs, tissue-specific subnetworks of these KDs were evaluated using Meta-MSEA to rank the magnitude of their genetic association with CVD and T2D across cohorts, yielding 15 top-ranked KDs at FDR<10% in Meta-MSEA for CVD and T2D separately (combined FDR<1% for both diseases simultaneously) **(Figure 3.1B, Table 3.2)**. The top KD subnetworks were related to

similar pathogenic processes highlighted in the previous section, including cholesterol biosynthesis, respiratory electron transport, immune system and ECM. We further inferred the directionality of the effects of each specific KD on both diseases using GWAS signals mapped to each KD based on eQTLs or chromosomal distance (details in **Methods**; results in **Figure 3.8**). This analysis differentiated the KDs into those showing consistent direction of association for both CVD and T2D (*ACLY, CAV1, SPARC, COL6A2, IGF1*), inverse directions with CVD and T2D (*HMGCR, IDI1*), and uncertain directions (**Table 3.2**). Therefore, the shared KDs do not necessarily affect the risks for the two diseases in the same direction.

Shared KDs and Subnetworks Orchestrate Known CVD and T2D Genes

The KDs and subnetworks were identified based on the full spectrum of genetic evidence (from strong to moderate and subtle) from the various GWAS datasets examined in the current study. To assess whether the top KD subnetworks were enriched for previously known disease genes that mostly represent the strong and replicated genes as a means of cross-validation, we manually curated previously reported genes associated with CVD, T2D, and intermediate metabolic traits related to CVD, T2D (glucose, insulin, lipids, obesity) from DisGeNET [91] and the NHGRI GWAS Catalog [69] (**Figure 3.1C,** genes listed in **Table S3.3**). The connection between the top 15 KDs and known genes for CVD, T2D and relevant cardiometabolic traits was confirmed by the significant over-representation of the known disease genes in KD subnetworks, with fold enrichment as large as 8, confirming the strong biological importance of these KDs (**Figure 3.9A**). Further, the top 15 KDs showed direct connections to 28 GWAS hits reaching genome-wide significance ($p < 5e-8$) for CVD and 16 for T2D, which account for 35% (fold = 3.35, p =

7.18e-10) and 22% (fold = 2.16, p = 8.08e-4) of all reported significant GWAS signals for CVD and T2D in GWAS catalog, respectively. Two of the 15 top KDs, namely *HMGCR* and *IGF1*, were previously identified as signals of genome-wide significance for obesity, lipids and T2D, all risk factors of CVD. Additionally, network visualization revealed tissue-specific KDs and interactions of CVD and T2D genes in many disease-relevant tissues including adipose, adrenal gland, artery, blood, digestive tract (small intestine, colon), hypothalamus, islet, liver, lymphocyte, skeletal muscle, thyroid, and vascular endothelium (**Figure 3.9B**). *PCOLCE* represents an intriguing hypothalamus-specific KD that interacts with important energy homeostasis genes like leptin receptor *LEPR*, suggesting a role of neurohormonal control in CVD and T2D pathogenesis. In contrast, *CAV1* appeared to interact extensively with other KDs in peripheral tissues, especially in the adipose tissue.

Experimental Validation of CAV1 Subnetworks using an in vitro Adipocyte Model and in vivo Knockout Mouse Model

CAV1 is a robust KD for CVD- and T2D-associated modules across multiple tissues, with the adipose tissue subnetwork of *CAV1* containing the largest number of neighboring genes (**Figure 3.9B**). In addition, adipose tissue is the only tissue where *CAV1* is a KD in both the Bayesian networks and GIANT networks. These lines of evidence implicate the potential importance of *CAV1* adipose subnetwork in the shared pathogenesis for both diseases. Indeed, *Cav1-/-* mice have been shown to alter the lipid profile, susceptibility to atherosclerosis, and insulin resistance [92, 93]. To assess whether perturbation of this potential KD induces changes in the subnetwork genes as predicted by our network modeling, we performed validation by conducting siRNA-

mediated knock down of *Cav1* in differentiating mouse 3T3-L1 adipocytes and by evaluating the whole transcriptome alteration in mouse gonadal adipose tissue between wild type and *Cav1*$^{-/-}$ mice [92] (**Figure 3.1C**; details in **Methods**). Of the 12 adipose network neighbors of *Cav1* that were tested *in vitro*, 6 exhibited significant changes in expression level on day 2 after ~60% *Cav1* knockdown using two siRNAs against *Cav1*. In contrast, none of the 5 negative controls, which were randomly selected among adipocyte genes that are not connected to *Cav1* or its first level neighbors in the adipose network, were affected after *Cav1* perturbation (**Figure 3.10A**). *Cav1* knockdown also led to decreased expression of *Pparg*, a major adipocyte differentiation regulator (**Figure 3.11**), supporting a role of *Cav1* in adipocyte differentiation as previously observed [94].

In 3-month-old *Cav1*$^{-/-}$ mice which showed perturbed lipid and insulin sensitivity profiles, we observed 1,474 differentially expressed genes (DEGs) at FDR<1%. We found that the first and second level neighbors of *CAV1* in our predicted subnetwork showed significant enrichment for DEGs in adipose tissue induced by *Cav1* knockout, with the degree of fold enrichment increasing as the statistical cutoff used to define DEGs became more stringent (**Figure 3.10B**; subnetwork view with DEGs in **Figure 3.12**). On the contrary, the third and fourth level neighbors of *CAV1* in our predicted subnetwork did not exhibit such enrichment of DEGs (**Figure 3.10B**). These experimental findings support that *CAV1* is a key regulator of the subnetwork and the network structure predicted by our network modeling is reliable, although it is difficult to discern whether the network changes are related to alterations in adipocyte differentiation status. We also observed strong enrichment for the focal adhesion pathway in both

the predicted *Cav1* adipose subnetwork (p=9.6e-14 by Fisher's exact test, fold enrichment = 6.0) and the differential adipose genes in *Cav1*-/- mice (p = 6.9e-9, fold enrichment = 3.5).

Shared KDs Are Associated with CVD and T2D Traits in Experimental Mouse Models

We further assessed the transcriptomic profiling in adipose (relevant to T2D and CVD) and aorta tissue (main site of CVD) in relation to 7 cardiometabolic phenotypes including adiposity, lipid levels (Triglyceride, LDL, HDL), fasting glucose, fasting insulin and HOMA-IR, across >100 mouse strains in two HMDP panels [54, 82, 83]. HMDP is a systems genetics resource that comprises more than 100 commercially available mouse strains differing in genetic composition, and has emerged as a power tool to study complex human diseases [82, 95]. The biological relevance of HMDP to human pathophysiology has been reproducibly demonstrated [96-98]. Moreover, HMDP data was completely independent of the human-focused genetic datasets and the network datasets used in our primary integrative analysis (**Figure 3.1C**). Here we selected two specific HMDP panels, high-fat (HF) and atherogenic (ATH), in which mice were either fed with a high-fat high-sucrose diet or underwent transgenic expression of human APOE-Leiden and CETP gene as a pro-atherogenic background, respectively. These two panels were chosen for their representativeness of human T2D (the HF panel) and CVD (the ATH panel) pathology. First, we investigated the correlation between the expression of 14 top KDs (no probe for KD *MSMO1* in HMDP) and cardiometabolic traits in the adipose and aorta tissues assessed in HMDP. All 14 KDs displayed significant trait association in HMDP, with the association for 11 KDs replicated in both the HF and ATH HMDP panels (**Figure 3.13A**). Next, we retrieved the adipose and aorta gene-trait correlation statistics for the top KD subnetwork genes, and used

MSEA to test whether genes in the KD subnetworks displayed an overall overrepresentation of strong trait association in HMDP. Again, the 14 KD subnetworks showed significant trait association after Bonferroni correction (**Figure 3.13B**). These findings support that the close involvement of the KDs in cardiometabolic trait perturbation we predicted based on human datasets can be cross-validated in mouse models.

Causal Implication of the Shared KD Subnetworks in Experimental Mouse Models

Cav1 knockout in mice led to dysreuglation of the predicted subnetwork (**Figure 3.10B**) and significant alterations in cardiometabolic phenotypes [92, 93], supporting the causal role of *CAV1* in both CVD and T2D. To further investigate the potential causal role of the top KDs and their subnetworks in CVD and T2D, we conducted integrative analysis of the KD subnetworks to assess their disease association using GWAS results for the 7 cardiometabolic traits from HMDP and tissue-specific cis-eQTLs (**Figure 3.1C**). By mapping GWAS signals to genes using adipose or aorta eQTLs and testing for enrichment of genetic association with cardiometabolic traits within the KD subnetwork genes using MSEA, we found consistent and significant association between cardiometabolic traits and the subnetworks of KDs *ACAT2, CAV1*, *COL6A2, IGF1,* *PCOLCE,* and *SPARC* across adipose and aorta (**Figure 3.13C**). These results informed by mouse GWAS support a potential causal role of these top KDs in perturbing gene networks in multiple tissues to trigger CVD and T2D.

3.3 Discussion

CVD and T2D are highly correlated complex diseases and share many common risk factors. Multiple genetic variants may individually exert subtle to strong effects on disease pathogenesis, and in aggregate perturb diverse pathogenic pathways [18, 70, 74, 80, 81, 99]. In this systems-level, data-driven analysis of GWAS from several large and high-quality cohorts of diverse ethnicities, integrated with functional data (from ENCODE, eQTLs, tissue-specific co-expression and regulatory networks constructed from human and mouse experiments), we identified both known and novel pathways and gene subnetworks that were genetically linked to both CVD and T2D across cohorts and ethnicities. Further, KDs in tissue-specific subnetworks appear to regulate many known disease genes for increased risk of CVD and T2D. Lastly, we experimentally validated the network topology using *in vitro* adipocyte and data from *in vivo* gene knockout models, and confirmed the role of the top KDs and subnetworks in both CVD and T2D traits in independent sets of mouse studies.

The data-driven nature of the current study offers several strengths. First, we incorporated the full-scale of genetic variant-disease association from multiple cohorts, ethnicities and disease endpoints, allowing for the detection of subtle to moderate signals, as well as comparison and replication of results across diseases and populations. More importantly, by focusing on results that demonstrate consistent significance at pathway and network level, we overcome the difficulties in harmonizing independent datasets that are complicated by substantial heterogeneity due to platform differences and population substructure. This is because disease signals across populations are more conserved at pathway level than at individual variant and

gene levels [73, 75, 77]. Second, the comprehensive incorporation of tissue-specific eQTLs, coupled with the use of tissue-specific networks, enhances our ability to achieve better functional mapping between genetic variants and genes, and uncover systems-level regulatory circuits for CVD and T2D in a tissue-specific fashion. Third, data-driven modules and networks used in this study increase the potential for novel discovery as gene-gene interactions are defined by data rather than prior knowledge. As the network models were from many independent studies reflecting diverse physiological conditions, leveraging these datasets and network models offers more comprehensive coverage of biological interactions than any given dataset can provide and has proven a valuable approach to unveil novel biological insights [11, 18, 74]. While some of our findings confirmed those from previous canonical pathway-based analysis on disease processes including ECM-receptor interaction and cell-adhesion, and KDs such as *SPARC* [70], our data-driven approach in the current study uncovered numerous novel genes, pathways, and gene subnetworks. A likely reason for the enhanced discovery potential of co-expression modules is that several interacting pathways could be co-regulated in a single module, or a pathway could interact with other poorly annotated processes in a module to together confer disease risk. The use of modules capturing such interactions improves the statistical power, in contrast to testing the pathways individually. Lastly, we conducted cross-validation studies in support of the functional roles of specific KDs and subnetworks in CVD and T2D using independent experimental models.

We acknowledge the following limitations in our study. First, our analyses were constrained by the coverage of functional datasets that are currently available, which causes uneven tissue coverage between data types and statistical bias towards more commonly profiled tissues such as

adipose and liver, making it difficult to achieve precise inference for all relevant tissues. Although we believe this does not necessarily undermine the validity of the main findings from our study, we acknowledge that we likely have missed relevant biology from tissues with fewer studies and smaller sample sizes. Further investigation is needed when additional relevant datasets become available. Secondly, our FDR estimates in MSEA do not take into consideration the gene overlap structure among co-expression modules, due to the challenge in accurately adjusting for the various degrees of overlaps between module pairs. To alleviate this limitation, we focus on modules and pathways demonstrating consistency across datasets and merge overlapping modules subsequently. Thirdly, although we conducted validation experiments on the *CAV1* subnetwork in both *in vitro* and *in vivo* models and cross-validated the importance of the predicted top key drivers and subnetworks in two independent large-scale mouse population studies, further experiments are warranted to thoroughly test the causality of the predicted KDs and elucidate the detailed tissue-specific mechanisms of the KDs on CVD and T2D. This is particularly important considering the limited overlaps in the modules and KDs identified from our study and the ones identified in two recent multi-tissue network analysis of cardiometabolic diseases [71, 72]. Only 7 KDs overlapped including *APOA1, CD2, CEBPD, CENPF, CSF1R, CTSS, UBE2S*. Methodological differences in network inference and key driver analysis and differences in the pathophysiological conditions of the study populations could explain the discrepancies. Lastly, ethnic-specific and sex-specific mechanisms await future exploration.

There are several direct implications that can be drawn from the results of our integrative analyses of both observational and experimental data. First, it appears that pathogenic pathways for CVD and T2D are indeed common in ethnically diverse populations. These shared pathways

capture most of the critical processes that have been previously implicated in the development of either T2D or CVD, including metabolism of lipids and lipoproteins, glucose, fatty acids, bile acids metabolism, biological oxidation, coagulation, immune response, cytokine signaling, and PDGF signaling. Second, BCAA metabolism and ECM are among the top and common pathways identified. Our finding on BCAA is consistent with recent work relating serum levels of BCAA to risk of CVD and T2D in large prospective cohorts [100, 101], although whether BCAA is a "pathophenotype" or strong pathogenic factor has been debated [88, 102]. Our findings support a causal role of BCAA because 1) both CVD and T2D risk variants were enriched in the co-expression modules related to BCAA degradation, and 2) 15 genes in the BCAA pathway were part of the top KD subnetworks, representing a significant enrichment of BCAA genes (fold enrichment = 3.02, Fisher's exact test p = 1.4e-5). Of note, BCAA genes themselves carry few genetic risk variants for CVD and T2D, albeit their network neighboring genes are highly enriched for disease variants, which may result from negative evolutionary pressure due to the critical role of BCAA. More recently, Jang and colleagues have shown BCAA catabolism can cause insulin resistance, providing further support for the causal role of BCAA for both CVD and T2D [103]. Our finding on the role of ECM in both CVD and T2D is also in line with recent reports [70, 74, 89, 90, 104]. In the top enriched subnetworks, ECM genes appear to exert strong effect (**Figure 3.9B**) coordinating other processes such as cholesterol metabolism, energy homeostasis, and immune response across a wide range of peripheral tissues and endocrine axis. This substantiates the importance of ECM modeling as a mechanistic driver for CVD and T2D.

Secondly, our comprehensive network modeling identified critical disease modulators and key targets whose functional roles were subsequently supported by multiple cross-validation efforts. This supports the use of network modeling to unravel and prioritize promising top targets that may have high pathogenic potential for both CVD and T2D. The KDs we identified can be considered as "highly confident" for the following reasons: 1) they are KDs for both CVD and T2D associated modules, 2) the tissue-specific subnetworks of these KDs show significant and replicable association with both diseases, 3) their subnetworks are highly enriched with known CVD and T2D genes, 4) *in vitro* siRNA knockdown and *in vivo* knockout mouse experiments confirm the role of a central KD *CAV1* in regulating the downstream genes as predicted in our network model, and 5) both the expression levels of KDs and the genetic variants mapped to the KD subnetworks are significantly associated with CVD and T2D relevant traits in independent mouse populations with naturally occurring genetic variations.

Thirdly, most KDs are not GWAS signals reaching genome-wide significance, nor are they rare-variant carrying genes, indicating that standard genetic studies miss important genes that orchestrate known CVD and T2D genes. The phenomenon may reflect a negative evolutionary pressure experienced by the KDs due to their crucial functions. In support of this hypothesis, we found a significant enrichment of human essential genes lacking functional variations among the 162 KDs identified in our study [105] (Fold = 1.41, p = 9.02e-3). This is consistent with previous findings [18, 70, 74] reaffirming the power and reliability of our approach in uncovering hidden biological insights particularly in a systematic integrative manner.

The connections between KDs and other disease genes revealed by our study warrant future investigation into the potential gene-gene interactions. Indeed, a closer examination of the biological functions from the top shared KDs further corroborates their disease relevance. For instance, our network modeling identified *HMGCR* as a top KD, consistent with its primary role as the target for cholesterol-lowering HMG-CoA inhibitors, namely statins. Our directionality inference analysis indicates that *HMGCR* is associated with CVD and T2D in opposite directions. This is consistent with the recent findings that genetic variations in *HMGCR* that decrease CVD risk cause slightly increased T2D risk, and statin drugs targeting *HMGCR* reduces CVD risk but increases T2D risk [106-108]. *CAV1* and *IGF1* represent two tightly connected multi-functional KDs. *CAV1* null mice were found to have abnormal lipid levels, hyperglycemia, insulin resistance and atherosclerosis [92, 93]. Consistent with these observations, we found strong association of *CAV1* expression levels as well as *CAV1* network with diverse cardiometabolic traits in both human studies and mouse HMDP panels. Our data-driven approach also revealed the central role of *CAV1* in adipose tissue by elucidating its connection to a large number of CVD and T2D GWAS genes and to genes involved in focal adhesion and inflammation (**Figure 3.9**), which could drive adipocyte insulin resistance [109, 110]. The regulatory effect of *CAV1* on neighboring genes was subsequently validated using *in vitro* adipocyte and *in vivo* mouse models. Moreover, our network modeling also captured the central role of *CAV1* in muscle and artery tissues, suggesting multi-tissue functions of *CAV1* in the pathogenic crossroads for CVD and T2D. The other multi-functional KD, *IGF1*, is itself a GWAS hit for fasting insulin and HOMA-IR. Despite being primarily secreted in liver, in our study *IGF1* demonstrated an adrenal gland and muscle specific regulatory circuit with CVD and

T2D genes, suggesting that it may confer risk to these diseases through the adrenal endocrine function and muscle insulin sensitivity. The three ECM KDs we identified, *SPARC*, *PCOLCE* and *COL6A2*, were especially interesting due to their consistent and strong impact on diverse cardiometabolic traits shown in our cross-validation analyses in HMDP (**Figure 3.9, Figure 3.13**). *SPARC* encodes osteonectin, which is primarily circulated by adipocytes. It inhibits adipogenesis and promotes adipose tissue fibrosis[50]. *SPARC* is also associated with insulin resistance and coronary artery lesions[51, 52]. *PCOLCE* (procollagen C-endopeptidase enhancer) represents a novel regulator for hypothalamus ECM that could potentially disrupt the neuroendocrine system. *COL6A2*, on the other hand, provides new insights into how collagen may affect cardiometabolic disorders: in adrenal tissue *COL6A2* is connected to *IGF1R*, the direct downstream effector for KD *IGF1*. Importantly, our directionality analysis suggests that while some KDs such as *CAV1* may have similar directional effects on CVD and T2D, cases like *HMGCR* that show opposite effects on these diseases are also present. Therefore, it is important to test the directional predictions to prioritize targets that have the potential to ameliorate both diseases and deprioritize targets with opposite effects on the two diseases.

3.4 Conclusions

In summary, through integration and modeling of a multitude of genetics and genomics datasets, we identified key molecular drivers, pathways, and gene subnetworks that are shared in the pathogenesis of CVD and T2D. Our findings offer a systems-level understanding of these highly clustered diseases and provide guidance on further basic mechanistic work and intervention

studies. The shared key drivers and networks identified may serve as more effective therapeutic targets to help achieve systems-wide alleviation of pathogenic stress for cardiometabolic diseases, due to their central and systemic role in regulating scores of disease genes. Such network-based approach represents a new avenue for therapeutic intervention targeting common complex diseases.

3.5 Methods

Identification of qualified SNPs from GWAS of CVD and T2D

Detailed GWAS information including sample size, ethnicity and genotyping platform was described in **Table 3.3** and **Appendix**. Briefly, p-values of qualified single nucleotide polymorphisms (SNPs) at minor allele frequency > 0.05 and imputation quality > 0.3 for CVD and T2D were collected for all available GWAS datasets (WHI-SHARe, WHI-GARNET, JHS, FHS, CARDIoGRAMplusC4D [80], and DIAGRAM [81]). SNPs meeting the following criteria were further filtered out: 1) ranked in the bottom 50% (weaker association) based on disease association p-values and 2) in strong linkage disequilibrium (LD) (r2 > 0.5) according to ethnicity-specific LD data from Hapmap V3 [111] and 1000 Genomes[112]. For each GWAS dataset, LD filtering was conducted by first ranking SNPs based on the association p values and then checking if the next highest ranked SNP was in LD with the top SNP. If the r^2 was above 0.5, the SNP with lower rank was removed. The step was repeated by always checking if the next SNP was in LD with any of the already accepted ones.

Curation of Data-driven Gene Co-expression Network Modules

Using the Weighted Gene Co-expression Network Analysis (WGCNA)[113], we constructed gene co-expression modules capturing significant co-regulation patterns and functional relatedness among groups of genes in multiple CVD- or T2D-related tissues (including aortic endothelial cells, adipose, blood, liver, heart, islet, kidney, muscle and brain) obtained from nine human and mouse studies (**Table S3.4**). Modules with size smaller than 10 genes were excluded

to avoid statistical artifacts, yielding 2,672 co-expression modules. These coexpression modules were used as a collection of data-driven sets of functionally connected genes for downstream analysis. The potential biological functions of each module were annotated using pathway databases Reactome and KEGG, and statistical significance was determined by Fisher's exact test with Bonferroni-corrected $p < 0.05$.

Curation of Functional Genomics from eQTLs and ENCODE

eQTLs establish biologically meaningful connections between genetic variants and gene expression, and could serve as functional evidence in support of the potential causal role of candidate genes in pathogenic processes[15, 31]. We therefore conducted comprehensive curation for significant eQTLs in a total of 19 tissues that have been identified by various consortia (including the Genotype-Tissue Expression (GTEx) [27], Muther [114] and Cardiogenics [115], and additional independent studies; **Table S3.5)**. Additional functional genomics resources from ENCODE were also curated to complement the eQTLs for SNP-gene mapping (See **Appendix**).

Identification of Genetically-driven CVD and T2D Modules using Marker Set Enrichment Analysis (MSEA)

MSEA was used to identify co-expression modules with over-representation of CVD- or T2D-associated genetic signals for each disease GWAS in each cohort/ethnicity in a study specific manner. MSEA takes into three input: 1) Summary-level results of individual GWAS (WHI, FHS, JHS, CARDIoGRAM+C4D, DIAGRAM) for the LD-filtered SNPs; 2) SNP-gene mapping

information, which could be determined by tissue-specific cis-eQTLs, ENCODE based functional annotation and chromosome distance based annotation. Cis-eQTLs is defined as eQTLs within 1MB of the transcription starting sites of genes. For ENCODE, we accessed the Regulome database and used the reported functional interactions to map SNPs to genes by chromosomal distance. Only SNPs within 50kb of the gene region and have functional evidence in Regulome database were kept; 3) Data-driven co-expression modules from multiple human and mouse studies as described above. Tissue-specificity was determined by the tissue origins of eQTLs and ethnic specificity was determined by the ethnicity of each GWAS cohort, since the human disease genetic signals and human eQTL mapping were the main driving factors to determine the significance of the modules. MSEA employs a chi-square like statistic with multiple quantile thresholds to assess whether a co-expression module shows enrichment of functional disease SNPs compared to random chance [75]. The varying quantile thresholds allows the statistic to be adoptable to studies of varying sample size and statistical power. For the list of SNPs mapped to each gene-set, MSEA tested whether the SNP list exhibited significant enrichment of SNPs with stronger association with disease using a chi-square like statistic: $\chi = \sum_{i=1}^{n} \frac{O_i - E_i}{\sqrt{E_i} + \kappa}$, where n denotes the number of quantile points (we used ten quantile points ranging from the top 50% to the top 99.9% based on the rank of GWAS p values), O and E denote the observed and expected counts of positive findings (i.e. signals above the quantile point), and κ = 1 is a stability parameter to reduce artefacts from low expected counts for small SNP sets. The null background was estimated by permuting gene labels to generate random gene sets matching the gene number of each co-expression module, while preserving the assignment of SNPs to genes, accounting for confounding factors such as gene size, LD block size and SNPs per loci.

For each co-expression module, 10000 permuted gene sets were generated and enrichment P-values were determined from a Gaussian distribution approximated using the enrichment statistics from the 10000 permutations and the statistics of the co-expression module. Finally, Benjami-Hochberg FDR was estimated across all modules tested for each GWAS.

To evaluate a module across multiple GWAS studies, we employed the Meta-MSEA analysis in Mergeomics, which conducts module-level meta-analysis to retrieve robust signals across studies. Meta-MSEA takes advantage of the parametric estimation of p-values in MSEA by applying Stouffer's Z score method to determine the meta-Z score, then converts it back to a meta P-value. The meta-FDR was calculated using Benjamini-Hochberg method. Co-expression modules with meta-FDR < 5% were considered significant and included in subsequent analyses.

Identification of Key Drivers and Disease Subnetworks

We used graphical gene-gene interaction networks including the GIANT networks [41] and Bayesian networks (BN) from 25 CVD and T2D relevant tissue and cell types **(Table S3.6**, See **Appendix**) to identify KDs. If more than one dataset was available for a given tissue, a network was constructed for each dataset and all networks for the same tissue were combined as a union to represent the network of that tissue, with the consistency of each network edge across datasets coded as edge weight. The co-expression modules genetically associated with CVD or T2D identified by Meta-MSEA were mapped onto these graphical networks to identify KDs using the weighted key driver analysis (wKDA) implemented in Mergeomics [75]. wKDA uniquely consider the edge weight information, either in the form of edge consistency score in the case of BNs or edge confidence score in the case of GIANT networks. Specifically, a network was first

screened for suitable hub genes whose degree (number of genes connected to the hub) is in the top 25% of all network nodes. Once the hubs have been defined, their local one-edge neighborhoods, or "subnetworks" were extracted. All genes in each of the CVD and T2D-associated gene sets that were discovered by meta-MSEA were overlaid onto the hub subnetworks to see if a particular subnetwork was enriched for the genes in CVD/T2D associated gene sets. The edges that connect a hub to its neighbors are simplified into node strengths (strength = sum of adjacent edge weights) within the neighborhood, except for the hub itself. The test statistic for the wKDA is analogous to the one used for MSEA: $\chi = \frac{O-E}{\sqrt{E}-\kappa}$, except that the values O and E represent the observed and expected ratios of genes from CVD/T2D gene sets in a hub subnetwork. In particular, $E = \frac{N_k N_p}{N}$ is estimated based on the hub degree N_k, disease gene set size N_p and the order of the full network N, with the implicit assumption that the weight distribution is isotropic across the network. Statistical significance of the disease-enriched hubs, henceforth KDs, is estimated by permuting the gene labels in the network for 10000 times and estimating the P-value based on the null distribution. To control for multiple testing, stringent Bonferroni adjustment was used to focus on the top robust KDs. KDs shared by CVD and T2D modules are prioritized based on the following criteria: i) Bonferroni-corrected $p< 0.05$ in wKDA, ii) replicated by both GIANT networks and Bayesian networks, and iii) the genetic association strength between the KD subnetworks (immediate network neighbors of the KDs) and CVD/T2D in Meta-MSEA. Finally, Cytoscape 3.3.0 [116] was used for disease subnetwork visualization.

Inference of the Direction of Genetic Effects of KD subnetworks

We used the genetic effect direction of KDs as a proxy for probable effect direction of the KD subnetworks. For each KD, we retrieved their tissue-specific eQTLs as well as variants within 50kb of the gene region, whose genetic association information was available in both CARDIoGRAMplusC4D and DIAGRAM, the two large meta-consortia of GWAS for CVD and T2D. CVD/T2D association beta-values of mapped variants of KDs were then extracted, and the signs of beta-values were examined to ensure they were based on the same reference alleles between GWAS. Lastly, for all mapped variants on each KD, the signs of the beta-value for CVD and T2D were compared and statistical significance of the proportion of variants with similar or opposite effect direction between diseases was determined by z-test.

Validation of KD Subnetwork Topology Using siRNA Knockdown in Adipocytes

We chose to validate the predicted adipose subnetwork of a top ranked KD of both CVD and T2D, *Cav1*, in 3T3-L1 adipocytes. Cells were cultured to confluence and adipocyte differentiation was induced using MDI differentiation medium (See **Appendix**). Two days after the initiation of differentiation, cells were transfected with 50 nM *Cav1* siRNAs (3 distinct siRNAs were tested and two of the strongest ones were chosen) or a scrambled control siRNA. For each siRNA, two separate sets of transfection experiments were conducted, with three biological replicates in each experiment. Two days after transfection, cells were collected for total RNA extraction, reverse transcription and quantitative PCR measurement of 12 predicted *Cav1* subnetwork genes and 5 random genes not within the subnetwork as negative controls (See **Appendix**). β-actin was used to normalize the expression level of target genes.

Validation of KD Subnetwork Topology Using Cav1 null mice

We accessed the gonadal white tissue gene expression data of 3-month-old wild type and *Cav1*$^{-/-}$ male mice (N=3/group) from Gene Expression Omnibus (GEO accession: GSE35431). Detailed description of the data collection procedures has been described previously [92]. Gene expression was profiled using Illumina MouseWG-6 v2.0 expression beadchip and normalized using robust spline. Differentially expressed genes (DEGs) between genotype groups were identified using linear model implemented in the R package Limma and false discovery rate was estimated using the Benjamini-Hochberg procedure [58]. DEGs at different statistical cutoffs were compared to *CAV1* subnetwork genes at different levels (i.e., 1, 2, 3, or 4 edges away from *CAV1*) to assess overlap and significance of overlap was evaluated using Fisher's exact test.

Validation of KD Subnetworks Using Mouse HMDP Studies

To further validate the role of KD subnetworks in CVD and T2D, we incorporated genetic, genomic and transcriptomic data from HMDP (comprised of >100 mouse strains differing by genetic composition) [54, 82, 83]. HMDP data was from two panels, one with mice fed with a high-fat diet (HF-HMDP)[82], and the other with hyperlipidemic mice made by transgenic expression of human APOE-Leiden and CETP gene (ATH-HMDP)[83]. For HF-HMDP, we retrieved gene-trait correlation data for adipose tissue (due to its relevance to both CVD and T2D) and 7 core cardiometabolic traits including adiposity, fasting glucose level, fasting insulin level, LDL, HDL, triglycerides and homeostatic model assessment-insulin resistance (HOMA-IR). For ATH-HMDP, we retrieved aorta gene-trait correlation (aorta tissue is the main site for CVD in mice) for all 7 traits. In addition to assessing the trait association strengths of individual

KDs, we also used MSEA to evaluate the aggregate association strength of the top CVD/T2D KD subnetworks with the traits at both transcription and genetic levels through transcriptome-wide association (TWAS) and GWAS in HF-HMDP and ATH-HMDP (See **Appendix**).

3.6 Tables

Table 3.1 Summary of top co-expression modules associated with CVD or T2D (FDR < 1% in Meta-MSEA, in column FDR_{meta})

Disease	Module ID	Annotation	CAR+C4D/ DIAGRAM	JHS	FHS	WHI	WHI	WHI	P_{meta}	FDR_{meta}
			Mixed	AA	EA	EA	AA	HA		
CVD	4406	NA	**3.32E-10**	NS	-	2.83E-02	4.41E-03	NS	5.73E-09	<0.01%
	4522	Signaling by FGFR mutants	**1.03E-04**	1.62E-02	-	3.80E-02	5.53E-03	2.86E-02	3.39E-08	<0.01%
	4540	NA	**9.72E-04**	NS	-	NS	1.50E-02	5.52E-04	5.07E-07	0.06%
	5242	Cholesterol Biosynthesis	**4.19E-06**	4.71E-02	-	NS	2.31E-02	NS	2.64E-06	0.08%
	4087	Carboxylic acid metabolic process	**2.34E-06**	NS	-	NS	8.63E-03	2.17E-02	4.24E-06	0.09%
	4019	Transmembrane transport of small molecules	**1.89E-03**	4.46E-02	-	NS	NS	**6.85E-04**	7.91E-06	0.20%
	4941	Establishment of localization	**8.97E-06**	1.52E-02	-	NS	NS	3.94E-02	2.72E-06	0.21%
	5023	TCA cycle and respiratory electron transport	NS	**6.37E-05**	-	**1.53E-03**	NS	1.50E-02	1.15E-05	0.22%
	blue	Cell cycle	1.08E-02	NS	-	NS	NS	**1.77E-04**	3.85E-06	0.30%
	5329	Biological oxidations	NS	2.32E-02	-	5.01E-03	3.26E-02	2.26E-02	2.21E-05	0.35%

	124	NA	NS	**1.48E-03**	-	NS	**7.05E-07**	NS	4.86E-06	0.55%
	4656	Cellular protein complex assembly	NS	NS	-	NS	3.64E-03	**2.27E-04**	8.85E-06	0.67%
	4147	NA	**1.55E-02**	**2.06E-04**	-	NS	8.85E-03	NS	5.72E-06	0.68%
	4989	Metabolism of amino acids and derivatives	**1.86E-03**	7.41E-03	-	NS	**3.71E-04**	NS	7.81E-05	0.82%
	5323	NA	**8.68E-04**	NS	NS	**2.25E-04**	**1.05E-03**	NS	1.58E-07	0.02%
	5250	NA	**4.78E-05**	NS	NS	3.01E-02	**3.46E-07**	NS	4.32E-07	0.03%
	4880	NA	8.96E-03	NS	1.18E-02	**5.06E-04**	NS	NS	1.61E-06	0.06%
	6872	NA	NS	**1.26E-03**	7.44E-03	**7.79E-03**	NS	NS	1.26E-06	0.06%
	4879	NA	3.18E-02	NS	**5.88E-04**	NS	**2.66E-03**	**2.20E-03**	1.19E-06	0.14%
	6533	Cholesterol biosynthesis	NS	5.02E-03	NS	NS	NS	**1.26E-06**	1.06E-05	0.25%
T2D	6977	NA	3.66E-02	NS	**4.01E-05**	NS	1.81E-02	4.05E-02	1.71E-06	0.39%
	6675	Cholesterol biosynthesis	**3.72E-03**	3.35E-02	NS	NS	NS	**2.06E-05**	2.56E-05	0.52%
	37	NA	**1.94E-03**	5.53E-03	NS	NS	**9.38E-04**	NS	4.95E-06	0.57%
	4302	NA	**2.07E-03**	NS	NS	4.80E-03	**4.05E-06**	NS	9.89E-06	0.71%
	6690	Complement and coagulation cascades	1.93E-02	**1.01E-04**	NS	2.24E-02	NS	NS	1.36E-05	0.86%
	4059	SLC mediated transmembrane	NS	3.05E-02	5.80E-03	NS	**1.50E-02**	NS	1.29E-05	0.86%

		transport								
	4937	Amino acid metabolic process	9.21E-03	NS	5.88E-03	NS	**1.37E-03**	NS	2.11E-05	0.89%
	5059	TCA cycle and respiratory electron transport	**7.31E-04**	NS	2.74E-02	**8.66E-04**	NS	NS	6.64E-06	0.95%

*Module IDs were randomly assigned IDs to co-expression modules. The annotation refers to the top functional category enriched in the co-expression modules (Bonferroni-corrected p< 0.05 based on Fisher's exact test, number of direct overlapping genes > 5). Numbers in scientific format were p-values from MSEA or Meta-MSEA analysis, and those reaching FDR < 20% in individual cohort analysis via MSEA (not the FDRmeta in Meta-MSEA) are highlighted in bold. CAR+C4D: CARDIoGRAMplusC4D; Mixed: mixed ethnicities; JHS: Jackson Heart Study; FHS: Framingham Heart Study; WHI: Women's Health Initiative; AA: African Americans; HA: Hispanic Americans; EA: European Americans; Pmeta and FDRmeta: p and FDR values from Meta-MSEA analysis across cohorts.

Table 3.2 Summary of the 15 key drivers and their corresponding subnetworks shared by CVD and T2D

Key drivers	Gene name	Sub-net size	Tissues	FDR_{CVD}	FDR_{T2D}	No. of CVD module	No. of T2D module	Suggestive genetic effect direction (CVD/T2D)	Subnetwork function
ACAT2	Acetyl-CoA Acetyltransferase 2	192	Adp, Dg, Lv, Ms, T	5.32%	4.35%	6	7	uncertain	Cell cycle; Cholesterol biosynthesis
ACLY	ATP Citrate Lyase	129	Adp, Dg, Lv, Ms	6.17%	0.47%	5	6	consistent	Cholesterol biosynthesis; Steroid biosynthesis
CAV1	Caveolin 1	954	Adp, Adr, Art, Dg, Ms, T, Ve	0.20%	0.32%	7	4	consistent	Immune system; Focal adhesion
COL6A2	Collagen Type VI Alpha 2 Chain	294	Adp, Adr, Dg, Ms, T	4.45%	0.40%	2	1	consistent	Extracellular matrix
COX7A2	Cytochrome C Oxidase Subunit 7A2	152	Adp, Adr, Art, Bld, Dg, Lv, Ly	3.79%	1.85%	1	4	uncertain	Respiratory electron transport
DBI	Diazepam Binding Inhibitor	181	Adp, Art, Bld, Dg, Is, Lv, Ly, Ms	7.70%	6.75%	5	5	uncertain	Respiratory electron transport
HMGCR	3-Hydroxy-3-Methylglutaryl-	75	Art, Dg, Lv, Ms	9.09%	4.87%	1	5	opposite	Cholesterol biosynthesis; Steroid

	CoA Reductase								biosynthesis
IDI1	isopentenyl-diphosphate delta isomerase 1	89	Adp, Art, Dg, Is, Lv, Ms, T	8.95%	3.46%	3	4	opposite	Cholesterol biosynthesis; Steroid biosynthesis
IGF1	insulin like growth factor 1	993	Adr, Ms	5.37%	1.20%	7	2	consistent	Immune system; Focal adhesion
MCAM	melanoma cell adhesion molecule	183	Adp, Adr, Art, Ms, T	7.16%	5.22%	4	2	uncertain	Extracellular matrix
MEST	mesoderm specific transcript	132	Adp, Adr, Lv, Ms	3.36%	1.58%	4	2	uncertain	Fibroblast growth factor signaling
MSMO1	methylsterol monooxygenase 1	133	Adp, Art, Dg, Lv, Ms, T,	7.70%	0.63%	1	4	uncertain	Cholesterol biosynthesis; Steroid biosynthesis
PCOLCE	procollagen C-endopeptidase enhancer	307	Adp, Adr, Art, Hy, Lv, Ms	6.17%	0.03%	2	2	uncertain	Extracellular matrix
SPARC	secreted protein acidic and cysteine rich	482	Adp, Adr, Art, Dg, Lv, Ms, Ve	9.63%	8.18%	5	3	consistent	Extracellular matrix
ZFP36	ZFP36 ring finger protein	176	Adp, Adr, Art, Lv, Ly, Ms	8.45%	7.69%	3	3	uncertain	Hypoxia-inducible factors; CD40 signaling

*FDR values were based on Meta-MSEA analysis of the KD subnetworks for enrichment of CVD or T2D GWAS signals across cohorts. The subnetwork size indicates the number of neighboring genes directly connected to a KD when all the tissue-specific networks where the KD was found are combined. No. of module columns indicate the number of CVD or T2D–associated co-expression modules from which each KD was identified. Suggestive genetic effect direction was designated "consistent" or "opposite" if the proportion of variants having consistent or opposite effect direction in CVD or T2D was statistically significant in either eQTL mapping or chromosomal distance mapping. Otherwise, "uncertain" was called. Subnetwork function was annotated based on KEGG and Reactome databases. Adp – adipose tissue; Adr - adrenal gland; Art – artery.

Table 3.3 Summary information of genome-wide association studies

Study	WHI-GARNET	WHI-SHARe	WHI-SHARe	Jackson Heart Study	Framingham Heart Study	CARDIoGRAM plusC4D	DIAGRAM
Origin	Caucasian	African American	Hispanic American	African American	Caucasian	Mostly Caucasian	Caucasian and Asian
Imputation	Hapmap3	Hapmap3	Hapmap3	1000G	1000G	1000G	Hapmap3
CVD Cases/Controls	545/2130	483/5880	131/2893	240/1908	471/5443	60801/123504	-
T2D Cases/Controls	1022/2130	1381/5739	581/2681	683/847	287/4657	-	26488/83964
Platform	Illuminia	Affymetrix	Affymetrix	Affymetrix	Affymetrix	-	-
N of SNPs	868609	836045	836060	8041765	5668200	6104813	2915012

3.7 Figures

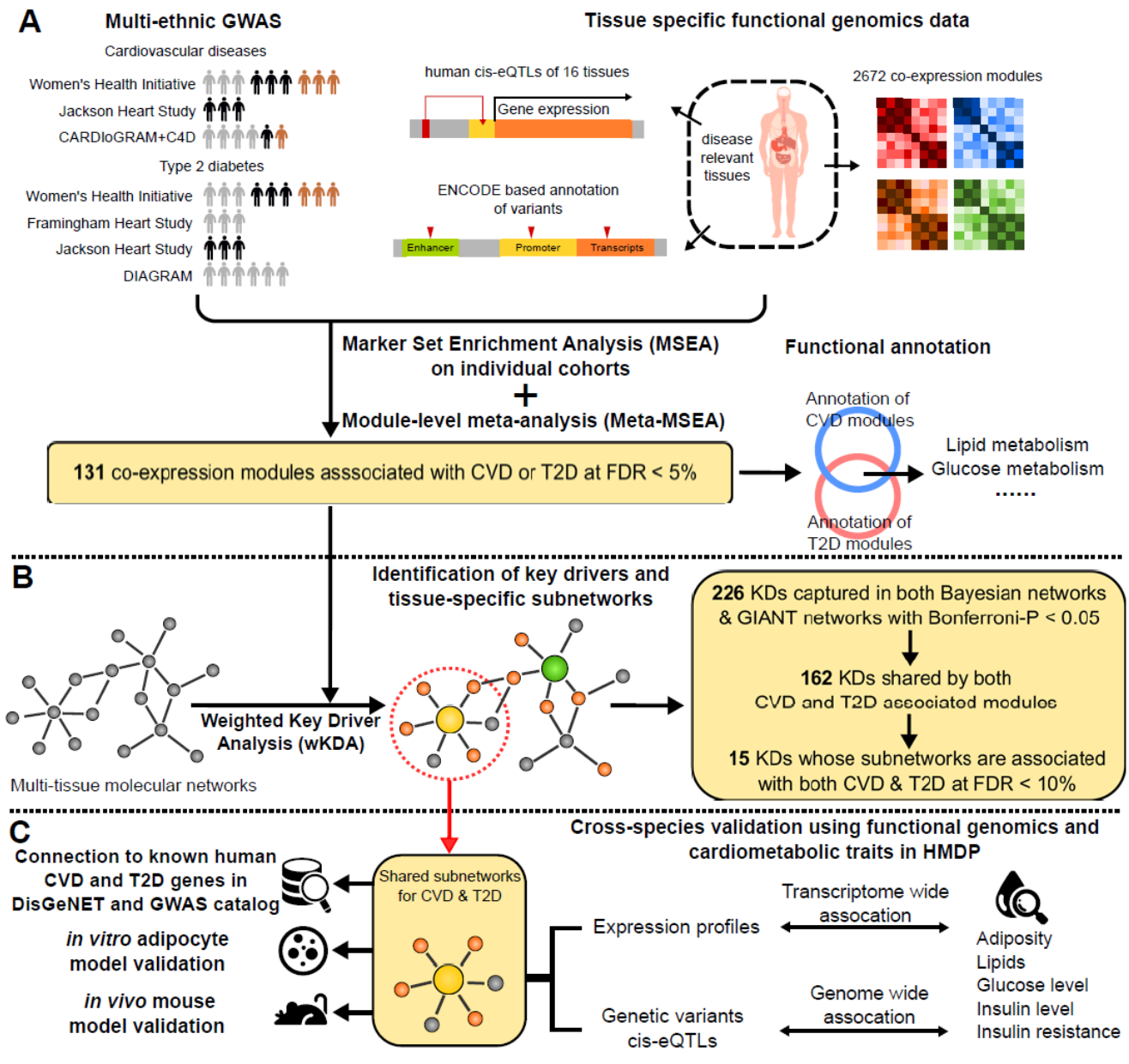


Figure 3.1 Framework of network-driven integrative genomics analyses.

A) Integration of genetics and functional genomics datasets to identify CVD and T2D associated co-expression modules. The GWAS studies for CVD and T2D were derived from three independent cohorts representing three ethnic populations: WHI (AA, EA, HA), FHS (EA), and

JHS (AA). These independent datasets were supplemented with GWAS of coronary artery disease from CARDIoGRAMplusC4D and T2D from DIAGRAM to increase power. We also curated a comprehensive list of tissue-specific functional genomics datasets, including 2672 co-expression modules, human eQTLs of various tissues, and ENCODE based variants annotation. The significant modules were identified by MSEA and Meta-MSEA, and then annotated to reveal shared pathways for CVD and T2D. In MSEA, the co-expression modules were used to define data-driven gene sets each containing functionally related genes, tissue-specificity was determined based on the tissue-origins of the human eQTLs, and ethnic specificity was determined based on the ethnicity of each GWAS cohort. **B**) Identification of disease key drivers and subnetworks. We utilized multi-tissue graphical networks to capture key drivers for disease associated co-expression modules using wKDA, then prioritized KDs based on consistency and disease relevance of the subnetworks. **C**) Validation of the top key drivers and their subnetworks via intersection with known human CVD and T2D genes from DisGeNET and GWAS catalog, in vitro adipocyte siRNA experiments, and cross-validation at both transcriptomic and genomic levels in the hybrid mouse diversity panels (HMDP).

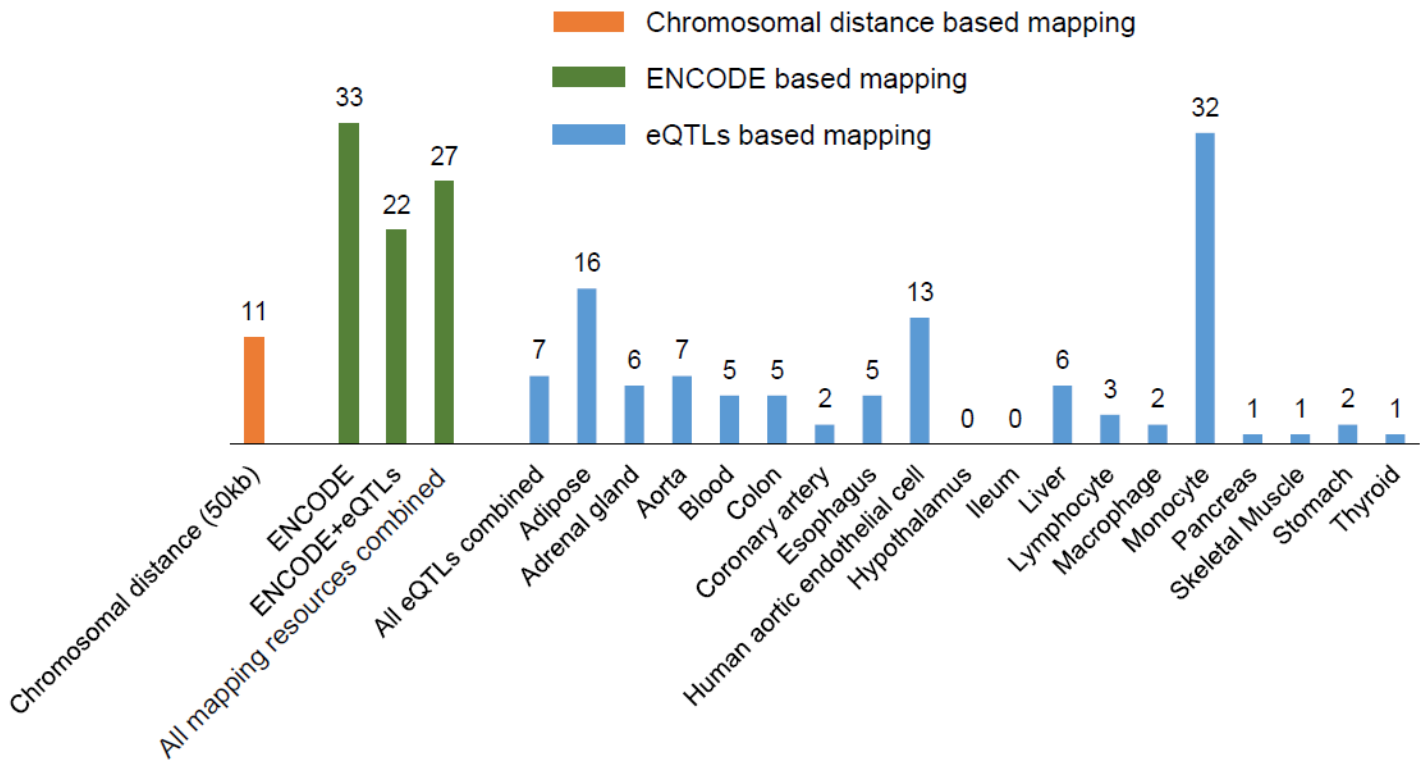


Figure 3.2 Number of significant co-expression modules found by different gene-SNP mapping types.

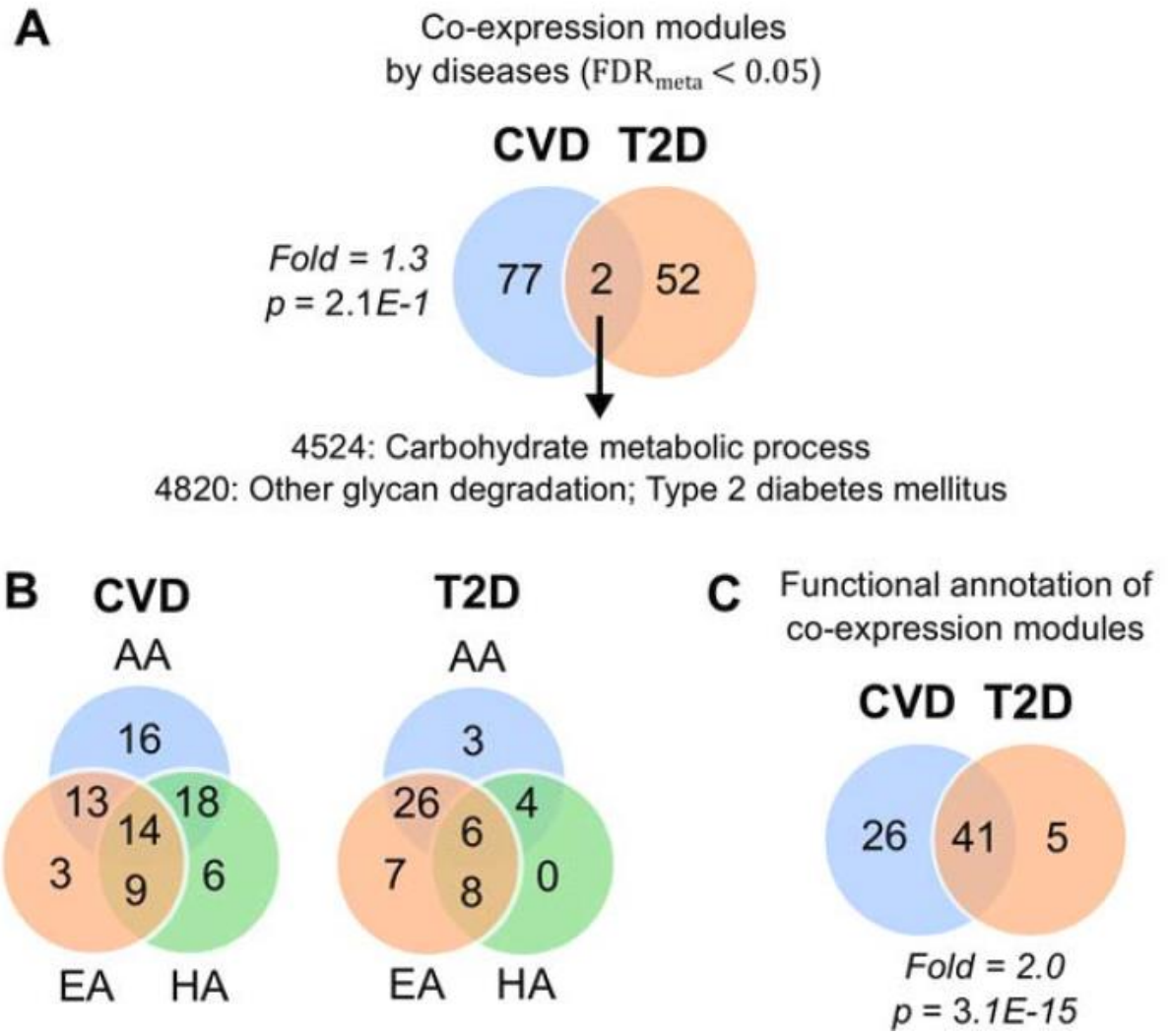


Figure 3.3 Venn Diagrams of overlap in significant co-expression modules and functional categories between diseases and ethnicities.

A) Count of module overlaps by disease based on Meta-MSEA; **B**) Count of module overlaps for each disease by ethnicity based on MSEA of individual studies. Co-expression modules captured in CARDIoGRAMplusC4D and DIAGRAM were not counted due to uncertain ethnic origin; **C**) Count of independent functional category overlaps by disease based on results from Meta-MSEA in panel A.

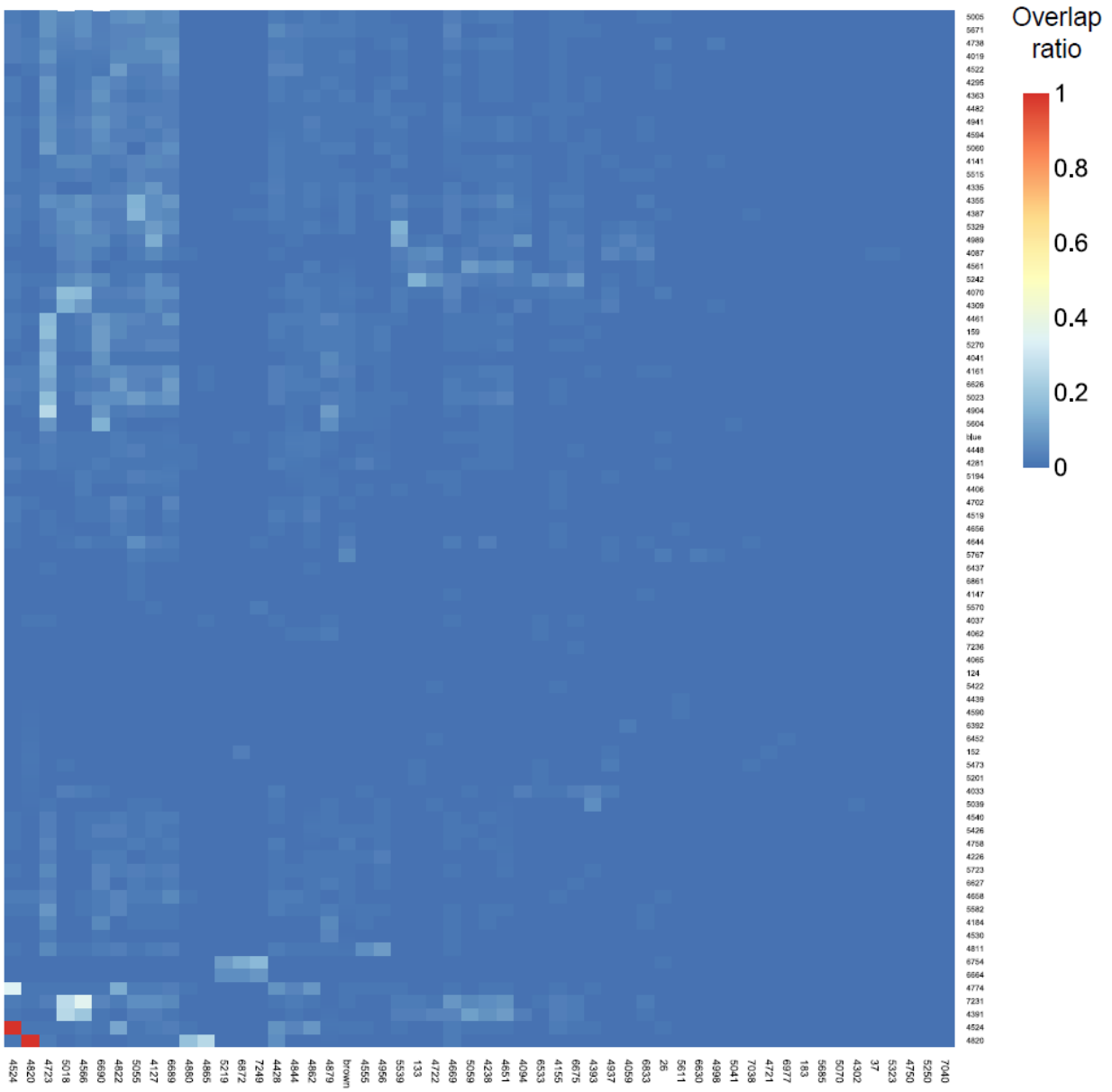


Figure 3.4 Heatmap of pair-wise overlapping ratio (Jaccard index) between the 79 co-expression modules associated with CVD (y-axis) and 54 modules (x-axis) associated with T2D.

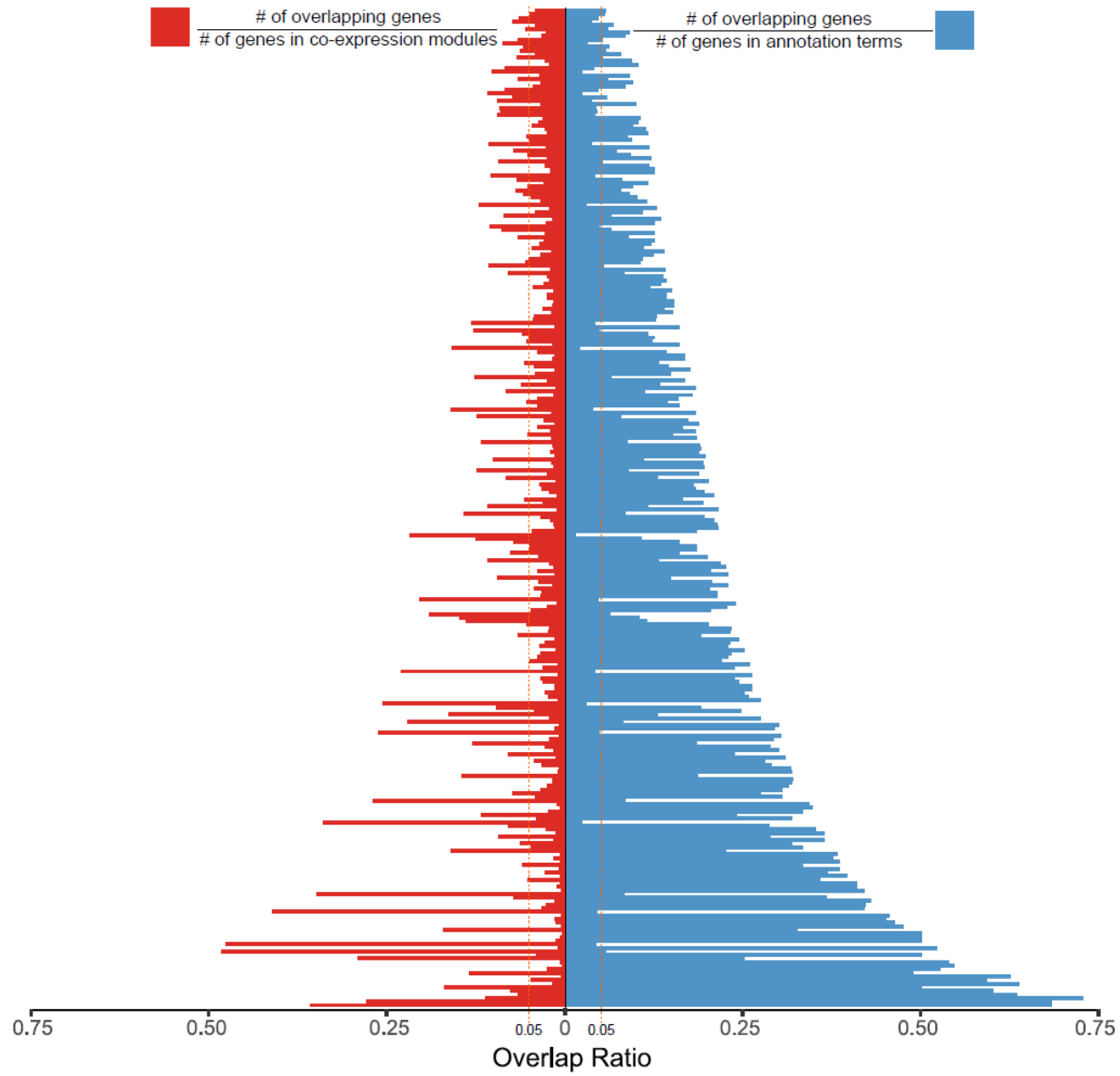


Figure 3.5 Overlap ratio plots between co-expression modules and the annotated functional terms.

All the annotated pathways reach >5% overlap ratio either on the pathway side or on the module side. Specifically, a majority (251 out of 278, 90.3%) of the annotated pathways had >=5% of genes overlapping with the modules to which they were assigned. For the 27 annotated pathways where <5% pathway genes were represented, these were the cases where the co-expression modules were small and the pathways were large, but all of them showed overlap of >5% module genes. The minimum, maximum, mean and median numbers of the overlapping genes for the annotations are 5, 170, 19 and 13, respectively.

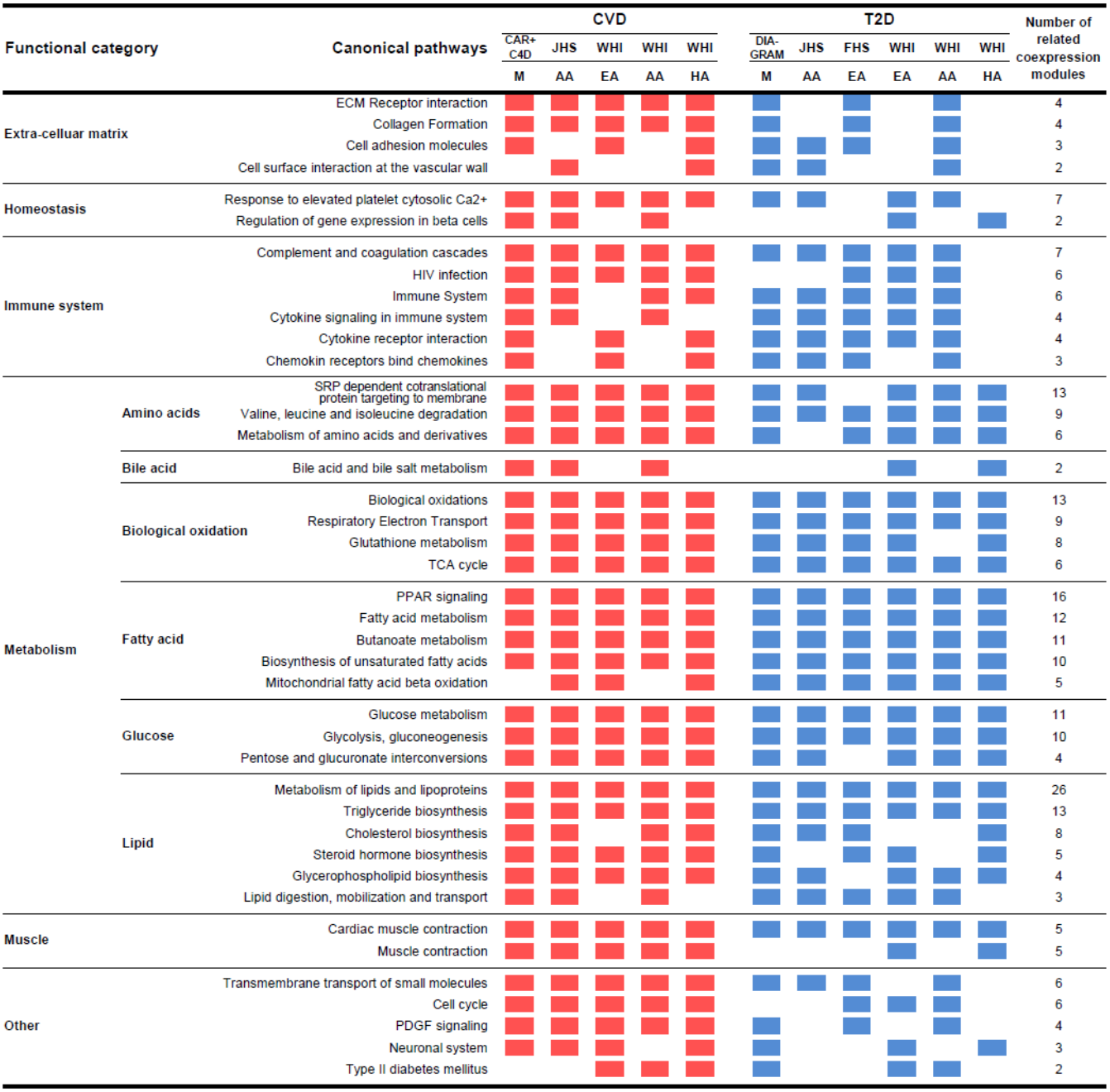


Figure 3.6 Summary of 41 independent functional categories enriched in both CVD and T2D co-expression modules (Bonferroni-corrected $p < 0.05$ based on Fisher's exact test, number of direct overlapping genes > 5).

Independent functional categories were defined as the categories with pair-wise overlapping ratio < 10%. Red and blue block indicates that the significant CVD or T2D co-expression modules identified from the study and ethnicity origin are enriched for the particular functional category

term. CAR+C4D: CARDIoGRAMplusC4D; M: mixed ethnicities; AA: African Americans; HA: Hispanic Americans; EA: European Americans.

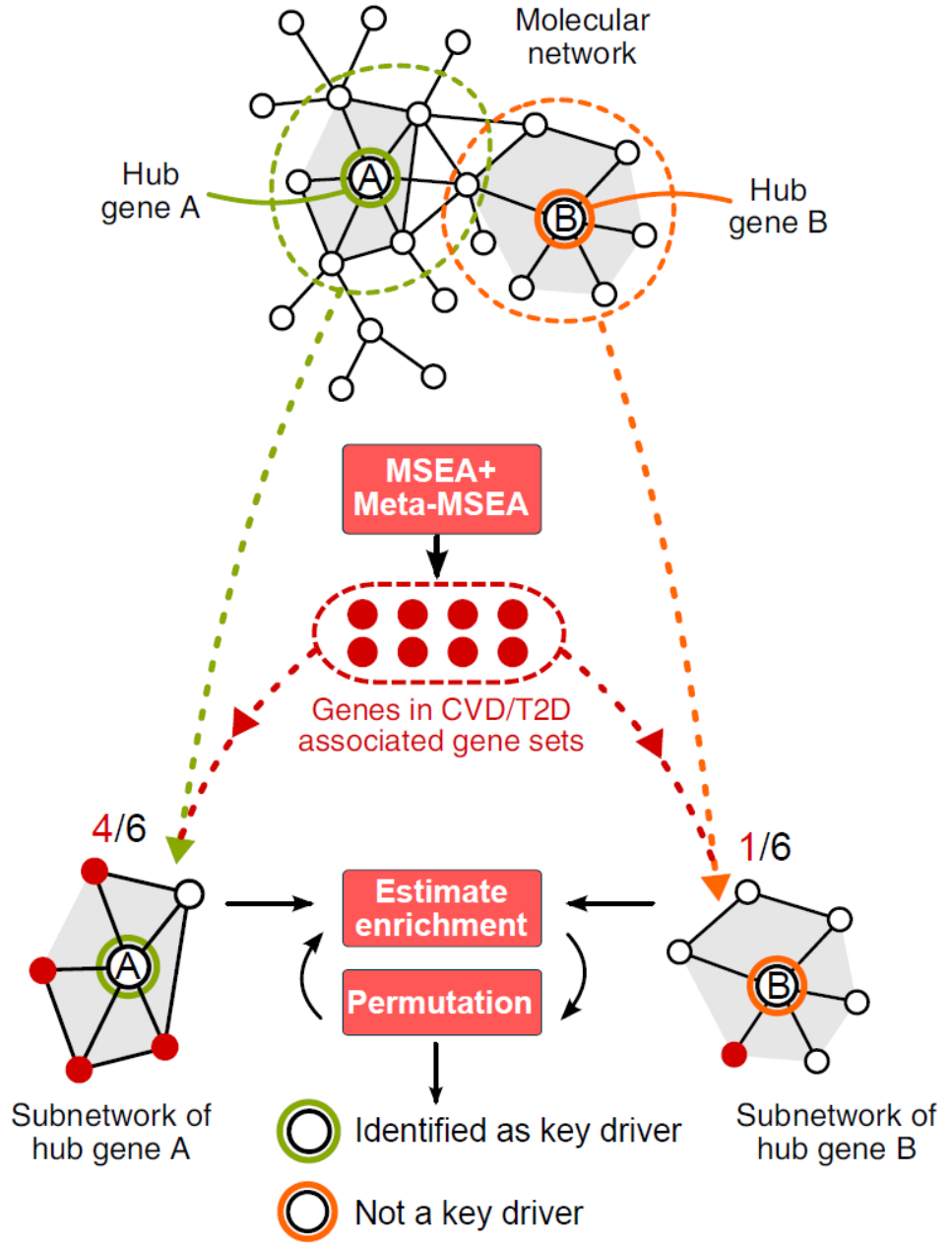


Figure 3.7 Concept of key driver analysis (KDA).

KDA requires gene regulatory networks capturing gene-gene interactions. Hub genes that show high degrees of connections to other networks genes are first identified, and their adjacent network neighbors (subnetworks) were extracted. All genes in each CVD/T2D associated module are used as input and mapped onto each hub subnetwork to assess whether a hub subnetwork was enriched for the genes in the input modules. The hubs whose subnetworks show significant enrichment of CVD/T2D module genes are defined as potential key drivers.

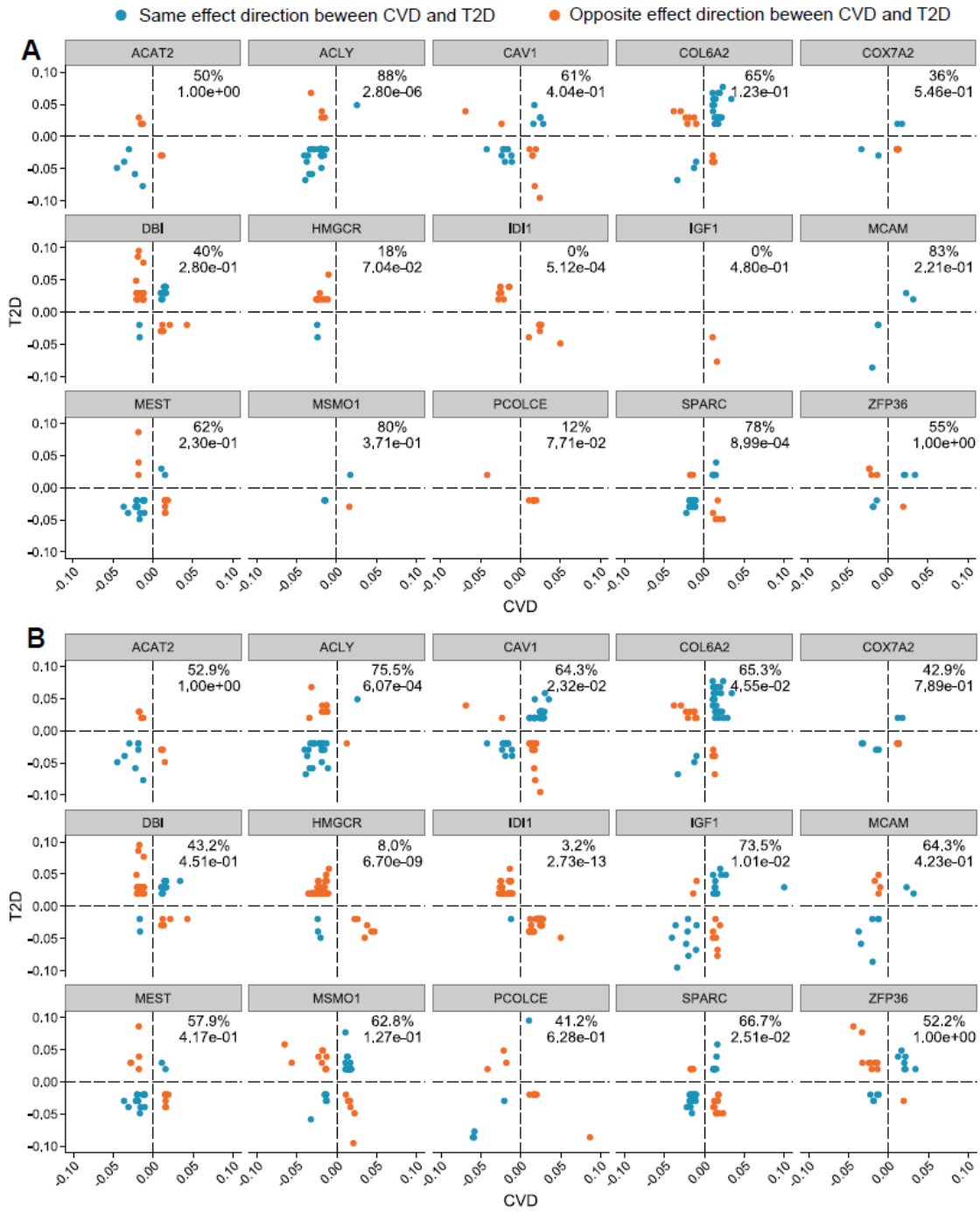


Figure 3.8 Scatter plots of the GWAS beta-values of variants mapped to the top 15 KDs.

A) Gene-variant mapping based on eQTLs only; **B**) Gene-variant mapping based on eQTLs and chromosomal distance. Percentage indicates the proportion of mapped variants with the same effect direction between CVD and T2D. Statistical significance of the difference of the proportion from random expectation is determined by z-test.

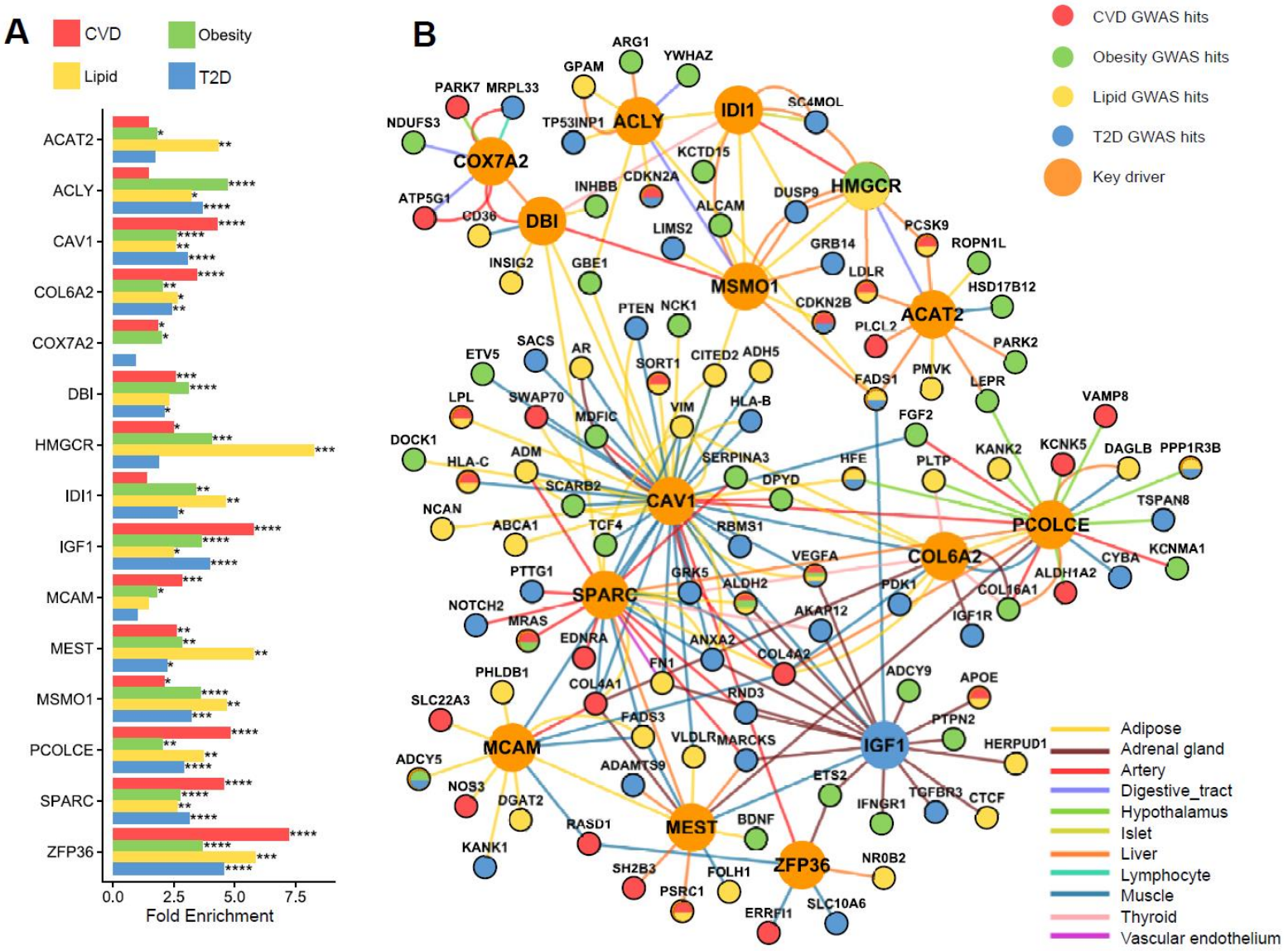

A
CVD
Obesity
Lipid
T2D
ACAT2
ACLY
CAV1
COL6A2
COX7A2
DBI
HMGCR
IDI1
IGF1
MCAM
MEST
MSMO1
PCOLCE
SPARC
ZFP36
0.0
2.5
5.0
7.5
Fold Enrichment
B
CVD GWAS hits
Obesity GWAS hits
Lipid GWAS hits
T2D GWAS hits
Key driver
Adipose
Adrenal gland
Artery
Digestive_tract
Hypothalamus
Islet
Liver
Lymphocyte
Muscle
Thyroid
Vascular endothelium

Figure 3.9 Subnetworks of the top 15 shared KDs orchestrate known genes for CVD, T2D, obesity and lipids.

A) Fold enrichment of KD subnetwork genes for known genes related to cardiometabolic traits reported in DisGeNET. *p<0.05, **p<0.01, ***p<0.001, ****p<0.0001. **B**) Top KD subnetworks with GWAS hits (p < 1e-5 as reported in GWAS Catalog) for cardiometabolic traits. KDs are large nodes. Edge color denotes tissue-origin. Only high-confidence edges (those with weight score in the top 20%) are visualized.

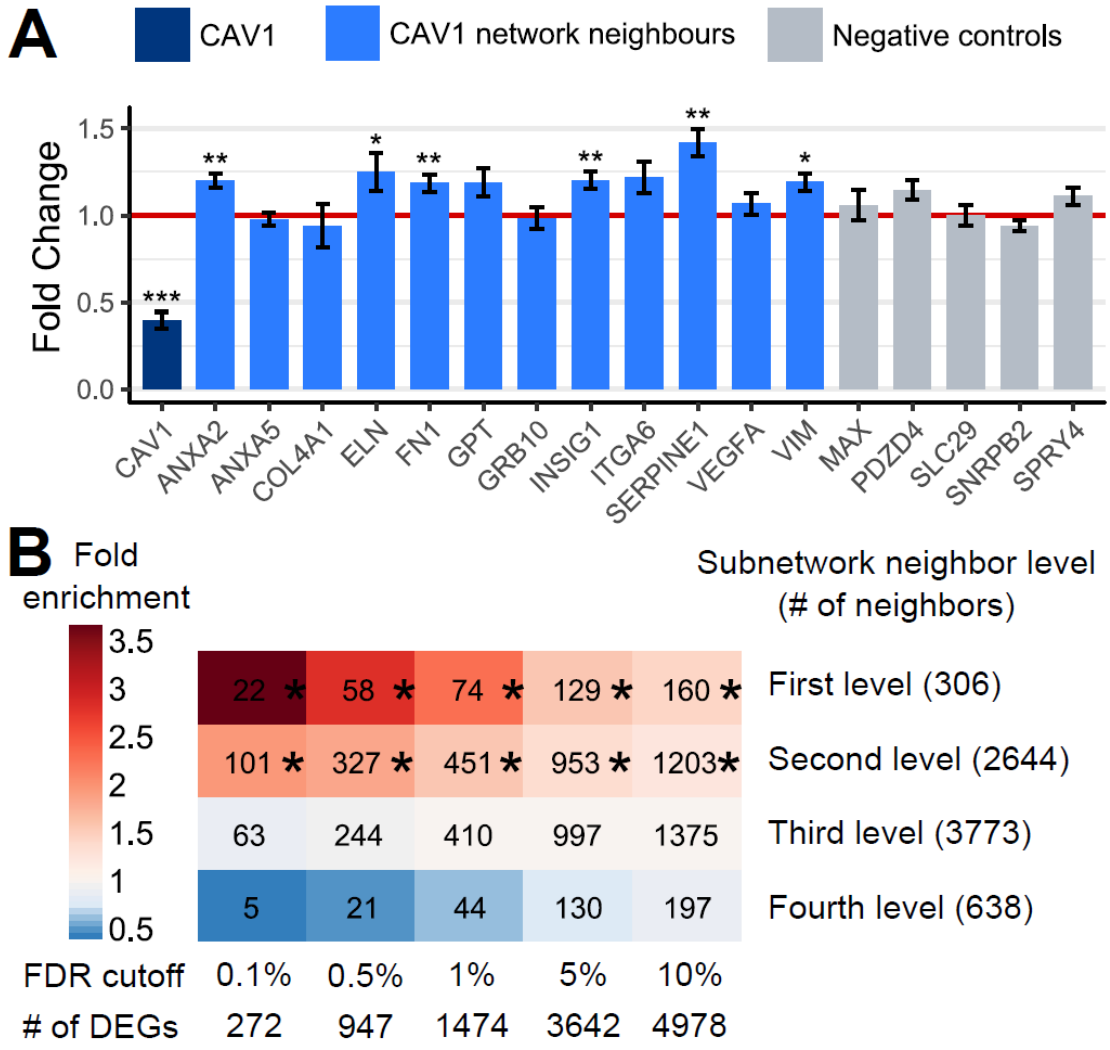


Figure 3.10 Validation of *CAV1* subnetwork using *in vitro* siRNA knockdown (**A**) and *in vivo* knockout mouse model (**B**).

A) Fold change of expression level for *CAV1* subnetwork and negative control genes 2 days after *Cav1* knockdown using two siRNAs separately. Twelve *CAV1* neighbors were randomly selected from the first and second level neighboring genes of *CAV1* in adipose network. Five negative controls were randomly selected from the genes not connected to *CAV1* or its first level networks in adipose network. Statistical significance of genes was determined by linear model, adjusting for batch effect and siRNA differences. N=12/group, mean ± SEM, *p < 0.05, **p < 0.01, ***p<0.001. **B**) Overlap of *CAV1* neighboring genes in the adipose tissue subnetwork at various distance levels with the differentially expressed genes in the gonadal adipose tissue in *Cav1* knockout mice (N=3/group). Overlap p-value is determined by Fisher's exact test. *Overlap p < 0.05 after Bonferroni correction.

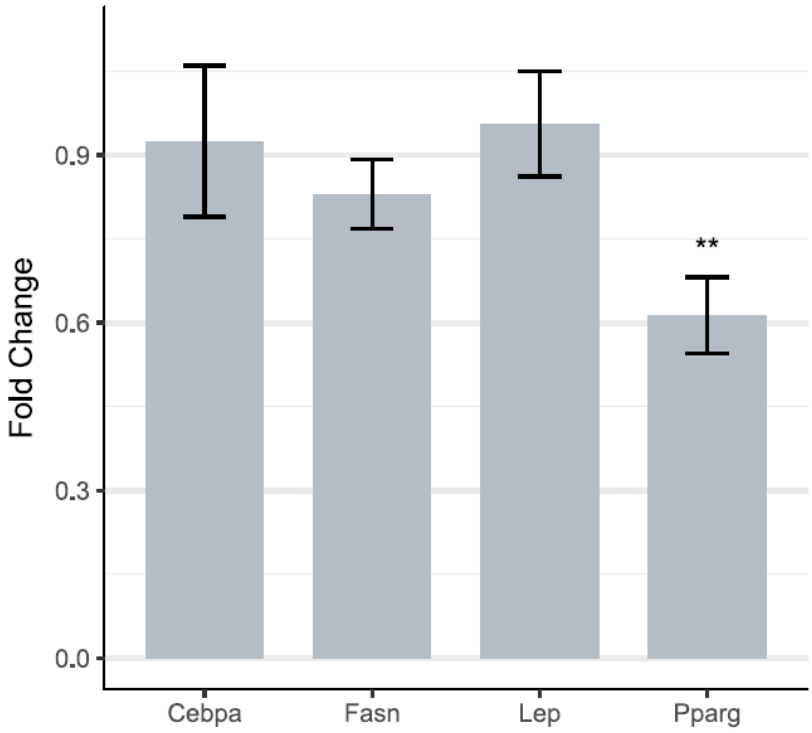


Figure 3.11 Expression changes in adipocyte differentiation state markers 3 days after the in vitro siRNA knockdown of Cav1.

Statistical significance of genes was determined by Student's t-test. N=3/group, mean ± SEM, **p < 0.01.

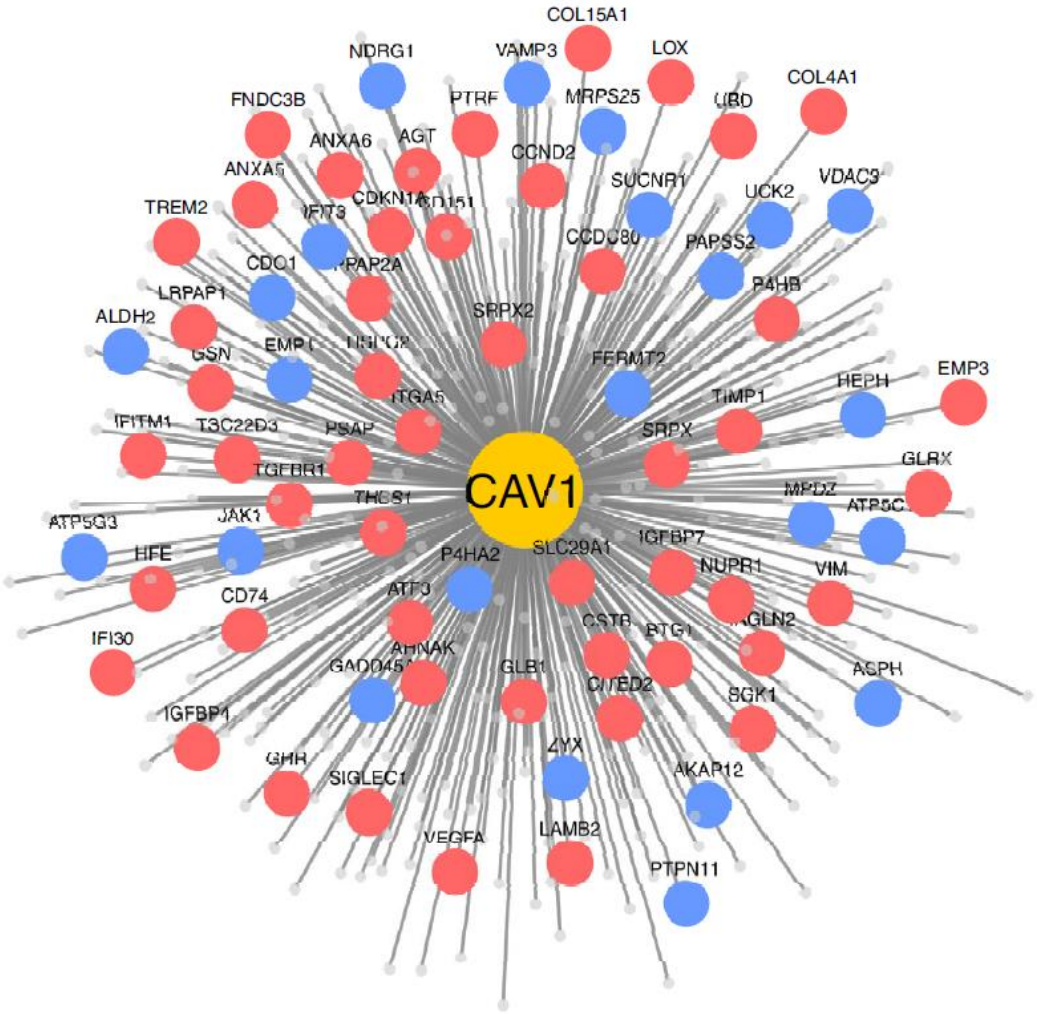


Figure 3.12 Visualization of CAV1 adipose subnetwork.

Red color indicates significantly up-regulated genes (FDR < 1%) in *Cav1*$^{-/-}$ mice, and blue color indicates significantly down-regulated genes (FDR < 1%) in *Cav1*$^{-/-}$ mice.

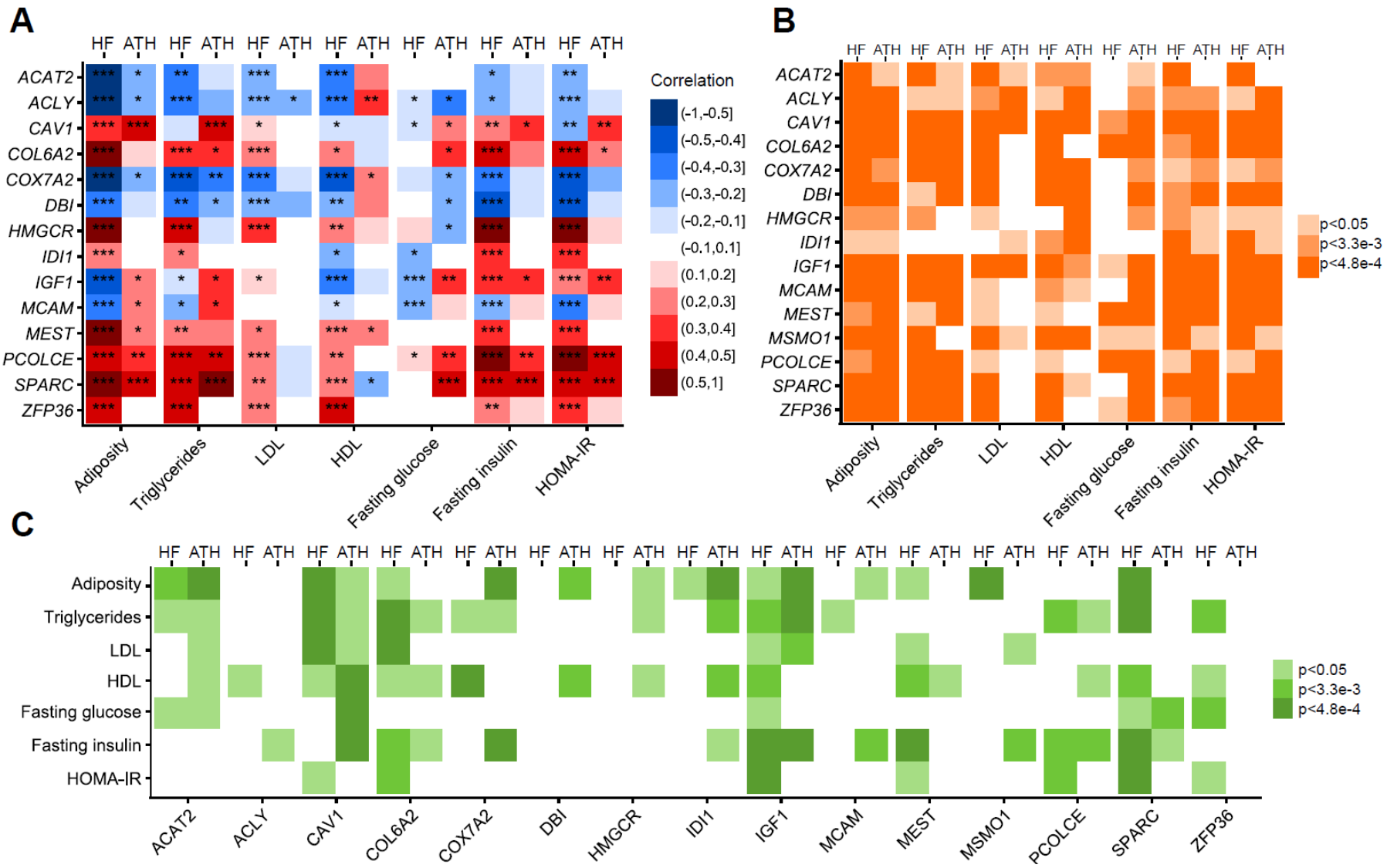

A
HF ATH
ACAT2
ACLY
CAV1
COL6A2
COX7A2
DBI
HMGCR
IDI1
IGF1
MCAM
MEST
PCOLCE
SPARC
ZFP36
Adiposity
Triglycerides
LDL
HDL
Fasting glucose
Fasting insulin
HOMA-IR
Correlation
(-1,-0.5]
(-0.5,-0.4]
(-0.4,-0.3]
(-0.3,-0.2]
(-0.2,-0.1]
(-0.1,0.1]
(0.1,0.2]
(0.2,0.3]
(0.3,0.4]
(0.4,0.5]
(0.5,1]
B
p<0.05
p<3.3e-3
p<4.8e-4
C
p<0.05
p<3.3e-3
p<4.8e-4

Figure 3.13 Associations of KDs and subnetworks with cardiometabolic traits in mice.

A) Association between KD expression and cardiometabolic traits in adipose tissue from HF-HMDP (HF) and aorta tissue from atherogenic-HMDP (ATH) as determined by Pearson correlation. *p< 0.05; **p< 0.05 after Bonferroni correction for the KD number; ***p< 0.05 after Bonferroni correction for the number of KDs and traits. **B**) Transcriptomic-wide association of KD subnetworks and cardiometabolic traits in adipose tissue from HF-HMDP, and aorta tissue from atherogenic-HMDP, as evaluated by MSEA. **C**) Genome-wide association of KD subnetworks and cardiometabolic traits based on adipose eQTL mapping in HF-HMDP, and aorta eQTL mapping in Atherogenic-HMDP, as determined by MSEA. p<0.05, p<3.3-3, and p<4.8e-4 correspond to uncorrected and Bonferroni-corrected p-values (correcting for the number of KDs or for the number of KD and trait combinations).

3.8 Appendix

Genome-wide Association Studies of CVD and T2D

A total of five GWAS were incorporated in the study, including WHI, JHS, FHS, CARDIoGRAMplusC4D and DIAGRAM. All studies were approved by institutional review committees and acquired informed consent from participants. WHI, JHS, and FHS are independent studies. Although FHS was included as a part of CARDIoGRAMplusC4D and DIAGRAM, the sample size of FHS (n=4461 for CVD, n=8338 for T2D) was only 2-7% of CARDIoGRAMplusC4D (n=184305) and DIAGRAM (n=110452). Thus, we consider all five GWAS as independent with each other. Individual summary level GWAS data used in the study was first pruned to only keep the SNPs in the top 50% range ranked by p-values to reduce noise from weak signals[75]. Remaining SNPs were further filtered for linkage disequilibrium using $r^2 > 0.5$ as the cutoff. Ethnicity-specific pairwise r^2 values of SNPs were retrieved from Hapmap v3 [111] and 1000 Genomes data [112].

Gene-SNP Mapping Resources

In additional to eQTL data, we included five additional gene-SNP mapping resources, 1) combining all tissue-specific eSNPs into a single multi-tissue eSNP set, 2) chromosomal distance mapping based on a 50kb window, 3) ENCODE project based Regulome SNPs[117], where only SNPs supported by at least two different kinds of RegulomeDB functional evidences are mapped to genes based on a 50kb window, 4) merging eSNP sets with Regulome data, 5) combining all

available mapping resources. As a result, a total of 21 sets of SNP-gene mapping were used for each co-expression module.

Mergeomics Package

"Mergeomics" is an open-source bioconductor package (https://bioconductor.org/packages/devel/bioc/html/Mergeomics.html) used to capture systems-level pathological perturbation from multiple layers of biological data such as genetics, genomics, transcriptomics, epigenomics, canonical pathways and gene networks.

Functional Annotation of Canonical Pathways

We collected 186 KEGG pathways and 674 Reactome pathways from MsigDB [118]. Functional categories were assigned to modules if Bonferroni-corrected enrichment reached $P < 0.05$ and there are more than five overlapping genes. For the over-represented functional categories, we considered those that had > 10% overlap ratio as redundant and only kept the categories with the highest number of associated disease modules. This was achieved by first ranking the annotations according to the number of linked modules, and then recursively removing the terms that have ≥ 10% overlapping ratio with the higher ranked annotation terms. The designation of trait, study and ethnicity to each of the functional categories was determined by the assigned modules.

Bayesian Networks of Gene-Gene Interactions

We utilized Bayesian network models of gene-gene interactions that take into account both the genotypes that affect gene expression (causal direction known), and the statistical relationships between gene expression levels (causal direction uncertain), using the established method by Zhu et al. [61, 62]. Bayesian network models from human and mouse studies were constructed based on genetics and gene expression data generated from multiple tissues from previously published studies, each involving hundreds of individuals (details and references in **Table S3.6**).

Bayesian networks are directed acyclic graphs in which the edges of the graph are defined by conditional probabilities that characterize the distribution of states of each gene given the state of its parents [119]. The network topology defines a partitioned joint probability distribution over all genes in a network. The likelihood of a Bayesian network model given observed genomic data is determined using Bayes formula. For each dataset, 1000 Bayesian networks, each using different random seeds, were reconstructed using Monte Carlo Markov Chain simulation [120]. Bayesian Information Criteria was used to determine the model with the best fit for each network. From the resulting set of 1000 networks, edges that appeared in greater than 30% of the networks were used to define a consensus network for a given dataset. To infer causal directions between genes in a network, genetic information was used as priors by allowing genes with *cis*-eSNPs to be parent nodes of genes without *cis*-eSNPs and preventing genes without *cis*-eSNPs to be parents of genes with *cis*-eSNPs [121]. Bayesian network provides a natural framework for integrating diverse data and reconstruct biological causal networks.

GIANT Networks of Gene-Gene Interactions

We downloaded the raw networks of relevant tissues from GIANT database [41]. The original connectivity of the GIANT networks was too high, which breaks the scale-free assumption of biological network topology and may lead to increased number of spurious results in subsequent key driver analysis. Thus we performed a filtering algorithm that step-wisely removed the networks edges with low confidence score and recursively evaluated the scale-freeness of the network. The removal process terminated as soon as the network reached boundary scale-freeness as defined by the criterion that at least 25% nodes has a degree (number of genes connected to the node] of 1.

Extraction of Disease/trait Associated Genes from DisGeNET and GWAS Catalog

Only genes with disease association score > 0.01 were extracted from DisGeNET [91] (accounting for ~ top 10% of genes reported in DisGeNET). The reported traits in DisGeNET were categorized into major groups as follows, 1) CVD: "Atherosclerosis", "Cardiovascular Diseases", "Acute myocardial infarction", "Myocardial Infarction", "Myocardial Ischemia", "Brain Ischemia", "Ischemia"; 2) T2D: "Insulin Resistance", "Impaired glucose tolerance", "Diabetes Mellitus, Non-Insulin-Dependent"; 3) Obesity: "Obesity", "Obesity, Abdominal", "Obesity, Visceral", "Obesity, Morbid", "Adult-onset obesity", "Moderate obesity", "Familial obesity", "Hypothalamic obesity", "Overweight and obesity", "Overweight", "Body Weight", "Moderate obesity", "Familial obesity", "Hypothalamic obesity", "Overweight and obesity", "Overweight", "Body Weight"; 4) Lipid: "Dyslipidemias", "Hyperlipidemia", "Hypercholesterolemia", "Hypercholesterolemia, Familial".

The reported traits in GWAS catalog [69] were categorized into major groups as follows, 1) CVD: "Coronary heart disease", "Coronary artery disease", "Coronary artery disease or ischemic stroke", "Coronary artery disease or large artery stroke", "Myocardial infarction", "Ischemic stroke", "Large artery stroke", "Stroke (ischemic)"; 2) T2D: Type 2 diabetes; 3) Obesity: "Body mass index", "Obesity", "Obesity (early onset extreme)", "Obesity (extreme)"; 4) Lipid: "HDL cholesterol", "Cholesterol, total", "Triglycerides", "LDL cholesterol". The latest results from CARDIoGRAM were not yet included in GWAS catalog and were manually added into the GWAS gene list for CVD.

The full list of CVD/T2D associated genes are shown in the **Table S3.3**.

Cell Culture of 3T3-L1 Cells and Differentiation

3T3-L1 cells were obtained from ATCC and cultured to confluence in high glucose and pyruvate DMEM containing 10% calf serum by changing every 2 days according to the manufacturer's instructions. Two days after confluence, the medium was changed to MDI differentiation medium (0.5 mM methylisobutylxanthine, 1.0 μM dexamethasone, 1.0 μg/ml bovine insulin in DMEM containing 10% fetal bovine serum). Two days later, the MDI was replaced with adipocyte maintenance medium containing 1.0 μg/ml bovine insulin and 10% fetal bovine serum in DMEM. The medium was replaced with fresh adipocyte maintenance medium every two to three days.

RNA Interference

Three siRNAs against *CAV1* were purchased from Sigma (siRNA ID: SASI_Mm01_00141141, St. Louis, MO) and two were chosen for further use based on the knockdown efficiency. A random sequence siRNA purchased from Sigma was used as negative control (siRNA ID: SIC001). Two days after induction with MDI, cells were transfected with 50 nM siRNA using Lipofectamine® RNAiMAX Reagent (ThermoFisher Scientific, Waltham, MA) according to the recommended manufacturer's protocol. For each siRNA, a total of six biological replicates (2 separate sets of transfection experiments, each with 3 replicates) were used and cells were collected two days after transfection to extract RNA for quantitative PCR analysis.

Reverse Transcription and Quantitative PCR

Total RNA was purified from 3T3-L1 cells using Direct-zol RNA MiniPrep Kit with DNase I (Zymo Research, Irvine, CA) according to the manufacturer's instructions. The first-strand cDNA were synthesized from 2 μg of total RNA with High-Capacity cDNA Reverse Transcription Kit (ThermoFisher Scientific, Waltham, MA). Amplification of each cDNA was performed with iTaq™ Universal SYBR® Green Supermix (Bio-Rad, Hercules, CA) and quantified by CFX96 qPCR System according to the protocol provided by manufacturer (Bio-Rad, Hercules, CA). The expression of the target genes was normalized to that of β-actin, which was selected as the housekeeping gene based on it stable expression during adipocyte differentiation. Linear regression was used to test for the statistical significance of each gene between control siRNA and *Cav1* siRNAs while adjusting for batch effect and siRNA differences.

Validation of KDs and Subnetworks using TWAS and GWAS Data from HMDP

To determine whether the subnetwork genes exhibited a global trend of having elevated association with cardiometabolic traits in HMDP, we performed enrichment test on the subnetworks of the top 15 KDs, using the MSEA module. In specific, the tissue and trait-specific TWAS data in HMDP was extracted. Genes were sorted by their association p-value and only genes with top 50% association were kept. The determination of enrichment statistics was the same as the one used for the SNPs, other than genes were used as markers directly without mapping SNPs to them.

We also took advantage of the HMDP GWAS and eQTL data in the aim of drawing causal relationship between the subnetworks and cardiometabolic traits. ~21000 mouse SNPs were reported in both the HF-HMDP and atherogenic-HMDP. The top 50% SNPs were retrieved and filtered by linkage disequilibrium threshold at 0.5. In addition, we collected eQTLs of adipose and aorta tissues from the two HMDP panels with $p < 1e\text{-}5$. eQTLs within the 1MB region of the transcription staring site of genes were categorized as “cis-eQTLs”, and were used to map SNPs to mouse genes. Lastly, MSEA was used to test for enrichment of association with cardiometabolic traits within the eQTLs mapped to the subnetwork genes.

Chapter 4. Prenatal Bisphenol A Exposure in Mice Induces Multi-tissue Multi-omics Disruptions Linking to Cardiometabolic Disorders

4.1 Introduction

A central concept in the Developmental Origin of Adult Health and Disease (DOHAD) states that adverse environmental exposure during early developmental stages is an important determinant for later onset adverse health outcomes, even in the absence of continuous exposure in adulthood [122-124]. BPA is one of the most influential environmental metabolic disruptors identified to date with widespread exposure in human populations and likely plays a role in DOHAD. BPA is used in the production of synthetic polymers, including epoxy resins and polycarbonates [125]. The advantageous mechanical properties of BPA have resulted in its ubiquitous use in everyday goods such as plastic bottles and inner coating of canned foods [126, 127]. BPA exposure has been confirmed in the majority of human populations [128] and has been linked to body weight, obesity, insulin resistance, diabetes, metabolic syndrome (MetS), and cardiovascular diseases in both human epidemiologic and animal studies [129-136]. Importantly, it has been suggested that the developing fetus is particularly vulnerable to BPA exposure [129, 137]. Intrauterine growth retardation (IUGR) has been consistently observed after developmental BPA exposure at intake doses below the suggested human safety level and has been associated with low birth weight, elevated adult fat weight and altered glucose homeostasis [129, 138-141]. As a result, BPA has been banned from baby products in Europe, Canada, and the US. However, BPA is still in use in non-baby products, posing continuous exposure to adults. Additionally, BPA has been associated with a transgenerational influence on obesity and MetS [142-144], contributing to a lingering effect of BPA exposure on future generations even under

usage restriction. Together these lines of evidence support an intriguing hypothesis that BPA may have been playing an important role in the rise of MetS and cardiometabolic diseases worldwide in the past decades [145-147].

Despite numerous findings connecting BPA with adverse health outcomes, there remain ample conflicting data, as summarized by the European Food Safety Agency [148] and the BPA Joint Emerging Science Working Group of the US FDA. Although inconsistencies across studies might be attributable to non-monotonic dose response, exposure window difference, and varying susceptibility between testing models [134, 149], there are also several additional layers of complexity and challenges hindering the full dissection of the biological effects of BPA. First, previous studies examining BPA in various cell types and tissues suggest a broad impact on biological systems [144, 150-152]. Second, BPA has been found to modulate multidimensional molecular events, such as gene expression and epigenetic changes, that are functionally important for processes such as metabolism and immune response [153-158]. However, due to most studies being designed to focus on one factor at a time as well as non-comparable study designs, it is difficult to directly compare effects across tissues or types of molecular data to derive the molecular rules of sensitivity to BPA exposures. In a recent NTP report, CLARITY-BPA, where multiple organs were examined, evidence of weight gain and cardiac dysfunctions were observed, however, the study was designed to be solely descriptive and no mechanism of action was proposed. These research gaps in our understanding of the pleiotropy of EDCs and toxicant biological actions necessitated the establishment of the NIEHS TaRGET consortium and a more recent call for the research community to systemically interrogate multiple omics in

multiple tissues to accelerate the discovery of key biological fingerprints of environmental exposure [159].

Here we present a multi-tissue, multi-omics systems biology study to examine the systems level influence of prenatal BPA exposure using advanced integrative genomics and network modeling approaches in a mouse model. We first utilized next-generation sequencing technologies to characterize perturbations in both the transcriptome and the epigenome across three tissues (gonadal white adipose tissue, hypothalamus, liver) in mouse offspring who had experienced *in utero* exposure to BPA. Based on mounting evidence that genes operate in highly complex tissue-specific regulatory networks, we hypothesized that prenatal BPA exposure induces genomic and epigenomic reprogramming in the offspring by affecting the organization and function of tissue-specific gene networks [18, 160-162]. Using both transcription factor (TF) networks and Bayesian networks, we modeled the dynamics of transcriptomic and epigenomic signatures and predicted potential regulators that govern the actions of BPA. Furthermore, the transcriptome, epigenome, and network information was layered upon metabolic phenotypes such as body weight, adiposity, circulating lipids, and glucose levels in the mouse offspring to evaluate disease association. Lastly, to assess the relevance of the BPA molecular targets identified in our mouse model for human diseases, we applied integrative genomics to bridge the mouse molecular signatures and genetic disease association data from human studies. Our study represents a comprehensive systems-level investigation of the molecular and health impact of BPA.

4.2 Results

Overall study design

As shown in **Figure 4.1A**, pregnant C57BL/6 mice were exposed to BPA during gestation via oral gavage at the dosage of 5mg/kg/day, situated below most reported no-observed-adverse-effect-level (NOAEL) according to toxicity testing (https://comptox.epa.gov/dashboard/dsstoxdb/results?search=Bisphenol+A). This dosage was typically used in previous studies [144, 163-165], and was chosen as a proof-of-concept for our systems biology study design and to facilitate comparison with previous studies. Male offspring (n = 9 for control, n = 11 for BPA) of weaning age (3-weeks) were examined for a spectrum of metabolic phenotypes (detailed below) and euthanized to collect key metabolic tissues including white adipose tissue, hypothalamus, and liver. We focused on examining males because we found stronger phenotypic changes in males at weaning age. We chose the weaning age in order to investigate early molecular and phenotypic changes in the offspring, which may predispose the offspring to late onset diseases. At the molecular level, we conducted RNA sequencing (RNA-seq) to evaluate transcriptomic alterations, and investigated perturbed biological pathways. We also used reduced representation bisulfite sequencing (RRBS) to uncover the epigenomic impact of prenatal BPA exposure at millions of methylation sites and analyzed the connection of the epigenomic alterations to changes in the transcriptome. We then integrated the transcriptomic and epigenomic signatures with two types of regulatory networks, namely, transcription factor networks to identify perturbed TF hotspots, and gene regulatory networks to identify non-TF regulatory genes. Finally, we interrogated the associations of the transcriptomic and epigenomic signatures of BPA obtained from our study with metabolic phenotypes in our

mouse offspring by correlative analysis, and with human diseases by querying top reported genes and full summary statistics from all publicly available genome-wide association studies (GWAS).

Prenatal BPA exposure induces tissue-specific transcriptomic alterations in male weaning offspring

We exposed pregnant C57BL/6 mice to BPA during gestation (day 1 to day 20 post-conception) at the dosage of 5mg/kg/day, and collected core tissues important for metabolism from male offspring at 3 weeks. Hypothalamus is the central regulator of endocrine and metabolic systems, whereas liver and white adipose tissues are critical for energy and metabolic homeostasis. To explore the molecular basis underlying the potential health impact of prenatal BPA exposure, we used RNA-seq to profile the transcriptome, and identified 86, 93, and 855 differentially expressed genes (DEGs) in the adipose tissue, hypothalamus, and liver tissue respectively, at FDR < 0.05 (**Figure 4.1B, Table S4.1**). This supports the ability of prenatal BPA exposure to induce large-scale transcriptomic disruptions in offspring, with the impact appearing to be more prominent in liver. Interestingly, the hypothalamic DEGs are predominantly up-regulated in the BPA group whereas the other two tissues did not show such direction bias (**Table 4.1**). The DEGs were highly tissue-specific, with only 12 out of the 86 adipose DEGs and 16 out of the 93 hypothalamus DEGs being found in liver. Only one gene, *Cyp51* (Sterol 14-Alpha Demethylase), was shared across all three tissues but with different directional changes (upregulated in hypothalamus and liver, downregulated in adipose) (**Figure 4.1C**). The Cyp51 protein catalyzes metabolic reactions including cholesterol and steroid biosynthesis and biological oxidation [166]. Previously, this gene was also found to be critical regulator for testicular spermatogenesis [167].

The consistent alteration of *Cyp51* across tissues suggests that this gene is a general target of BPA, with the potential to alter functions related to cholesterol, hormone, and energy metabolism.

Functional annotation of DEGs in adipose, hypothalamus, and liver tissues

To better understand the biological implications of the BPA exposure related DEGs in individual tissues, we evaluated the enrichment of DEGs for known biological pathways and functional categories (**Figure 4.1D-F,** full results in **Table S4.2**). We observed strong enrichment for pathways related to lipid metabolism (lipid transport, fatty acid metabolism, cholesterol biosynthesis) and energy metabolism (biological oxidation, TCA cycle) across all three tissues. Most of these pathways appeared to be upregulated in all three tissues, with the exception of downregulation of genes involved in biological oxidation in adipose tissue (**Figure 4.1D-F**). Individual tissues also showed perturbations of unique pathways: PPAR signaling and arachidonic acid pathways were altered in liver; extracellular matrix related processes were enriched among hypothalamic DEGs; core histone genes were upregulated in adipose DEGs (**Figure 4.1D-F**). In addition, triglyceride biosynthesis and glucose metabolism pathways were also moderately enriched among adipose DEGs, whereas few changes were seen for genes involved in adipocyte differentiation (**Figure 4.2**).

Prenatal BPA exposure induces tissue-specific epigenetic alterations in male weaning offspring

Consistent with the observed gene expression disruptions at the transcriptomic level, we observed numerous methylomic alterations using RRBS, which characterizes DNA methylation

states of millions of potential epigenetic sites at single base resolution. At FDR < 5%, 5136, 104, and 476 differentially methylated CpGs (DMCs) were found in adipose, hypothalamus, and liver tissues, respectively (**Figure 4.3A, Table 4.2**). Interestingly, BPA induced local methylation changes in *Gm26917* and *Yam1*, two lncRNAs with no previously known link to BPA, consistently across three tissues (**Figure 4.3B**). The majority of the DMCs are located in intergenic regions (32% - 38%), followed by introns (31% - 37%) and exons (13% - 15%), but there is a paucity of DMCs in the promoter region (3% - 5%) (**Figure 4.4**). Contrary to predictions that promoter regions may be more prone to epigenetic changes, we found that within-gene and intergenic methylation alterations in DNA methylation are more prevalent, a pattern consistently observed in previous epigenomic studies [160, 168]. In addition, 5.0%, 8.6%, and 8.1% DMCs overlap with repetitive DNA elements in adipose, hypothalamus, and liver, respectively, recapitulating previous report of the interaction between BPA and repetitive DNA [169].

For DMCs that are located within or adjacent to genes, we further tested whether the local genes adjacent to those DMCs show enrichment for known functional categories. Unlike DEGs, top processes enriched for DMCs concentrated on intra- and extra-cellular communication and signaling related pathways such as axon guidance, extracellular matrix organization and NGF signaling (**Figure 4.3C**, **full results in Table S4.3**). The affected genes in these processes are related to cellular structure, cell adhesion, and cell migration, indicating that these functions may be particularly vulnerable to BPA induced epigenetic modulation.

Potential regulatory role of DMCs in transcriptional regulation of BPA induced DEGs

To explore the role of DMCs in regulating DEGs, we evaluated the connection between transcriptome and methylome by correlating the expression level of DEGs with the methylation ratio of their local DMCs. For the DEGs in adipose, hypothalamus and liver tissue, we identified 42, 36, and 278 local DMCs whose methylation ratios were significantly correlated with the gene expression. At a global level, compared to non-DEGs, DEGs are more likely to contain local correlated DMCs (**Figure 4.5**). A closer look into the expression-methylation correlation by different chromosomal regions further revealed a context dependent correlation pattern (**Figure 4.3D**). In adipose and liver, the 3-5% of DMCs in promoter regions tend to show significant enrichment for negative correlation with DEGs, whereas gene body methylations for DEGs are more likely to show significant enrichment for positive correlation with gene expression. In hypothalamus, however, positive correlations between DEGs and DMCs are more prevalent across different gene regions. In addition, liver DMCs within lncRNAs were uniquely enriched for negative correlation with lncRNA expression, although the lack of a reliable mouse lncRNA target database prevented us from further investigating whether downstream targets of the lncRNAs were enriched in the DEGs. Specific examples of DEGs showing significant correlation with local DMCs include adipose DEG *Slc25a1* (Solute Carrier Family 25 Member 1, involved in triglyceride biosynthesis), hypothalamic DEG *Mvk* (Mevalonate Kinase, involved in cholesterol biosynthesis), and liver DEG *Gm20319* (a lncRNA with unknown function) (**Figure 4.6 and Table S4.4**). These results support a role of BPA-induced differential methylation in altering the expression levels of adjacent genes.

Pervasive influence of prenatal BPA exposure on the liver transcription factor network

BPA is known to bind to diverse types of nuclear receptors such as estrogen receptors and peroxisome proliferator-activated (PPAR) receptors that function as transcription factors (TFs), thus influencing the action of downstream genes [170, 171]. *PPARg* in particular has been shown to be a target of BPA in mouse and human and mechanistically linking BPA exposure with its associated effect on weight gain and increased adipogenesis [172-174]. To explore the TF regulatory landscape underlying BPA exposure based on our genome-wide data, we leveraged tissue-specific TF regulatory networks from the FANTOM5 project [175] and integrated it with our BPA transcriptome profiling data. No TF was found to be differentially expressed in adipose tissue, whereas 1 TF (*Pou3f1*) and 14 TFs (such as *Esrra*, *Hnf1a, Pparg, Tcf21, Srebf1*) were found to be differentially expressed in hypothalamus and liver, respectively. Due to the temporal nature of TF action, changes in TF levels may precede the downstream target genes and not be reflected in the transcriptomic profiles measured at the time of sacrifice. Therefore, we further curated the target genes of TFs from FANTOM5 networks and tested the enrichment for the target genes of each TF among our tissue-specific DEGs (**Table S4.5**). This analysis confirmed that BPA perturbs the activity of the downstream targets for estrogen receptors *Esrrg* (p = 1.4e-3, FDR = 1.9%) and *Esrra* (p = 0.03, FDR = 13%) in liver, as well as *Esr1* in both adipose (p = 7.2e-3, FDR = 10.6%) and liver (p = 7.2e-3, FDR = 4.7%). Targets of *Pparg* were also perturbed in liver (p = 4.1e-3, FDR = 3.8%). Therefore, we demonstrated that our data-driven network modeling is able to not only recapitulate results from previous *in-vitro* and *in-vivo* studies showing that BPA influences estrogen signaling and PPAR signaling [171], but also uniquely point to the tissue specificity of these BPA target TFs.

In addition to these expected TFs, we identified 14 adipose TFs and 61 liver TFs whose target genes were significantly enriched for BPA DEGs at FDR < 5%. Many of these TFs showed much stronger enrichment for BPA DEGs among their downstream targets than the estrogen receptors (**Table S4.5**). The liver TFs include multiple genes from the hepatocyte nuclear factors (HNF) family and the CCAAT-enhancer-binding proteins (CEBP) family, which are critical for liver development and function, suggesting a pervasive influence of BPA on liver TF regulation.

We further extracted the subnetwork containing 89 unique downstream targets of the significant liver TFs that are also liver DEGs. This subnetwork showed significant enrichment for genes involved in metabolic pathways such as steroid hormone biosynthesis and fatty acid metabolism. The regulatory subnetwork for the top liver TFs (FDR < 5%) revealed a highly interconnected TF subnetwork that potentially senses BPA exposure and in turn governs the expression levels of their targets (**Figure 4.7A**), with *Pparg* and *Hnf4* among the core TFs. Some of the TFs in this network, including *Esr1, Esrrg, Foxp1,* and *Tcf7l1,* also had local DMCs identified in our study, indicating that BPA may perturb this liver TF subnetwork via local modification of DNA methylation of key TFs.

Identification of potential non-TF regulators governing BPA induced molecular perturbations

To further identify regulatory genes that mediate the action of BPA on downstream targets through non-TF mechanisms, we leveraged data-driven tissue-specific Bayesian networks (BNs) generated from multiple independent human and mouse studies (**Table S4.1**). These data-driven networks are complementary to the TF networks used above and have proven valuable for accurately predicting gene-gene regulatory relationships and novel key drivers (KDs) [6, 18,

160-162]. KDs were defined as network nodes whose surrounding subnetworks are significantly enriched for BPA exposure related DEGs. At FDR < 1%, we identified 21, 1, and 100 KDs in adipose, hypothalamus, and liver, respectively (**Table S4.6**). The top KDs in adipose (top 5 KDs *Acss2, Pc, Agpat2, Slc25a1, Acly*), hypothalamus (*Fa2h*) and liver (top 5 KDs *Dhcr7, Aldh3a2, Fdft1, Mtmr11, Hmgcr*) were involved in cholesterol, fatty acid and glucose metabolism processes. In addition, three KDs, *Acss2* (Acetyl-Coenzyme A Synthetase 2), *Acat2* (Acetyl-CoA Acetyltransferase 2), and *Fasn* (Fatty Acid Synthase), were involved in the upregulation of DEGs in both adipose and liver, despite the fact that few DEG signatures overlap across tissues (**Figure 4.7B**). These KDs are consistent with the observed increased expression of several genes implicated in lipogenesis, including *Fasn*, and help explain the liver accumulation of triglycerides when mice are exposed to BPA [176]. Together, these results indicate that BPA may engage certain common regulators which have tissue-specific targets. The distinct upregulatory pattern within the subnetworks of individual KDs supports the potential functional importance of KDs in orchestrating the action of downstream genes. These KDs, along with the newly identified TFs from the above analysis, may represent novel regulatory targets which transmit the *in vivo* biological effects of BPA.

BPA transcriptomic and methylomic signatures are related to metabolic traits in mice

Aligning with the enrichment for metabolic genes in the above pathway and network analyses, we observed alterations in various metabolic phenotypes in the male offspring with prenatal BPA exposure. Compared with the control group (n = 9), offspring from the BPA group (n = 11) showed significantly lower body weight at weaning, indicative of post-natal growth retardation, a trait that is strongly associated with later life insulin resistance and obesity risk [177] (**Figure**

4.8A). There were also significant decreases in serum lipid parameters and an increase in serum glucose level at weaning age (**Figure 4.8B**). The decreases in the lipid parameters at this early developmental stage likely reflect the growth retardation phenotype observed and may provide feedback signals to increase the expression of *Fasn* and other lipid genes, which likely predispose the exposed offspring to increased triglyceride accumulation and other lipid dysregulation later in life.

To further assess the relationship between the BPA molecular signatures and metabolic traits in the mouse model, the DEGs and DMCs from individual tissues were tested for correlation with the measured metabolic traits: body weight, free fatty acids, total cholesterol, high density lipoprotein cholesterol, triglycerides and blood glucose. At $p < 0.05$, over two thirds of tissue-specific DEGs and over 60% DMCs were identified to be correlated with at least one metabolic trait (**Figure 4.8C, D**). Notably, liver DEGs exhibited stronger correlation with free fatty acid and triglycerides, whereas adipose DEGs were uniquely associated with glucose level, which is consistent with the pathway annotation results for these tissues. On the other hand, liver DMCs showed stronger correlations with metabolic traits than those from adipose and hypothalamus tissues.

Cross-examination of correlation across gene expression, DNA methylation, and metabolic traits revealed 35 consistent DEG-DMC-trait associations (3 in adipose, 4 in hypothalamus, and 28 in liver) (**Table S4.7**). For example, in adipose tissue, *Fasn* (also a shared KD in adipose and liver) was correlated with its exonic DMC at chr11:120816457, and both were correlated with triglyceride level; in hypothalamus, *Igf1r* (Insulin Like Growth Factor 1 Receptor) was correlated with its intronic DMC at chr7:68072768, and both were correlated with blood glucose

level; in liver, *Adh1* (Alcohol Dehydrogenase 1A) was correlated with its intronic DMC at chr3:138287690, and both were correlated with body weight (**Figure 4.9**). These results suggest that BPA alters local DMCs of certain genes to regulate gene expression, which may in turn regulate distinct metabolic traits.

Relevance of BPA signature to human complex traits/diseases

Human observational studies have associated developmental BPA exposure with a wide variety of human diseases ranging from cardiometabolic diseases to neuropsychiatric disorders [135, 136, 178]. Large-scale human genome-wide association studies offer an unbiased view of the genetic architecture for various human traits/diseases, and intersections of the molecular footprints of BPA in our mouse study with human disease risk genes can help infer the potential disease-causing properties of BPA in humans. From the GWAS Catalog [69], we collected associated genes for 161 human traits/diseases (traits with fewer than 50 associated genes were excluded), and evaluated the enrichment for the trait associated genes among DEG and DMC signatures. At FDR < 5%, no trait was found to be significantly enriched for BPA DEGs. Surprisingly, despite the difference between tissue-specific DMCs (**Figure 4.3B**), 19 out the 161 traits showed consistently strong enrichment for DMCs across all three tissues at FDR < 1%. The top traits include body mass index (BMI) and type 2 diabetes (**Table 4.3**). As DNA methylation status is known to determine long-term gene expression pattern instead of immediate dynamic gene regulation, the BMI and diabetes associated genes may be under long-term programming by BPA-induced differential methylation, thereby affecting later disease risks.

The above analysis involving the GWAS catalog focused only on small sets of the top candidate genes for various diseases and may have limited statistical power. To improve the statistical power, we curated the full summary statistics from 61 human GWAS that are publicly available (covering millions of SNP-trait associations in each GWAS), which enabled us to extend the assessment of disease association by considering additional human disease genes with moderate to low effect sizes (See **Methods**). This analysis showed that DEGs from all three tissues exhibited consistent enrichment for genes associated with lipid traits such as triglycerides, LDL, and HDL **(Figure 4.10A-C**). Interestingly, enrichment for birth weight and birth length was also observed for hypothalamus and liver signatures, respectively. Liver DEGs were also significantly associated with coronary artery disease, inflammatory bowel disease, Alzheimer's disease, and schizophrenia. Top DEGs driving the inflammatory bowel disease association involve immune and inflammatory response genes (*PSMB9, TAP1, TNF*), whereas association with Alzheimer's disease and schizophrenia involve genes related to cholesterol homeostasis (*APOA4, ABCG8, SOAT2*) and mitochondrial function (*GCDH, PDPR, SHMT2*), respectively. These results suggest that tissue-specific targets of BPA are connected to diverse human complex diseases through both the central nervous system and peripheral tissues.

4.3 Discussion

This multi-tissue, multi-omics integrative study represents one of the first systems biology investigations of prenatal BPA exposure. By integrating systematic profiling of the transcriptome and methylome of multiple metabolic tissues with phenotypic trait measurements, large-scale human association datasets, and network analysis, we uncovered insights into the molecular regulatory mechanisms underlying the health effect of prenatal BPA exposure. Specifically, we

identified tens to hundreds of tissue-specific DEGs and DMCs involved in diverse biological functions such as metabolic pathways (oxidative phosphorylation/TCA cycle, fatty acid, cholesterol, glucose metabolism, and PPAR signaling), extracellular matrix, focal adhesion, and inflammation (arachidonic acid), with DMCs partially explaining the regulation of DEGs. Network analysis helped reveal potential regulatory circuits post BPA exposure and pinpointed both tissue-specific and cross-tissue regulators of BPA activities, including TFs such as estrogen receptors, *PPARg*, and *HNF1A*, and non-TF key drivers such as *FASN*. Furthermore, the BPA gene signatures and the predicted regulators were found to be linked to a wide spectrum of disease-related traits in both mouse and human.

The large-scale disruption we observed in the transcriptome and methylome in adipose and liver was consistent with previous reports [153, 156, 179, 180]. However, our unique study design of examining multi-omics in multiple tissues in parallel yields higher comparability when integrating the results between data types and across tissues, as they were from the same set of animals and were profiled in the same conditions. Furthermore, our advanced multidimensional integrative approach provides deeper insights into the regulatory cascades within and across tissues. Across all three tissues at the transcriptome level, we found that lipid metabolism and energy homeostasis related processes were consistently perturbed, with the scale of perturbation being strongest in liver. This aligns well with the significant changes in the plasma lipid profiles we observed in the offspring, the reported perturbation of lipid metabolism in fetal murine liver [180], and the reported susceptibility for nonalcoholic fatty liver diseases following BPA exposure [181-183]. The only shared gene across tissues, *CYP51*, is involved in cholesterol and sterol biosynthesis and beta oxidation, again supporting that metabolism is a central target of

BPA. At the methylome level, we are able to replicate 5 out of 7 peak hypomethylated genes, and 6 out of 9 peak hypermethylated genes from a study focusing on the gonadal adipose tissue [156]. We also revealed an intriguing link between BPA and lncRNAs across tissues, whose functional importance in developmental processes, disease progression, and response to BPA exposure was increasingly recognized yet underexplored [184]. Our molecular data provides intriguing lncRNA candidates such as *Gm20319*, *Gm26917*, and *Yam1* for future in-depth functional analyses.

For adipose tissue, clusters of genes responsible for core histones were found to be uniquely altered. Along with the strong adipose-specific differential methylation status, our results revealed gonadal adipose tissues as an especially vulnerable site for BPA induced epigenetic reprogramming. Besides, developmental BPA exposure has been previously suggested to influence white adipocyte differentiation [185-187]. However, the adipocyte differentiation pathway was not significantly enriched in our study. This is consistent with the report by Angel et al. [187], where increased adipocyte number is only found in mouse offspring with prenatal BPA exposure at 5ug/kg/day and 500ug/kg/day, but not 5mg/kg/day. Additionally, we found significant changes in triglyceride biosynthesis and glucose metabolism genes, suggesting that prenatal BPA exposure affects fat storage and glucose homeostasis in the adipose tissue. Although here we mainly investigate gonadal adipose tissue as a surrogate for abdominal fat in the context of metabolic disorders, the information may be useful for exploring the relationship between this fat depot and the gonad.

With regards to the hypothalamus, our study is the first to investigate the effect of BPA on the hypothalamic transcriptome and DNA methylome. Hypothalamus is an essential brain region

that regulates the endocrine system, peripheral metabolism, and numerous brain functions. We identified BPA-induced DEGs and DMCs that were enriched for extracellular matrix related processes such as axon guidance, focal adhesion, and various metabolic processes. These hypothalamic pathways have been previously associated with metabolic [160, 161] and neurodegenerative diseases [160, 188], and they could underlie the reported disruption of hypothalamic melanocortin circuitry after BPA exposure [189]. Our study highlights the hypothalamus as another critical yet under-recognized target for BPA.

By interrogating both the transcriptome and DNA methylome in matching tissues, we were able to directly assess both global and specific correlative relationships between DEGs and DMCs (**Figure 4.3D**). Specifically, we found that DEGs are more likely to have correlated DMCs in the matching tissue, a trend that persists in non-promoter regions. Our results corroborate previous findings regarding the importance of gene body methylation in disease etiology [190, 191]. Given that over 90% of DMCs were found in non-promoter regions, closer investigation of the regulatory circuits involving these regions may unveil new insights into BPA response [168].

Known as an endocrine disrupting chemical, BPA has been speculated to exert its primary biological action by modifying the activity of hormone receptors, including estrogen receptors, *PPARg* and glucocorticoid receptors [171]. Indeed, the activity for the downstream targets of *Pparg* and three estrogen and estrogen-related receptors were found to be disrupted in liver by prenatal BPA exposure. Nevertheless, our unbiased data-driven analysis revealed many novel transcription factors and non-TF regulatory genes that also potentially mediate BPA effects. In fact, many of the newly identified TF targets of BPA, such as several hepatic nuclear factors, showed much higher ranking in our regulator prediction analyses. In liver, a tightly inter-

connected TF subnetwork was highly concentrated with BPA affected genes involved in metabolic processes such as cytochrome P450 system (*CYP3A4, CYP2A6, CYP1A2*), lipid (*APOA4, ABCG5, SOAT2*) and glucose (*HNF1A, ADRA1B, GCK*) regulation, with extensive footprints of altered methylation status in the TFs and other subnetwork genes (**Figure 4.7A**). Therefore, our results support a widespread impact of BPA on liver transcriptional regulation, and the convergence of differential methylation and gene expression in this TF subnetwork implies that BPA perturbs this subnetwork via epigenetic regulation of the TFs, which in turn trigger transcriptomic alterations in downstream genes. Of the non-TF regulators identified in our study, the cross-tissue KDs *Acss2*, *Acat2*, and *Fasn*, are of special interest. Based on our network prediction, these three KDs appeared to up-regulate distinct groups of lipid metabolism genes in adipose and liver post-BPA exposure (**Figure 4.7B**), supporting their role in mediating the BPA-induced lipid dysregulation at the systemic level. The significant correlation of gene expression and methylation for *Fasn* with triglyceride level furthers implicates its role as a network-level biomarker for BPA induced lipid dysregulation. Our observation of *Fasn* is consistent with evidences suggesting its susceptibility to methylation perturbation under obesogenic feeding [192] and its functional importance for fatty liver diseases [162, 193].

One unique aspect of this study is the linking of the molecular landscape of prenatal BPA exposure to traits/diseases in both mouse and human. In our mouse study, the observed changes in body weight, lipid profiles, and glucose level are highly concordant with the functions of the molecular targets. For instance, prenatal BPA exposure perturbs both the expression levels and the local DNA methylation status of *Fasn*, *Igf1r*, and *Adh1*. These DEGs and their local DMCs also significantly correlate with phenotypic outcomes, thus serving as examples of how DNA

methylation and gene expression bridge the gap between BPA exposure and phenotypic manifestation. To further enhance the translatability of our findings from mouse to human, we searched for human diseases linked to the BPA-affected genes. An intriguing discovery is the prominent overrepresentation of differential methylation signals in adipose, hypothalamus, and liver within known genes related to obesity and type 2 diabetes, supporting that BPA may impact obesity and diabetes risk through systemic reprogramming of DNA methylation. More sophisticated analysis incorporating the BPA differential gene expression and the full statistics of human genome-wide association studies corroborated the potential of prenatal BPA exposure to affect lipid homeostasis [194], birth weight [195], and coronary artery disease [135] reported in observational studies. Moreover, our findings suggest a potential involvement of prenatal BPA exposure in the development of inflammatory bowel syndrome, schizophrenia, and Alzheimer's disease. These associations warrant future investigations.

One limitation of our work is the restriction of study scope to weaning age male mice with *in utero* BPA exposure below the NOAEL (5mg/kg/d) as a proof-of-concept for our systems biology framework. Considering that the effects of early-life exposure to BPA is highly variable and dependent on factors such as the dose, window, route, and frequency of exposure as well as genetic background, age, and sex [134], future studies testing these additional variables are necessary to generate a comprehensive understandings of BPA risks under various exposure conditions.

4.4 Conclusions

Our study represents the first multi-tissue, multi-omics integrative investigation of prenatal BPA exposure. The systems biology framework we applied revealed how BPA triggers cascades of regulatory circuits involving numerous transcription factors and non-TF regulators that coordinate diverse molecular processes within and across core metabolic tissues, thereby highlighting that BPA exerts its biological functions via much more diverse targets than previously thought. As such, our findings offer a comprehensive systems-level understanding of tissue sensitivity and molecular perturbations elicited by prenatal BPA exposure and offer promising novel candidates for targeted mechanistic investigation as well as much-needed network-level biomarkers of prior BPA exposure. The strong influence of BPA on metabolic pathways and cardiometabolic phenotypes merits it characterization as a general metabolic disruptor posing systemic health risks.

4.5 Methods

Mouse model of prenatal BPA exposure

Inbred C57BL/6 mice were maintained on a special diet 5V01 (LabDiet), certified to contain less than 150ppm estrogenic isoflavones, and housed under standard housing conditions (room temperature 22–24°C) with 12:12 hr light:dark cycle before mating at 8-10 weeks of age. From 1-day post-conception (dpc) to 20 dpc, BPA (Sigma-Aldrich, St. Louis, MO) dissolved in corn oil was administered to pregnant female mice via oral gavage (mimicking common exposure route in humans) at 5mg/kg/day on a daily basis. Control mice were fed the same amount of empty vehicle. BPA exposure was restricted to experimental manipulation through the use of

polycarbonate-free water bottles and cages. Offspring from each treatment were maintained on a standard chow diet (Newco Distributors Inc, Rancho Cucamonga, CA).

Characterization of cardiometabolic phenotypes and tissue collection

Body weight of offspring was measured daily from postnatal day 5 up to the weaning age of 3 weeks. Mice were fasted overnight before sacrifice, and plasma samples were collected through retro-orbital bleeding. Serum lipid and glucose traits including total cholesterol, high density lipoprotein cholesterol (HDL), un-esterified cholesterol (UC), triglyceride (TG), free fatty acid (FFA), and glucose were measured by enzymatic colorimetric assays at UCLA GTM Mouse Transfer Core as previously described [160]. Gonadal white adipose tissue, hypothalamus, and liver tissues were collected from each animal, flash frozen in liquid nitrogen, and stored at –80°C. For white adipose tissue, we chose the gonadal depot mainly due to its similarity to abdominal fat, established relevance to cardiometabolic risks, tissue abundance, and the fact that it is the most well-studied adipose tissue in mouse models. All mouse experiments were conducted in accordance with and approved by the Institutional Animal Care and Use Committee at University of California, Los Angeles.

RNA sequencing (RNA-seq) and data analysis

A total of 18 RNA samples were isolated from gonadal adipose, hypothalamus and liver tissues (n = 3 per group per tissue) from male offspring using the AllPrep DNA/RNA Mini Kit (QIAGEN GmbH, Hilden, Germany). Samples were processed for library preparation using TruSeq RNA Library Preparation Kit (Illumina, San Diego, CA) for poly-A selection, fragmentation, and reverse transcription using random hexamer-primers to generate first-strand

cDNA. Second-strand cDNA was generated using RNase H and DNA polymerases, and sequencing adapters were ligated using the Illumina Paired-End sample prep kit. Library products of 250-400bp fragments were isolated, amplified, and sequenced with Illumina Hiseq2500 System. After quality control using FastQC [196], the HISAT-StringTie pipeline [197] was used for sequence alignment and transcript assembly. Identification of differentially expressed genes (DEGs) were conducted using DEseq2 [198]. To account for multiple testing, we used the q-value method [199]. After excluding genes with extremely low expression levels (FPKM < 1), only DEGs demonstrating differential expression comparing the BPA and control groups per tissue at a false discovery rate (FDR) < 5% were used for biological pathway analysis, network analysis, and phenotypic data integration, as described below.

Reduced representation bisulfite sequencing (RRBS) and data analysis

We constructed RRBS libraries for 18 DNA samples from adipose, hypothalamus and liver tissues (n = 3 per group per tissue). The DNA samples were quantified using the dsDNA BR assay (Qubit, Waltham, MA) and 100ng of DNA was used for library preparation. After digestion of the DNA with the MspI enzyme, samples underwent an end-repair and adenylation process, followed by adapter ligation using the Truseq barcode adapter (Illumina, San Diego, CA), size selection using AMPure Beads (Beckman Coulter, Brea, CA), and bisulfite treatment using the Epitect Kit (Qiagen, Germantown, MD). Bisulfite-treated DNA was then amplified using the Truseq Library Prep Kit (Illumina, San Diego, CA) and sequenced with the Illumina Hiseq2500 System. Bisulfite-converted reads were processed and aligned to the reference mouse genome (GRCm38/mm10 build) using the bisulfite aligner BSMAP [200]. We then used MOAB [201] for methylation ratio calling and identification of differentially methylated CpGs (DMCs).

FDR was estimated using the q-value approach. Loci with methylation level changes of > 5% between BPA and control groups and FDR < 0.05 for each tissue were considered statistically significant DMCs. To annotate the locations of the identified DMCs in relation to gene regions and repetitive DNA elements accessed from UCSC genome browser, we used the Bioconductor package "annotatr" [202]. Specifically, gene regions were categorized into 1) 1-5kb upstream of the transcription start site (TSS), 2) promoter (< 1kb upstream of the TSS), 3) 5' untranslated region (UTR), 4) exons, 5) introns, and 6) 3'UTR. The "annotatr" package was also used to annotate DMCs for known long non-coding RNAs (lncRNAs) based on GENCODE Release M16. Over-representation of DMCs within each category was calculated using Fisher's exact test. We further evaluated the link between DEGs and their local DMCs (DMCs annotated as any of the 6 above mentioned gene regions) by correlating the methylation ratio of DMCs with the expression level of DEGs.

Pathway, network, and disease association analyses of DEGs and DMCs using the Mergeomics R package

To investigate the functional connections among the BPA-associated DEGs or DMCs (collectively referred to as molecular signatures of BPAs) and to assess the potential association of BPA affected genes with diseases in human populations, we utilized the Mergeomics package [75], an open-source bioconductor package (https://bioconductor.org/packages/devel/bioc/html/Mergeomics.html) designed to perform various integrative analyses in multi-omics studies. Mergeomics consists of two main libraries, Marker Set Enrichment Analysis (MSEA) and Weighted Key Driver Analysis (wKDA). In the current study, we used MSEA to assess 1) whether known biological processes, pathways or

transcription factor targets were enriched for BPA molecular signatures as a means to annotate the potential functions or regulators of the molecular signatures, and 2) whether the BPA signatures demonstrate enrichment for disease associations identified in human genome-wide association studies (GWAS) of various complex diseases (**Figure 4.11**). wKDA leverages gene network topology (interactions or regulatory relations among genes) and edge weight (strength or reliability of interactions and regulatory relations) information of graphical gene networks to predict potential key regulators of a given group of genes, in this case, the BPA-associated DEGs. Both MSEA and wKDA were built around a chi-square like statistics (See **Appendix**) that yields robust findings that have been experimentally validated [75, 161, 162]. Details of each usage of the Mergeomics package are discussed below.

Functional annotation of DEGs and DMCs

To infer the functions of the DEGs and DMCs affected by BPA, we used MSEA to annotate the DEGs or local genes adjacent to the DMCs with known biological pathways curated from the Kyoto Encyclopedia of Genes and Genomes (KEGG) [49] and Reactome [48]. In brief, we extracted the differential expression p-values of genes in each pathway from the differential expression or methylation analyses and compared these p-values against the null distribution of p-values from random gene sets with matching gene numbers. If genes in a given pathway collectively show more significant differential expression or differential methylation p-values compared to random genes based on a chi-square like statistic, we annotate the DEGs or DMCs using that pathway (See **Appendix**). DEGs and DMCs can have multiple over-represented pathways.

Identification of transcription factor (TF) hotspots perturbed by BPA

To dissect the regulatory cascades of BPA, we first assessed whether BPA-associated DEGs were downstream targets of specific transcription factors. The hypothesis behind this analysis is that BPA first affects TFs which in turn regulate the expression of downstream genes. We used TF regulatory networks for adipose, brain, and liver tissue retrieved from the FANTOM5 database [175]. Note that only a whole brain (instead of hypothalamus) TF network was available, which may only partially represent hypothalamic gene regulation. Each TF network was processed to keep the edges with high confidence (See **Appendix**). To identify TFs whose targets were perturbed by BPA, the downstream nodes of each TF in the network were pooled as the target genes for that TF. We then assessed the enrichment for BPA exposure related DEGs among the target genes of each TF using MSEA. TFs with FDR < 5% were considered statistically significant. Cytoscape software was used for TF network visualization [116].

Bayesian network and Weighted Key Driver Analysis (wKDA) to identify potential non-TF regulators

To further identify non-TF regulators that sense BPA and then perturb downstream genes, we used Bayesian networks (BN) of adipose, hypothalamus and liver tissues constructed from genetic and transcriptomic data from several large-scale mouse and human studies (**Table S4.1,** See **Appendix**). wKDA was used to identify network key drivers (KDs), which are defined as network nodes whose neighboring subnetworks are significantly enriched for BPA-associated DEGs. Briefly, wKDA takes gene set G (i.e. BPA DEGs) and directional gene network N (i.e. BNs) as inputs. For every gene K in network N, neighboring genes within 1-edge distance were

tested for enrichment of genes in G using a chi-square like statistics followed by FDR assessment by permutation (See **Appendix**). Network genes that reached FDR < 0.05 were reported as potential KDs.

Association of BPA DEGs and DMCs with mouse phenotypes and human diseases/traits

To assess whether the BPA molecular signatures were related to phenotypes examined in the mouse offspring, we calculated the Pearson correlation coefficient among expression level of DEGs, methylation ratio of DMCs, and the measurement of metabolic traits. For human diseases or traits, we accessed the GWAS catalog database [69] and collected the lists of candidate genes reported to be associated with 161 human traits/diseases (P < 1e-5). These genes were tested for enrichment of the BPA DEGs and DMCs in our mouse study using MSEA. We further curated all publicly available full summary statistics for 61 human traits/diseases from various public repositories (**Table S4.2,** See **Appendix**). This allowed us to apply MSEA to comprehensively assess the enrichment for human disease association among BPA transcriptomic signatures using the full-spectrum of large-scale human GWAS. For each tissue-specific gene signature, we used the SNPs within a 50kb chromosomal distance as the representing SNPs for that gene. The trait/disease association p-values of the SNPs were then extracted from each GWAS and compared to the p-values of SNPs of random sets of genes to assess whether the BPA signatures were more likely to show stronger disease association in human GWAS (**Figure 4.11,** See **Appendix**). This strategy has been successfully used in our previous animal model studies to assess the connection of genes affected by environmental perturbations such as diets and trauma to various human diseases [160, 203].

4.6 Tables

Table 4.1 Count of differentially expressed genes in adipose, hypothalamus and liver tissue following prenatal exposure to BPA (n = 3).

	FDR < 5%			P < 0.01		
	All	Up-regulated	Down-regulated	All	Up-regulated	Down-regulated
Adipose	86	46	40	213	123	90
Hypothalamus	93	65	28	375	234	141
Liver	855	439	416	980	493	487

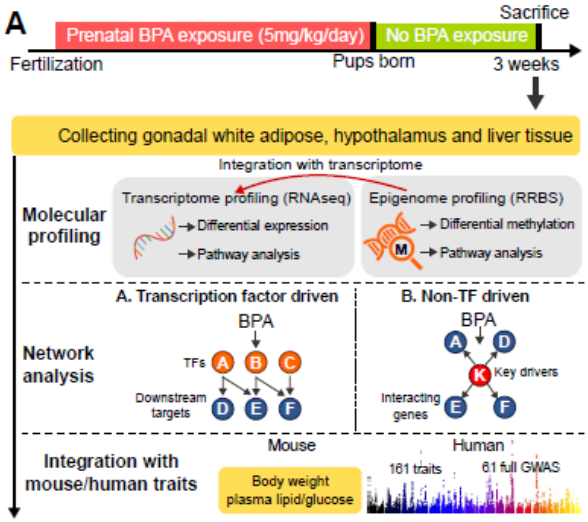
Table 4.2 Count of differentially methylated regions in hypothalamus and liver tissue following prenatal exposure to BPA (n = 3).

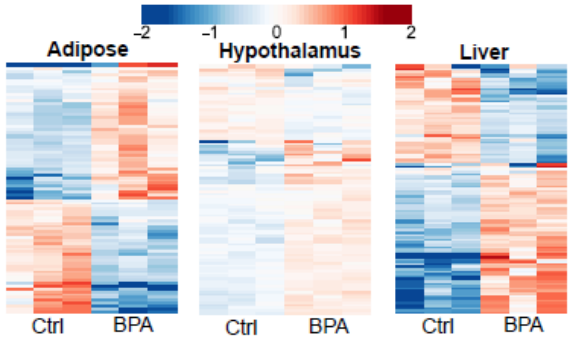
	FDR < 5%			P < 0.01		
	All	Hyper methylated	Hypo methylated	All	Hyper methylated	Hypo methylated
Adipose	5136	2413	2723	38771	17928	20846
Hypothalamus	104	49	55	8321	3862	4459
Liver	476	326	150	14025	7413	6612

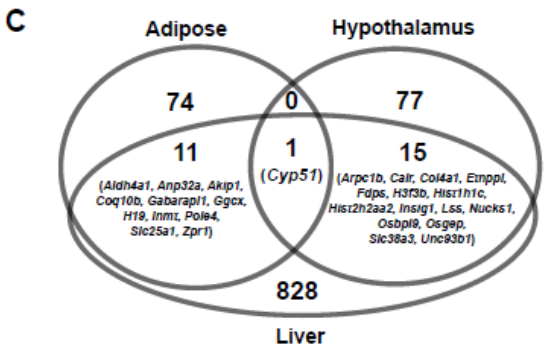
Table 4.3 Top 5 human traits whose associated genes in genome-wide association studies are enriched for differentially methylated CpGs (DMCs) across adipose, hypothalamus and liver at FDR < 1% in MSEA.

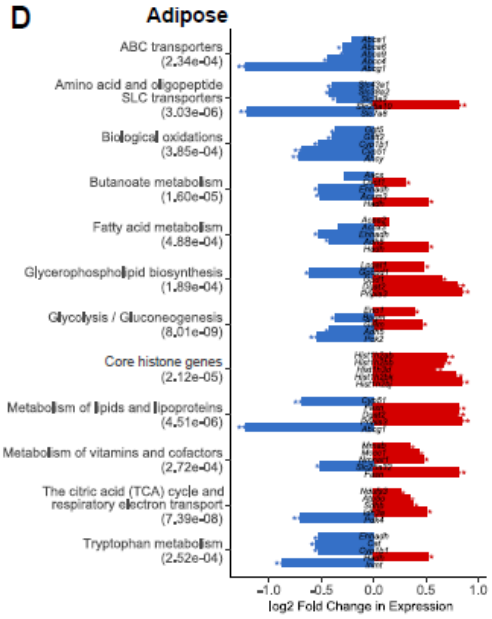
Human trait	Adipose		Hypothalamus		Liver	
	P	FDR	P	FDR	P	FDR
Obesity-related traits	1.28E-16	0.00%	3.03E-15	0.00%	2.71E-19	0.00%
Body mass index	1.30E-13	0.00%	3.74E-07	0.00%	9.66E-12	0.00%
Post bronchodilator FEV1/FVC ratio	8.17E-09	0.00%	1.45E-08	0.00%	3.67E-07	0.00%
Type 2 diabetes	1.21E-05	0.03%	8.97E-09	0.00%	0.001243	0.92%
Platelet distribution width	8.16E-08	0.00%	7.62E-05	0.16%	5.20E-05	0.12%

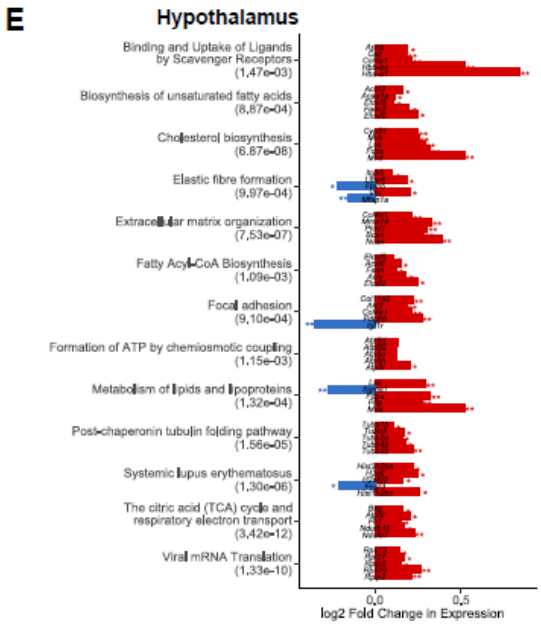
4.7 Figures


A
Prenatal BPA exposure (5mg/kg/day)
No BPA exposure
Sacrifice
Fertilization
Pups born
3 weeks
Collecting gonadal white adipose, hypothalamus and liver tissue
Integration with transcriptome
Molecular profiling
Transcriptome profiling (RNAseq)
Differential expression
Pathway analysis
Epigenome profiling (RRBS)
Differential methylation
Pathway analysis
Network analysis
A. Transcription factor driven
BPA
TFs
Downstream targets
B. Non-TF driven
BPA
Key drivers
Interacting genes
Integration with mouse/human traits
Mouse
Body weight
plasma lipid/glucose
Human
161 traits
61 full GWAS


B
Adipose
Hypothalamus
Liver
Ctrl
BPA
Ctrl
BPA
Ctrl
BPA


C
Adipose
Hypothalamus
74
0
77
11
1
(Cyp51)
15
(Aldh4a1, Anp32a, Akip1, Coq10b, Gabarapl1, Ggcx, H19, Inmt, Pole4, Slc25a1, Zpr1)
(Arpc1b, Calr, Col4a1, Etnppl, Fdps, H3f3b, Hist1h1c, Hist2h2aa2, Insig1, Lss, Nucks1, Osbpl9, Osgep, Slc38a3, Unc93b1)
828
Liver


D
Adipose
ABC transporters (2.34e-04)
Amino acid and oligopeptide SLC transporters (3.03e-06)
Biological oxidations (3.85e-04)
Butanoate metabolism (1.60e-05)
Fatty acid metabolism (4.88e-04)
Glycerophospholipid biosynthesis (1.89e-04)
Glycolysis / Gluconeogenesis (8.01e-09)
Core histone genes (2.12e-05)
Metabolism of lipids and lipoproteins (4.51e-06)
Metabolism of vitamins and cofactors (2.72e-04)
The citric acid (TCA) cycle and respiratory electron transport (7.39e-08)
Tryptophan metabolism (2.52e-04)
log2 Fold Change in Expression


E
Hypothalamus
Binding and Uptake of Ligands by Scavenger Receptors (1.47e-03)
Biosynthesis of unsaturated fatty acids (8.87e-04)
Cholesterol biosynthesis (6.87e-08)
Elastic fibre formation (9.97e-04)
Extracellular matrix organization (7.53e-07)
Fatty Acyl-CoA Biosynthesis (1.09e-03)
Focal adhesion (9.10e-04)
Formation of ATP by chemiosmotic coupling (1.15e-03)
Metabolism of lipids and lipoproteins (1.32e-04)
Post-chaperonin tubulin folding pathway (1.56e-05)
Systemic lupus erythematosus (1.30e-06)
The citric acid (TCA) cycle and respiratory electron transport (3.42e-12)
Viral mRNA Translation (1.33e-10)
log2 Fold Change in Expression

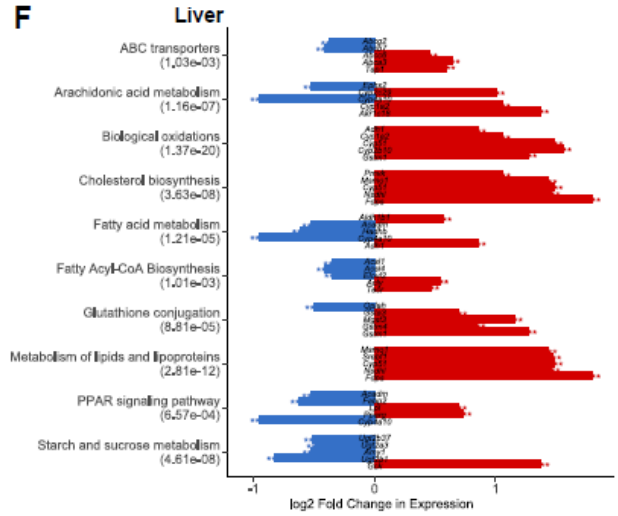

F
Liver
ABC transporters (1.03e-03)
Arachidonic acid metabolism (1.16e-07)
Biological oxidations (1.37e-20)
Cholesterol biosynthesis (3.63e-08)
Fatty acid metabolism (1.21e-05)
Fatty Acyl-CoA Biosynthesis (1.01e-03)
Glutathione conjugation (8.81e-05)
Metabolism of lipids and lipoproteins (2.81e-12)
PPAR signaling pathway (6.57e-04)
Starch and sucrose metabolism (4.61e-08)
log2 Fold Change in Expression

Figure 4.1 Overall study design and transcriptomic alterations in adipose, hypothalamus and liver.

A) Framework of multi-omics approaches to investigate the impact of prenatal BPA exposure. **B**) Heatmap of expression changes in adipose, hypothalamus and liver for the top 100 differentially expressed genes (DEGs) affected by BPA. Color indicates fold change of expression, with red and blue indicating upregulation and downregulation by BPA. **C**) Venn Diagram demonstrating tissue-specific and shared DEGs between tissues. **D-F**) Significantly enriched pathways (FDR < 5%) among DEGs from each tissue. Enrichment p-value (shown in parenthesis following the name of functional annotation) is determined by MSEA. The fold change and statistical significance for the top 5 differentially expressed genes in each pathway are shown. *, $p < 0.05$; **, FDR < 5% in differential expression analysis using DEseq2.

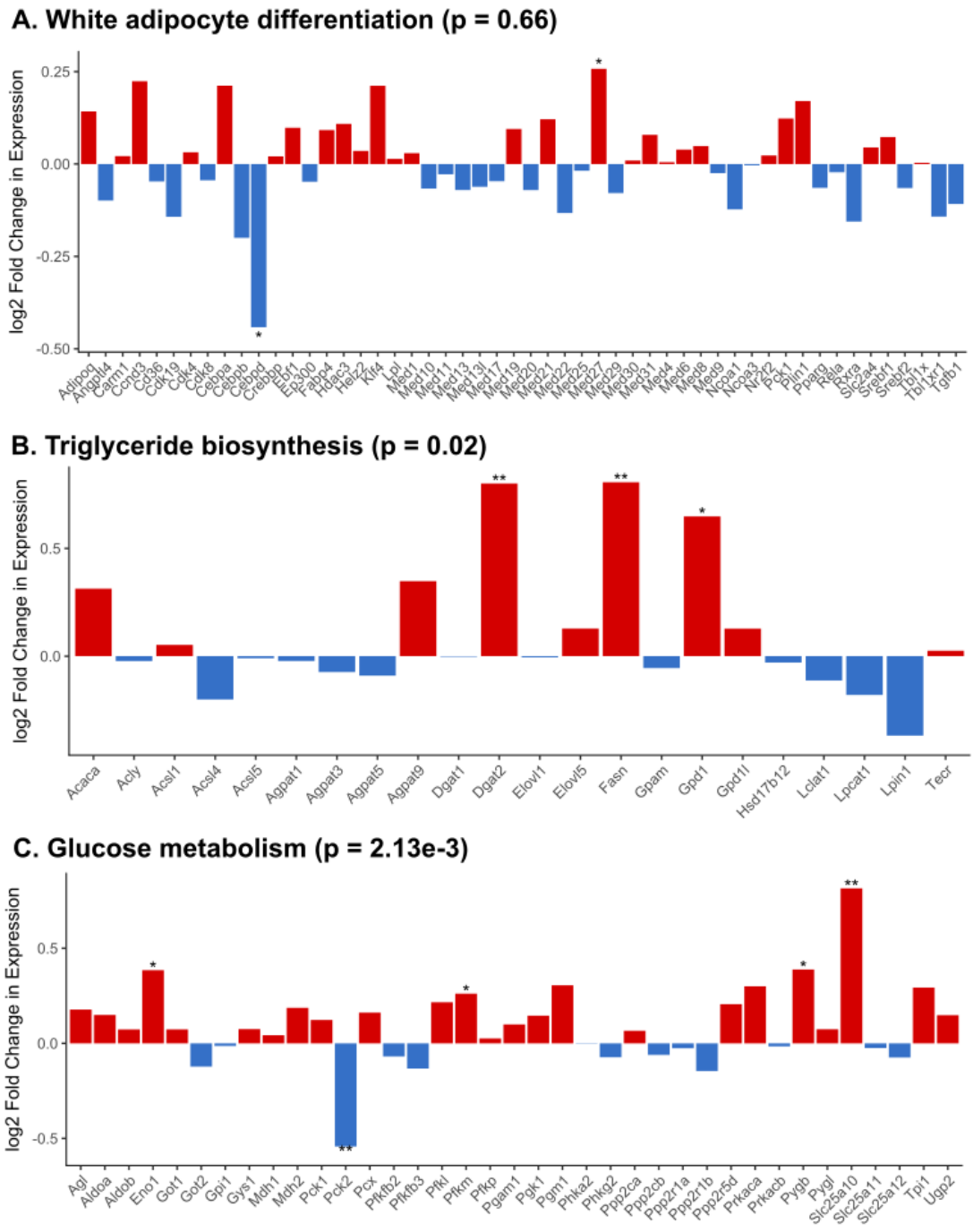


Figure 4.2 Prenatal BPA exposure induced expression change for genes from the adipocyte differentiation (**A**), triglyceride biosynthesis (**B**) and glucose metabolism (**C**) in the adipose tissue.

P-values for enrichment of pathway genes among DEGs (shown in parenthesis in each panel heading) were determined by MSEA. *p < 0.05 in differential expression tests for individual genes by DEseq2; **FDR < 5% in differential expression tests for individual genes by DEseq2.

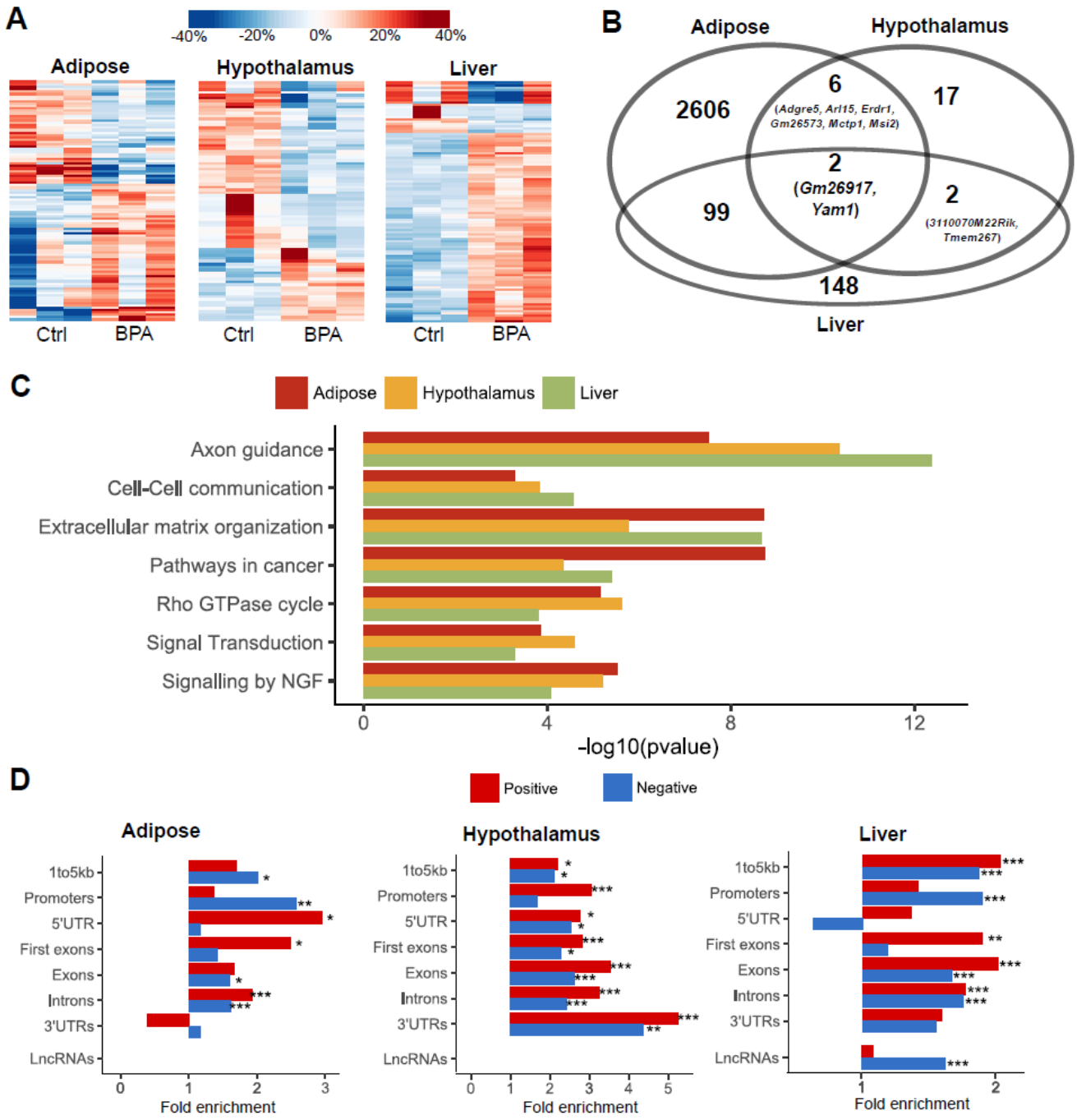


Figure 4.3 Prenatal BPA exposure induced methylomic level alteration in adipose, hypothalamus and liver.

A) Heatmap of methylation level changes for the top 100 differentially methylated CpGs (DMCs). Color indicates change in methylation ratio, with red and blue indicating upregulation and downregulation by BPA. **B**) Venn Diagram of genes with local DMCs between tissues shows tissue-specific and shared genes mapped to DMCs. **C**) Significantly enriched pathways that satisfied FDR < 1% across DMCs from adipose, hypothalamus, and liver tissues.

Enrichment p-value is determined by MSEA. **D**) Fold enrichment for positive correlations (red bars) or negatively correlations (blue bars) between DMCs and local DEGs, assessed by different gene regions. *, $p < 0.05$; **, $p < 0.01$; ***, $p < 0.0001$; enrichment p-values were determined using Fisher's exact test.

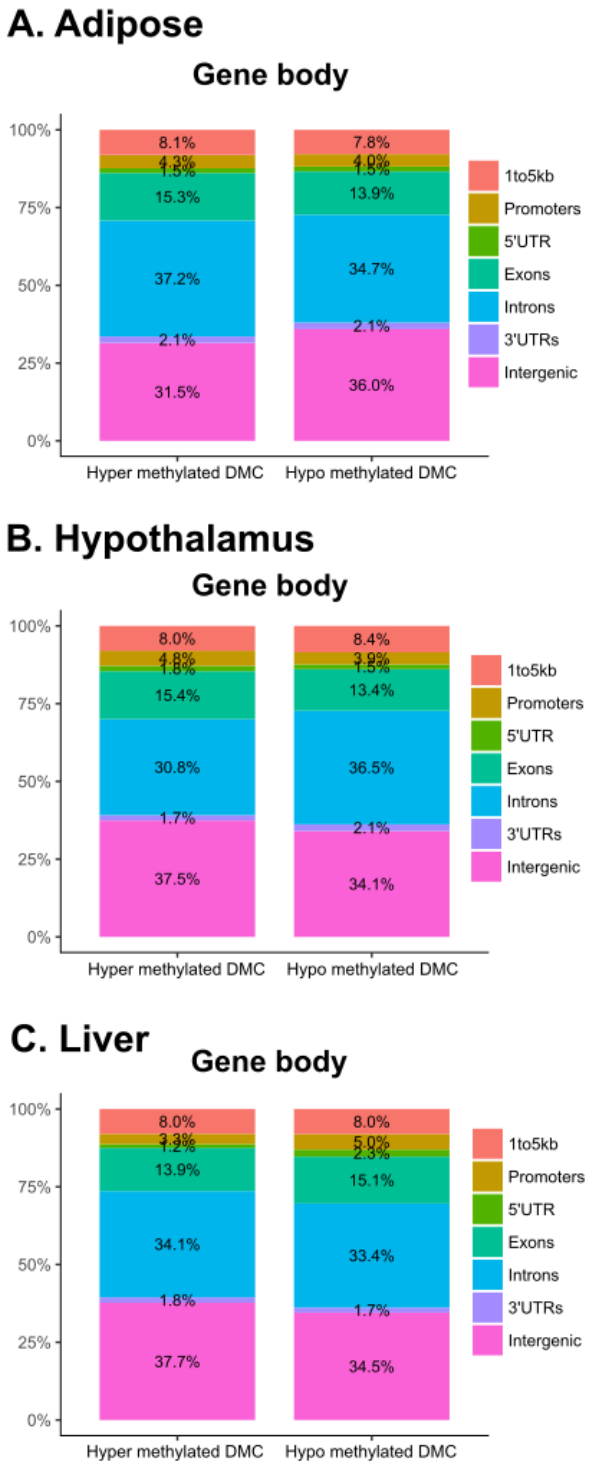


Figure 4.4 Gene body location distribution for hyper- and hypo- methylated DMC s in adipose (**A**), hypothalamus (**B**), and liver (**C**).

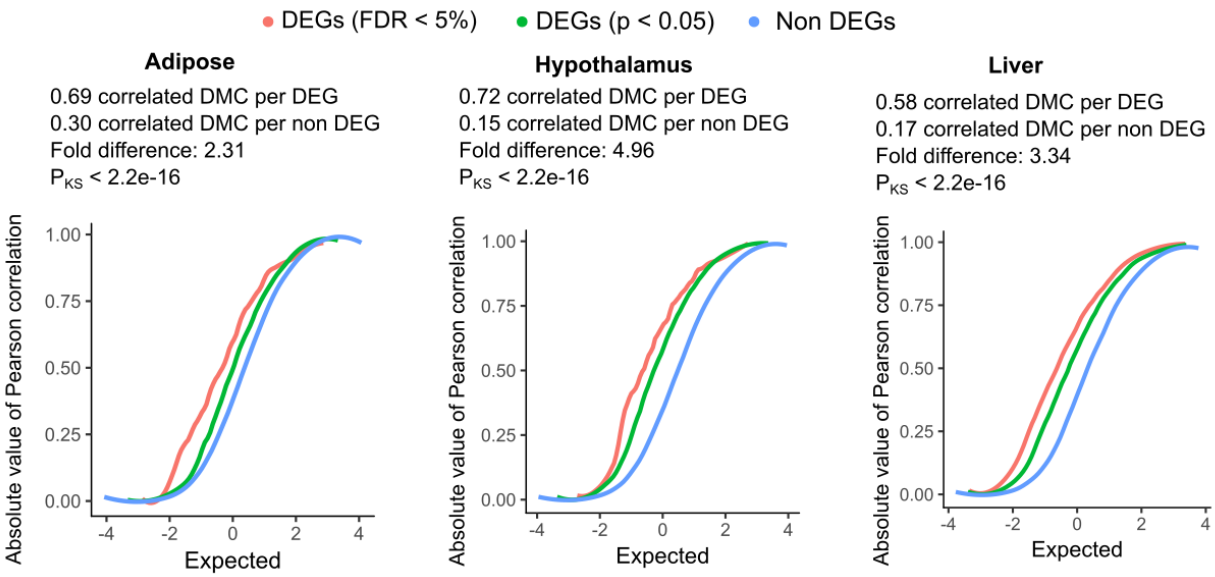


Figure 4.5 Quantile-quantile plots for the absolute Pearson correlation with local DMC for DEGs and Non DEGs in adipose, hypothalamus, and liver tissue.

Statistical difference of the distribution of correlation value between DEGs (FDR < 5%) and non DEGs is determined by the Kolmogorov–Smirnov test.

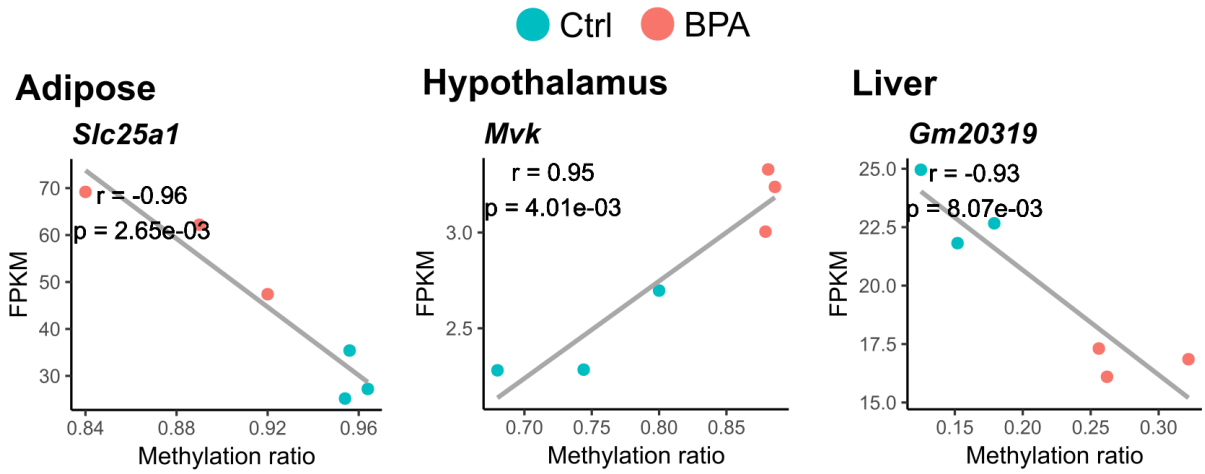


Figure 4.6 Scatter plots of correlations between DEG expression levels and DMC methylation ratios for *Slc25a1* in adipose, *Mvk* in hypothalamus, and *Gm20319* in liver.

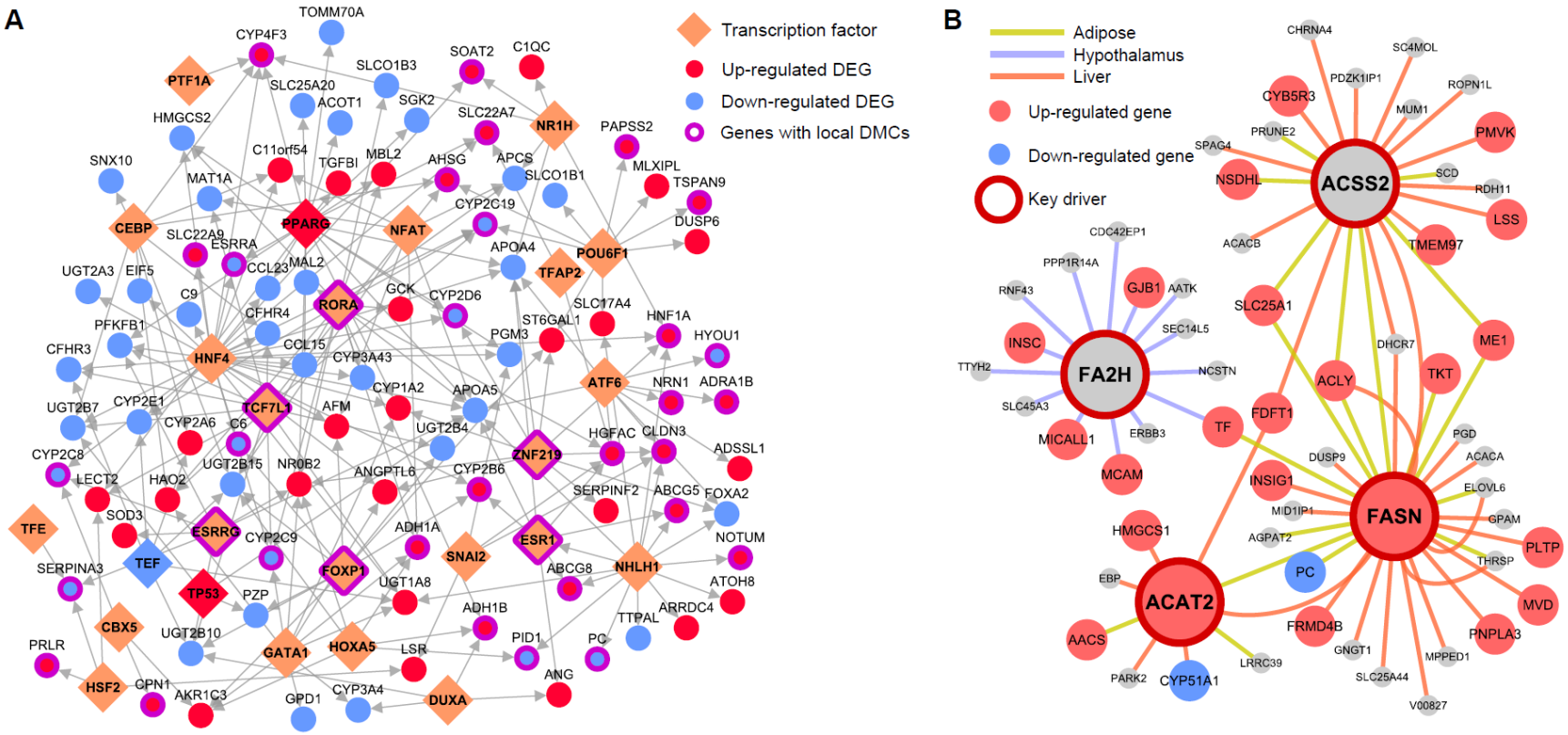


Figure 4.7 Transcription factors and key drivers orchestrate BPA induced gene expression level changes.

A) Liver transcription factor regulatory networks for the top ranked transcription factors (FDR < 5%) based on enrichment of liver DEGs among TF downstream targets. Network topology was based on FANTOM5. For TFs with > 20% overlapping downstream targets, only the TF with the lowest FDR is shown. **B**) Gene-gene regulatory subnetworks (Bayesian networks) for cross-tissue key drivers. Network topology was based on Bayesian network modeling of each tissue using genetic and transcriptome datasets from mouse and human populations. For each tissue, if >= 2 datasets were available for a given tissue, a network for each dataset was constructed and a consensus network was derived by keeping only the high confidence network edges between genes (edges appearing in >= 2 studies).

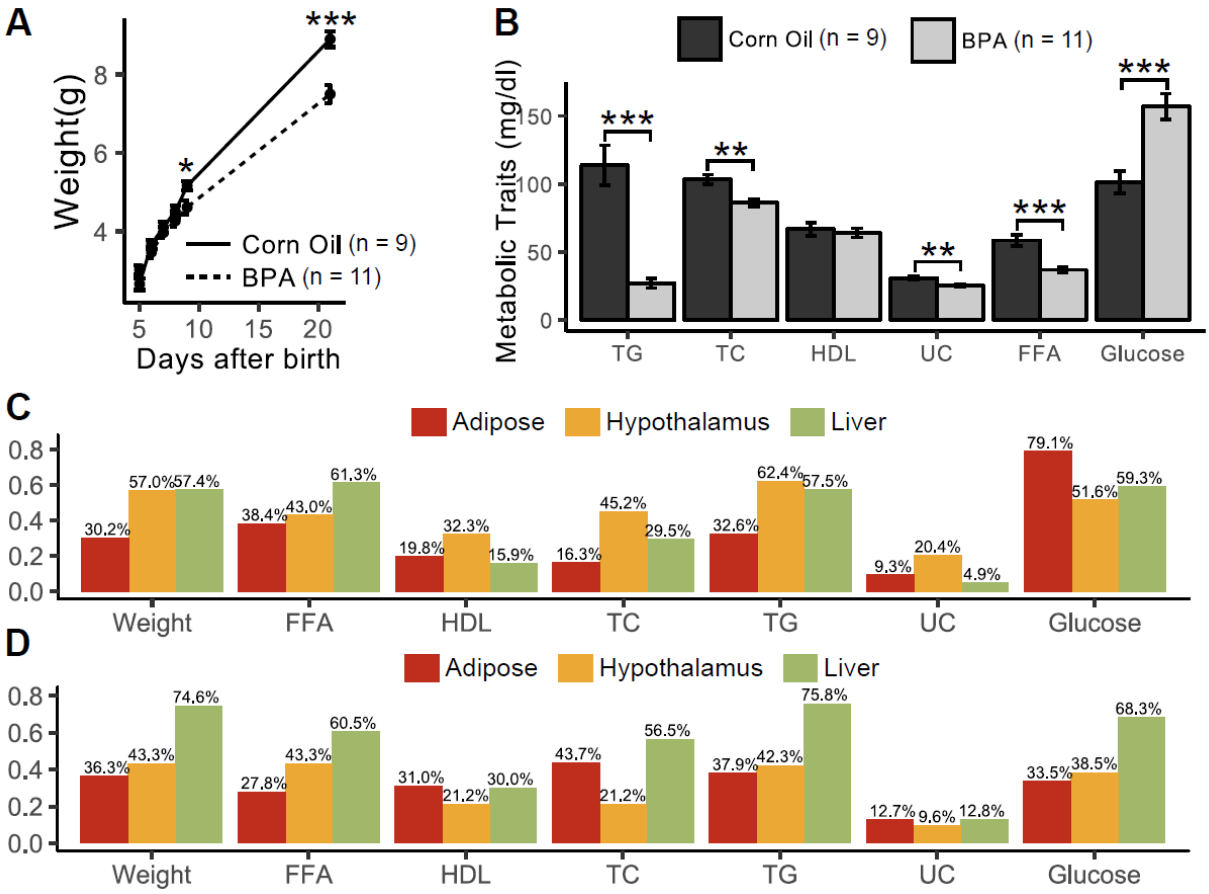


Figure 4.8 Measurements of metabolic traits in male offspring and the correlation between gene expression, methylation and metabolic traits.

A) Body weight for male offspring with prenatal exposure to corn oil (n = 9) or BPA (n = 11). **B**) Lipid and glucose profiles for male offspring with prenatal exposure to corn oil (n = 9) or BPA (n = 11) measured at 3 weeks of age. **C**) Percentage of tissue-specific DEGs that are correlated with metabolic traits ($p < 0.05$). **D**) Percentage of tissue-specific DMCs that are correlated with metabolic traits ($p < 0.05$). (A-B) p-values were determined using Student's t-test. (C-D) p-values were determined using Pearson correlation test.

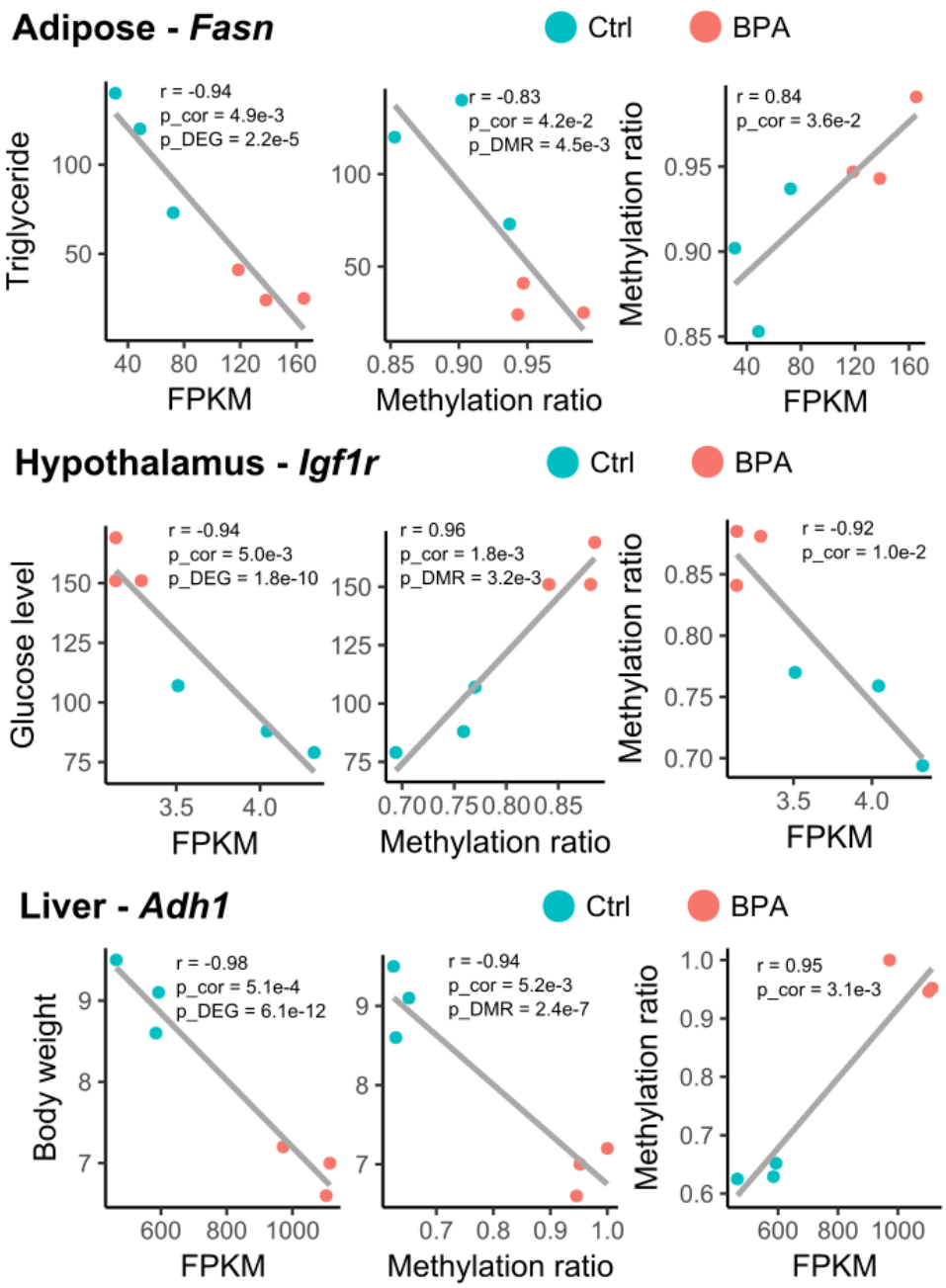


Figure 4.9 Pair-wise correlation between expression level, methylation ratio and metabolic profiles (triglyceride, glucose level, body weight) for *Fasn, Igf1r and Adh1*.

P_cor, p-value was determined using Pearson correction test; P_DEG was determined using differential expression test; P_DMC was determined using differential methylation test.

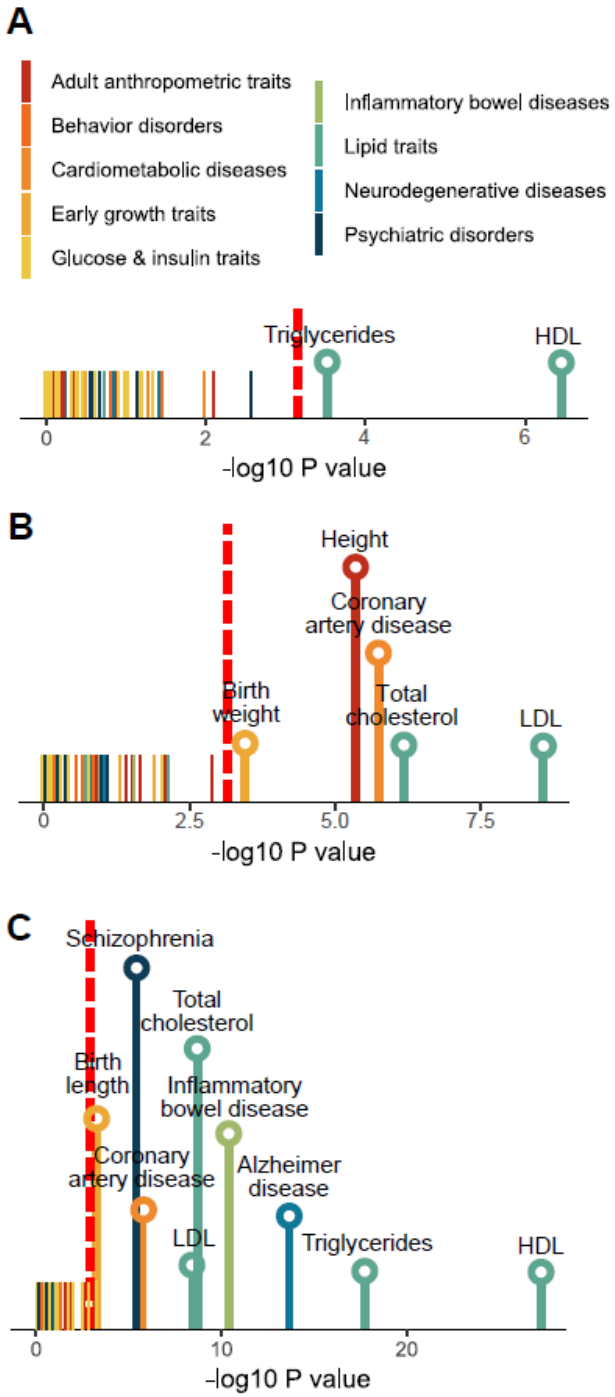


Figure 4.10 Association of differential expression signatures from adipose (**A**), hypothalamus (**B**) and liver (**C**) with 61 human traits/diseases, color coded into nine primary categories.

P-values are determined using MSEA. Red dashed line indicates the cutoff for Bonferroni-corrected p = 0.05. Names of traits/diseases whose p-values didn't pass Bonferroni-corrected cutoff were not shown.

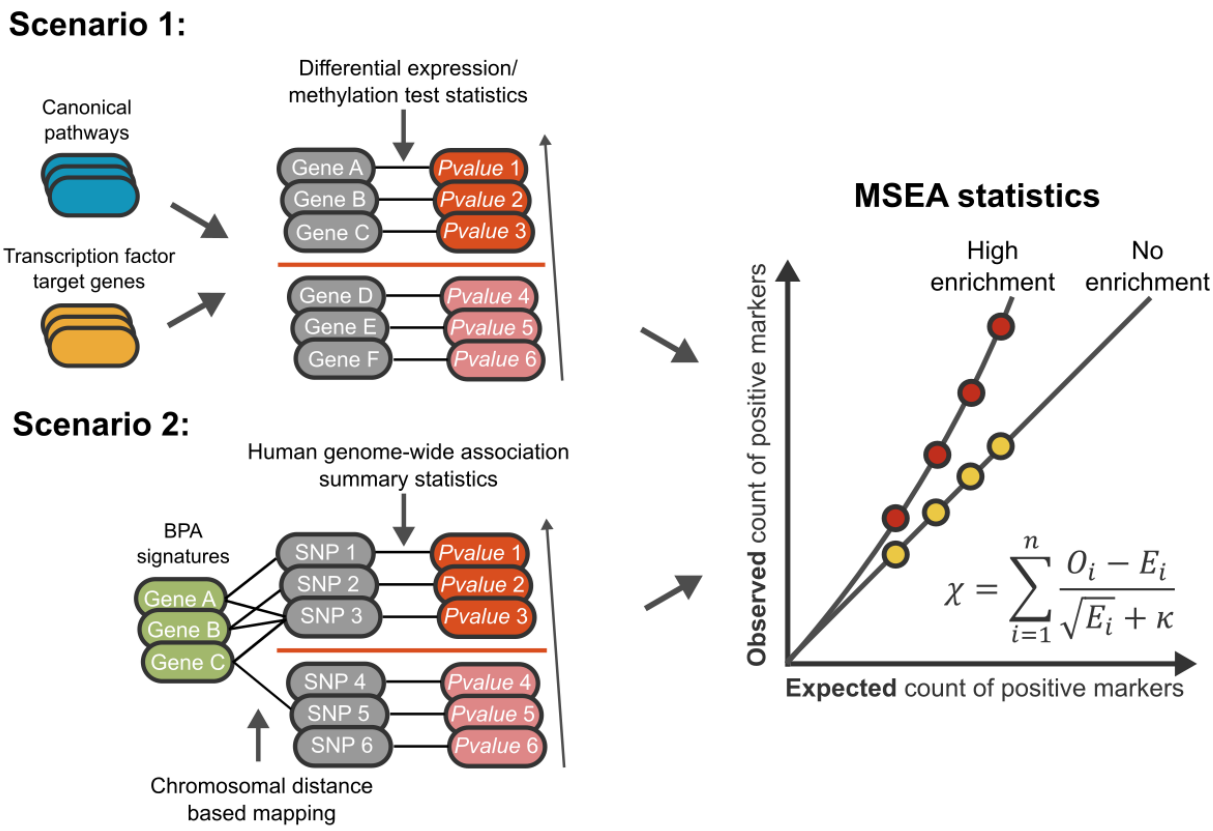


Figure 4.11 Schematic illustration of MSEA.

4.8 Appendix

Marker-set Enrichment Analysis (MSEA)

MSEA assesses whether a defined group of genes (members of canonical pathways, TF target genes, DEGs, etc.) shows enrichment of molecular markers (DEGs, DMCs, disease-associated SNPs, etc.) compared to random chance [75]. For the list of markers mapped to each gene-set, MSEA tested whether the marker list exhibited significant enrichment of markers with stronger biologically relevant signatures (differential expression, differential methylation, disease association, etc.) using a chi-square like statistic: $\chi = \sum_{i=1}^{n} \frac{O_i - E_i}{\sqrt{E_i} + \kappa}$, where n denotes the number of quantile points (we used ten quantile points ranging from the top 50% to the top 99.9% based on the rank of p-values), O and E denote the observed and expected counts of positive findings (i.e. signals above the quantile point), and $\kappa = 1$ is a stability parameter to reduce artefacts from low expected counts for small gene sets. The varying quantile thresholds allow the statistic to be adoptable to studies of varying sample size and statistical power. The null background was estimated by permuting gene labels to generate random gene sets matching the gene number of each gene set, while preserving the assignment of markers to genes. For each gene set, 10000 permuted gene sets were generated and enrichment p-values were determined from a Gaussian distribution approximated using the enrichment statistics from the 10000 permutations and the statistics of the gene sets. Finally, Benjami-Hochberg FDR was estimated across all modules tested for each GWAS.

In the study, MSEA was used in two analytical scenarios (**Figure 4.11**), 1) to assess the enrichment for differential expression/methylation signals among gene members of canonical pathways and TF downstream targets, and 2) to assess the enrichment for association signals from 61 human GWAS (**Table S4.2**) among tissue-specific BPA DEGs. In the second scenario, GWAS reported SNPs within the 50kb up/down-stream chromosomal distance of a gene were mapped to the corresponding gene.

FANTOM5 transcription factor network

We accessed the transcription factor networks for "adipose_tissue_adult", "brain_adult" and "liver_adult" from the FANTOM5 database [175]. The original connectivity of the FANTOM5 networks was too high, which violates the scale-free assumption of biological network topology and may lead to spurious results in subsequent analysis. Thus, we performed step-wise filtering by removing the networks edges with the lowest confidence score and recursively evaluated the scale-freeness of the resulting network. The removal process terminated as soon as the network reached boundary scale-freeness as defined by the criterion that at least 25% nodes has a degree (number of genes connected to the node) of 1. The downstream genes of each TF in the filtered network were pooled as the target genes for that TF.

Bayesian networks

BNs were constructed using a previously established method [61, 62]. BN are directed acyclic graphs with edges defined by conditional probabilities characterized by the distribution of states of each gene given the state of its parents. A joint probability distribution over all network genes can be defined based on network topology, and the likelihood of a BN network model can be

determined with given observed transcriptomic data (**Table S4.1**) using Bayes' formula [204]. For each data set, 1000 BN with different random seeds were reconstructed using Monte Carlo Markov Chain simulation, and the model with the best fit for each network was determined. In the resulting set of 1000 networks, edges appearing in over 30% of the networks were included in a consensus network. Causal direction was inferred between genes with genetic information as priors for the network [205]. For each tissue, the union of nodes and edges from BNs of multiple mouse and human studies were used as tissue-specific networks.

Weighted Key Driver Analysis (wKDA)

BPA induced DEGs were mapped onto constructed BNs to identify key drivers (KDs) using the weighted key driver analysis (wKDA) implemented in Mergeomics [75]. wKDA uniquely considers the edge weight information in the form of edge consistency score for BNs. Specifically, a network was first screened for suitable hub genes whose degree (number of genes connected to the hub) is in the top 25% of all network nodes. Once the hubs have been defined, their local one-edge neighborhoods, or "subnetworks" were extracted. All genes in each of the tissue-specific DEG sets were overlaid onto the hub subnetworks to see if a particular subnetwork was enriched for the genes from the DEG sets. The edges that connect a hub to its neighbors are simplified into node strengths (strength = sum of adjacent edge weights) within the neighborhood, except for the hub itself. The test statistic for the wKDA is analogous to the one used for MSEA: $\chi = \frac{O-E}{\sqrt{E}-\kappa}$, except that the values O and E represent the observed and expected ratios of genes from DEG sets in a hub subnetwork. In particular, $E = \frac{N_k N_p}{N}$ is estimated based on the hub degree N_k, disease gene set size N_p and the order of the full network N, with the implicit

assumption that the weight distribution is isotropic across the network. Statistical significance of the disease-enriched hubs, henceforth KDs, is estimated by permuting the gene labels in the network for 10000 times and estimating the p-value based on the null distribution. To control for multiple testing, stringent Bonferroni adjustment was used to focus on the top robust KDs.

Human GWAS summary statistics

We accessed and curated full summary statistics of 61 different human GWAS from diverse public data repositories (**Table S4.2**). For each GWAS, we removed SNPs with a minor allele frequency < 0.05. For SNPs that are in linkage disequilibrium with $r^2 > 0.5$, only the SNP with the strongest disease association was kept. The 61 GWAS traits/diseases were further categorized in nine categories, namely adult anthropometric traits, behavior disorders, cardiometabolic diseases, early growth traits, glucose and insulin traits, inflammatory bowel diseases, lipid traits, neurodegenerative diseases and psychiatric disorders.

Reference

1. Cowie, C.C., et al., *Full accounting of diabetes and pre-diabetes in the U.S. population in 1988-1994 and 2005-2006.* Diabetes Care, 2009. **32**(2): p. 287-94.

2. Blackwell, D.L., J.W. Lucas, and T.C. Clarke, *Summary health statistics for US adults: national health interview survey, 2012.* Vital and health statistics. Series 10, Data from the National Health Survey, 2014(260): p. 1-161.

3. Hunter, D.J., *Gene-environment interactions in human diseases.* Nat Rev Genet, 2005. **6**(4): p. 287-98.

4. Marbach, D., et al., *Wisdom of crowds for robust gene network inference.* Nat Methods, 2012. **9**(8): p. 796-804.

5. Huan, T., et al., *A systems biology framework identifies molecular underpinnings of coronary heart disease.* Arterioscler Thromb Vasc Biol, 2013. **33**(6): p. 1427-34.

6. Zhang, B., et al., *Integrated systems approach identifies genetic nodes and networks in late-onset Alzheimer's disease.* Cell, 2013. **153**(3): p. 707-20.

7. De Smet, R. and K. Marchal, *Advantages and limitations of current network inference methods.* Nat Rev Microbiol, 2010. **8**(10): p. 717-29.

8. Rhinn, H., et al., *Integrative genomics identifies APOE epsilon4 effectors in Alzheimer's disease.* Nature, 2013. **500**(7460): p. 45-50.

9. Schadt, E.E., et al., *An integrative genomics approach to infer causal associations between gene expression and disease.* Nat Genet, 2005. **37**(7): p. 710-7.

10. Tu, Z., et al., *Integrative analysis of a cross-loci regulation network identifies App as a gene regulating insulin secretion from pancreatic islets.* PLoS Genet, 2012. **8**(12): p. e1003107.

11. Wang, I.M., et al., *Systems analysis of eleven rodent disease models reveals an inflammatome signature and key drivers.* Mol Syst Biol, 2012. **8**: p. 594.

12. Yang, X., et al., *Validation of candidate causal genes for obesity that affect shared metabolic pathways and networks.* Nat Genet, 2009. **41**(4): p. 415-23.

13. Yang, X., et al., *Systematic genetic and genomic analysis of cytochrome P450 enzyme activities in human liver.* Genome Res, 2010. **20**(8): p. 1020-36.

14. Joyce, A.R. and B.Ø. Palsson, *The model organism as a system: integrating'omics' data sets.* Nature Reviews Molecular Cell Biology, 2006. **7**(3): p. 198-210.

15. Zhong, H., et al., *Integrating Pathway Analysis and Genetics of Gene Expression for Genome-wide Association Studies.* American Journal of Human Genetics, 2010. **86**(4): p. 581-591.

16. Yang, X., *Use of Functional Genomics to Identify Candidate Genes Underlying Human Genetic Association Studies of Vascular Diseases.* Arteriosclerosis Thrombosis and Vascular Biology, 2012. **32**(2): p. 216-222.

17. Civelek, M. and A.J. Lusis, *Systems genetics approaches to understand complex traits.* Nature Reviews Genetics, 2014. **15**(1): p. 34-48.

18. Makinen, V.P., et al., *Integrative genomics reveals novel molecular pathways and gene networks for coronary artery disease.* PLoS Genet, 2014. **10**(7): p. e1004502.

19. Kasarskis, A., X. Yang, and E. Schadt, *Integrative genomics strategies to elucidate the complexity of drug response.* Pharmacogenomics, 2011. **12**(12): p. 1695-715.

20. Yang, X., B. Zhang, and J. Zhu, *Functional genomics- and network-driven systems biology approaches for pharmacogenomics and toxicogenomics.* Curr Drug Metab, 2012. **13**(7): p. 952-67.

21. Schadt, E.E., S.H. Friend, and D.A. Shaywitz, *A network view of disease and compound screening.* Nat Rev Drug Discov, 2009. **8**(4): p. 286-95.

22. Hunter, D.J., *Gene-environment interactions in human diseases.* Nature Reviews Genetics, 2005. **6**(4): p. 287-98.

23. Mailman, M.D., et al., *The NCBI dbGaP database of genotypes and phenotypes.* Nat Genet, 2007. **39**(10): p. 1181-6.

24. Barrett, T. and R. Edgar, *Gene expression omnibus: microarray data storage, submission, retrieval, and analysis.* Methods Enzymol, 2006. **411**: p. 352-69.

25. Parkinson, H., et al., *ArrayExpress--a public database of microarray experiments and gene expression profiles.* Nucleic Acids Research, 2007. **35**(Database issue): p. D747-50.

26. Consortium, E.P., et al., *Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project.* Nature, 2007. **447**(7146): p. 799-816.

27. G. TEx Consortium, *The Genotype-Tissue Expression (GTEx) project.* Nat Genet, 2013. **45**(6): p. 580-5.

28. Kellis, M., et al., *Defining functional DNA elements in the human genome.* Proc Natl Acad Sci U S A, 2014. **111**(17): p. 6131-8.

29. Freedman, M.L., et al., *Principles for the post-GWAS functional characterization of cancer risk loci.* Nat Genet, 2011. **43**(6): p. 513-8.

30. Visscher, P.M., et al., *Five years of GWAS discovery.* American Journal of Human Genetics, 2012. **90**(1): p. 7-24.

31. Emilsson, V., et al., *Genetics of gene expression and its effect on disease.* Nature, 2008. **452**(7186): p. 423-8.

32. Schadt, E.E., et al., *An integrative genomics approach to infer causal associations between gene expression and disease.* Nat Genet, 2005. **37**(7): p. 710-717.

33. Boyle, A.P., et al., *Annotation of functional variation in personal genomes using RegulomeDB.* Genome Res, 2012. **22**(9): p. 1790-1797.

34. Schaub, M.A., et al., *Linking disease associations with regulatory information in the human genome.* Genome Res, 2012. **22**(9): p. 1748-59.

35. Segre, A.V., et al., *Common inherited variation in mitochondrial genes is not enriched for associations with type 2 diabetes or related glycemic traits.* PLoS Genet, 2010. **6**(8).

36. Zhang, K., et al., *i-GSEA4GWAS: a web server for identification of pathways/gene sets associated with traits by applying an improved gene set enrichment analysis to genome-wide association study.* Nucleic Acids Research, 2010. **38**(Web Server issue): p. W90-5.

37. Langfelder, P. and S. Horvath, *WGCNA: an R package for weighted correlation network analysis.* BMC Bioinformatics, 2008. **9**: p. 559.

38. Hiersche, M., F. Ruhle, and M. Stoll, *Postgwas: advanced GWAS interpretation in R.* PLoS One, 2013. **8**(8): p. e71775.

39. Wang, Q., et al., *EW_dmGWAS: edge-weighted dense module search for genome-wide association studies and gene expression profiles.* Bioinformatics, 2015. **31**(15): p. 2591-4.

40. Rossin, E.J., et al., *Proteins encoded in genomic regions associated with immune-mediated disease physically interact and suggest underlying biology.* PLoS Genet, 2011. **7**(1): p. e1001273.

41. Greene, C.S., et al., *Understanding multicellular function and disease with human tissue-specific networks.* Nat Genet, 2015. **47**(6): p. 569-76.

42. Wang, X., et al., *An R package suite for microarray meta-analysis in quality control, differentially expressed gene analysis and pathway enrichment detection.* Bioinformatics, 2012. **28**(19): p. 2534-2536.

43. Kettunen, J., et al., *Genome-wide association study identifies multiple loci influencing human serum metabolite levels.* Nat Genet, 2012. **44**(3): p. 269-76.

44. Ma, L., et al., *Genome-wide association analysis of total cholesterol and high-density lipoprotein cholesterol levels using the Framingham heart study data.* BMC medical genetics, 2010. **11**(1): p. 55.

45. Global Lipids Genetics Consortium, et al., *Discovery and refinement of loci associated with lipid levels.* Nat Genet, 2013. **45**(11): p. 1274-83.

46. Lango Allen, H., et al., *Hundreds of variants clustered in genomic loci and biological pathways affect human height.* Nature, 2010. **467**(7317): p. 832-8.

47. Subramanian, A., et al., *Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.* Proc Natl Acad Sci U S A, 2005. **102**(43): p. 15545-50.

48. Croft, D., et al., *The Reactome pathway knowledgebase.* Nucleic Acids Res, 2014. **42**(Database issue): p. D472-7.

49. Kanehisa, M. and S. Goto, *KEGG: kyoto encyclopedia of genes and genomes.* Nucleic Acids Res, 2000. **28**(1): p. 27-30.

50. Zhou, Z.H., et al., *Cidea-deficient mice have lean phenotype and are resistant to obesity.* Nat Genet, 2003. **35**(1): p. 49-56.

51. Eppig, J.T., et al., *The Mouse Genome Database (MGD): facilitating mouse as a model for human biology and disease.* Nucleic acids research, 2015. **43**(D1): p. D726-D736.

52. Koscielny, G., et al., *The International Mouse Phenotyping Consortium Web Portal, a unified point of access for knockout mice and related phenotyping data.* Nucleic Acids Res, 2014. **42**(Database issue): p. D802-9.

53. Dupuis, J., et al., *New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk.* Nat Genet, 2010. **42**(2): p. 105-16.

54. Bennett, B.J., et al., *A high-resolution association mapping panel for the dissection of complex traits in mice.* Genome Res, 2010. **20**(2): p. 281-90.

55. Orozco, L.D., et al., *Epigenome-wide association of liver methylation patterns and complex metabolic traits in mice.* Cell Metab, 2015. **21**(6): p. 905-17.

56. Canetti, L., H. Werner, and A. Leikin-Frenkel, *Linoleic and alpha linolenic acids ameliorate streptozotocin-induced diabetes in mice.* Arch Physiol Biochem, 2014. **120**(1): p. 34-9.

57. Hedges, L.V. and I. Olkin, *Statistical methods for meta-analysis*. 2014: Academic press.

58. Benjamini, Y. and Y. Hochberg, *Controlling the false discovery rate: a practical and powerful approach to multiple testing.* Journal of the Royal Statistical Society. Series B (Methodological), 1995: p. 289-300.

59. International HapMap Consortium, *The International HapMap Project.* Nature, 2003. **426**(6968): p. 789-96.

60. Genomes Project, C., et al., *An integrated map of genetic variation from 1,092 human genomes.* Nature, 2012. **491**(7422): p. 56-65.

61. Zhu, J., et al., *Increasing the power to detect causal associations by combining genotypic and expression data in segregating populations.* PLoS Comput Biol, 2007. **3**(4): p. e69.

62. Zhu, J., et al., *Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks.* Nat Genet, 2008. **40**(7): p. 854-61.

63. Huang, D.W., B.T. Sherman, and R.A. Lempicki, *Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources.* Nature Protocols, 2009. **4**(1): p. 44-57.

64. National Center for Health Statistics, *Health, United States, 2015: with special feature on racial and ethnic health disparities.* 2016.

65. Grundy, S.M., et al., *Diabetes and cardiovascular disease: a statement for healthcare professionals from the American Heart Association.* Circulation, 1999. **100**(10): p. 1134-46.

66. Beckman, J.A., M.A. Creager, and P. Libby, *Diabetes and atherosclerosis: epidemiology, pathophysiology, and management.* JAMA, 2002. **287**(19): p. 2570-81.

67. Wilson, P.W.F., et al., *Metabolic syndrome as a precursor of cardiovascular disease and type 2 diabetes mellitus.* Circulation, 2005. **112**(20): p. 3066-3072.

68. Mazzone, T., A. Chait, and J. Plutzky, *Cardiovascular disease risk in type 2 diabetes mellitus: insights from mechanistic studies.* Lancet, 2008. **371**(9626): p. 1800-9.

69. Welter, D., et al., *The NHGRI GWAS Catalog, a curated resource of SNP-trait associations.* Nucleic Acids Res, 2014. **42**(Database issue): p. D1001-6.

70. Chan, K.H., et al., *Shared molecular pathways and gene networks for cardiovascular disease and type 2 diabetes mellitus in women across diverse ethnicities.* Circ Cardiovasc Genet, 2014. **7**(6): p. 911-9.

71. Talukdar, H.A., et al., *Cross-Tissue Regulatory Gene Networks in Coronary Artery Disease.* Cell Syst, 2016. **2**(3): p. 196-208.

72. Franzen, O., et al., *Cardiometabolic risk loci share downstream cis- and trans-gene regulation across tissues and diseases.* Science, 2016. **353**(6301): p. 827-30.

73. von Scheidt, M., et al., *Applications and Limitations of Mouse Models for Understanding Human Atherosclerosis.* Cell Metab, 2016.

74. Zhao, Y., et al., *Network-Based Identification and Prioritization of Key Regulators of Coronary Artery Disease Loci.* Arterioscler Thromb Vasc Biol, 2016. **36**(5): p. 928-41.

75. Shu, L., et al., *Mergeomics: multidimensional data integration to identify pathogenic perturbations to biological systems.* BMC Genomics, 2016. **17**(1): p. 874.

76. Arneson, D., et al., *Mergeomics: a web server for identifying pathological pathways, networks, and key regulators via multidimensional data integration.* BMC Genomics, 2016. **17**(1): p. 722.

77. Mooney, M.A., et al., *Functional and genomic context in pathway analysis of GWAS data.* Trends Genet, 2014. **30**(9): p. 390-400.

78. Meigs, J.B., et al., *Genome-wide association with diabetes-related traits in the Framingham Heart Study.* BMC Med Genet, 2007. **8 Suppl 1**: p. S16.

79. Lettre, G., et al., *Genome-wide association study of coronary heart disease and its risk factors in 8,090 African Americans: the NHLBI CARe Project.* PLoS Genet, 2011. **7**(2): p. e1001300.

80. Nikpay, M., et al., *A comprehensive 1,000 Genomes-based genome-wide association meta-analysis of coronary artery disease.* Nat Genet, 2015. **47**(10): p. 1121-30.

81. Mahajan, A., et al., *Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility.* Nature Genetics, 2014. **46**(3): p. 234-+.

82. Parks, B.W., et al., *Genetic control of obesity and gut microbiota composition in response to high-fat, high-sucrose diet in mice.* Cell Metab, 2013. **17**(1): p. 141-52.

83. Bennett, B.J., et al., *Genetic Architecture of Atherosclerosis in Mice: A Systems Genetics Analysis of Common Inbred Strains.* PLoS Genet, 2015. **11**(12): p. e1005711.

84. Libby, P. and P. Theroux, *Pathophysiology of coronary artery disease.* Circulation, 2005. **111**(25): p. 3481-8.

85. Bornfeldt, K.E. and I. Tabas, *Insulin resistance, hyperglycemia, and atherosclerosis.* Cell Metab, 2011. **14**(5): p. 575-85.

86. Ceriello, A. and E. Motz, *Is oxidative stress the pathogenic mechanism underlying insulin resistance, diabetes, and cardiovascular disease? The common soil hypothesis revisited.* Arteriosclerosis, thrombosis, and vascular biology, 2004. **24**(5): p. 816-823.

87. Haffner, S.M., *The metabolic syndrome: inflammation, diabetes mellitus, and cardiovascular disease.* Am J Cardiol, 2006. **97**(2A): p. 3A-11A.

88. Lynch, C.J. and S.H. Adams, *Branched-chain amino acids in metabolic signalling and insulin resistance.* Nat Rev Endocrinol, 2014. **10**(12): p. 723-36.

89. Williams, A.S., L. Kang, and D.H. Wasserman, *The extracellular matrix and insulin resistance.* Trends Endocrinol Metab, 2015. **26**(7): p. 357-66.

90. Chistiakov, D.A., I.A. Sobenin, and A.N. Orekhov, *Vascular extracellular matrix in atherosclerosis.* Cardiol Rev, 2013. **21**(6): p. 270-88.

91. Pinero, J., et al., *DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes.* Database-the Journal of Biological Databases and Curation, 2015.

92. Asterholm, I.W., et al., *Altered mitochondrial function and metabolic inflexibility associated with loss of caveolin-1.* Cell Metab, 2012. **15**(2): p. 171-85.

93. Frank, P.G., et al., *Genetic ablation of caveolin-1 confers protection against atherosclerosis.* Arterioscler Thromb Vasc Biol, 2004. **24**(1): p. 98-105.

94. Razani, B., et al., *Caveolin-1-deficient mice are lean, resistant to diet-induced obesity, and show hypertriglyceridemia with adipocyte abnormalities.* J Biol Chem, 2002. **277**(10): p. 8635-47.

95. Lusis, A.J., et al., *The Hybrid Mouse Diversity Panel: a resource for systems genetics analyses of metabolic and cardiovascular traits.* J Lipid Res, 2016. **57**(6): p. 925-42.

96. von Scheidt, M., et al., *Applications and Limitations of Mouse Models for Understanding Human Atherosclerosis.* Cell Metab, 2017. **25**(2): p. 248-261.

97. Rau, C.D., A.J. Lusis, and Y. Wang, *Genetics of common forms of heart failure: challenges and potential solutions.* Curr Opin Cardiol, 2015. **30**(3): p. 222-7.

98. Parks, B.W., et al., *Genetic architecture of insulin resistance in the mouse.* Cell Metab, 2015. **21**(2): p. 334-46.

99. Manolio, T.A., et al., *Finding the missing heritability of complex diseases.* Nature, 2009. **461**(7265): p. 747-53.

100. Lee, C., et al., *Plasma Branched-Chain Amino Acids, Insulin Metabolism, and Incident Type 2 Diabetes-The Insulin Resistance Atherosclerosis Study (IRAS).* Diabetes, 2014. **63**: p. A382-A382.

101. Bhattacharya, S., et al., *Validation of the association between a branched chain amino acid metabolite profile and extremes of coronary artery disease in patients referred for cardiac catheterization.* Atherosclerosis, 2014. **232**(1): p. 191-196.

102. Lotta, L.A., et al., *Genetic Predisposition to an Impaired Metabolism of the Branched-Chain Amino Acids and Risk of Type 2 Diabetes: A Mendelian Randomisation Analysis.* PLoS Med, 2016. **13**(11): p. e1002179.

103. Jang, C., et al., *A branched-chain amino acid metabolite drives vascular fatty acid transport and causes insulin resistance.* Nat Med, 2016. **22**(4): p. 421-6.

104. Meng, Q., et al., *Systems Nutrigenomics Reveals Brain Gene Networks Linking Metabolic and Brain Disorders.* EBioMedicine, 2016.

105. Lek, M., et al., *Analysis of protein-coding genetic variation in 60,706 humans.* Nature, 2016. **536**(7616): p. 285-91.

106. Sattar, N., et al., *Statins and risk of incident diabetes: a collaborative meta-analysis of randomised statin trials.* Lancet, 2010. **375**(9716): p. 735-42.

107. Ference, B.A., et al., *Variation in PCSK9 and HMGCR and Risk of Cardiovascular Disease and Diabetes.* New England Journal of Medicine, 2016. **375**(22): p. 2144-2153.

108. Ridker, P.M., et al., *Cardiovascular benefits and diabetes risks of statin therapy in primary prevention: an analysis from the JUPITER trial.* Lancet, 2012. **380**(9841): p. 565-71.

109. Kusminski, C.M. and P.E. Scherer, *Mitochondrial dysfunction in white adipose tissue.* Trends Endocrinol Metab, 2012. **23**(9): p. 435-43.

110. Coletta, D.K. and L.J. Mandarino, *Mitochondrial dysfunction and insulin resistance from the outside in: extracellular matrix, the cytoskeleton, and mitochondria.* Am J Physiol Endocrinol Metab, 2011. **301**(5): p. E749-55.

111. Thorisson, G.A., et al., *The International HapMap Project Web site.* Genome Res, 2005. **15**(11): p. 1592-3.

112. 1000 Genomes Project Consortium, *An integrated map of genetic variation from 1,092 human genomes.* Nature, 2012. **491**(7422): p. 56-65.

113. Zhang, B. and S. Horvath, *A general framework for weighted gene co-expression network analysis.* Stat Appl Genet Mol Biol, 2005. **4**: p. Article17.

114. Nica, A.C., et al., *The architecture of gene regulatory variation across multiple human tissues: the MuTHER study.* PLoS Genet, 2011. **7**(2): p. e1002003.

115. Garnier, S., et al., *Genome-wide haplotype analysis of cis expression quantitative trait loci in monocytes.* PLoS Genet, 2013. **9**(1): p. e1003240.

116. Shannon, P., et al., *Cytoscape: a software environment for integrated models of biomolecular interaction networks.* Genome Res, 2003. **13**(11): p. 2498-504.

117. Boyle, A.P., et al., *Annotation of functional variation in personal genomes using RegulomeDB.* Genome Research, 2012. **22**(9): p. 1790-1797.

118. Liberzon, A., et al., *Molecular signatures database (MSigDB) 3.0.* Bioinformatics, 2011. **27**(12): p. 1739-40.

119. Pearl, J., *Probabilistic reasoning in intelligent systems : networks of plausible inference*. The Morgan Kaufmann series in representation and reasoning. 1988, San Mateo, Calif.: Morgan Kaufmann Publishers. xix, 552 p.

120. Madigan, D.a.Y., J., *Bayesian graphical models for discrete data.* International Statistical Review, 1995. **63**: p. 215-232.

121. Zhu, J., et al., *An integrative genomics approach to the reconstruction of gene networks in segregating populations.* Cytogenet Genome Res, 2004. **105**(2-4): p. 363-74.

122. Barouki, R., et al., *Developmental origins of non-communicable disease: implications for research and public health.* Environ Health, 2012. **11**: p. 42.

123. Boekelheide, K., et al., *Predicting later-life outcomes of early-life exposures.* Environ Health Perspect, 2012. **120**(10): p. 1353-61.

124. Heindel, J.J. and L.N. Vandenberg, *Developmental origins of health and disease: a paradigm for understanding disease etiology and prevention.* Current opinion in pediatrics, 2015. **27**(2): p. 248.

125. Vandenberg, L.N., et al., *Human exposure to bisphenol A (BPA).* Reprod Toxicol, 2007. **24**(2): p. 139-77.

126. Tsai, W.T., *Human health risk on environmental exposure to Bisphenol-A: a review.* J Environ Sci Health C Environ Carcinog Ecotoxicol Rev, 2006. **24**(2): p. 225-55.

127. Sun, C., et al., *Single laboratory validation of a method for the determination of Bisphenol A, Bisphenol A diglycidyl ether and its derivatives in canned foods by reversed-phase liquid chromatography.* J Chromatogr A, 2006. **1129**(1): p. 145-8.

128. Calafat, A.M., et al., *Exposure of the U.S. population to bisphenol A and 4-tertiary-octylphenol: 2003-2004.* Environ Health Perspect, 2008. **116**(1): p. 39-44.

129. Rubin, B.S., et al., *Perinatal BPA exposure alters body weight and composition in a dose specific and sex specific manner: The addition of peripubertal exposure exacerbates adverse effects in female mice.* Reprod Toxicol, 2017. **68**: p. 130-144.

130. Hao, M., et al., *Urinary bisphenol A concentration and the risk of central obesity in Chinese adults: A prospective study.* J Diabetes, 2017.

131. Beydoun, H.A., et al., *Sex differences in the association of urinary bisphenol-A concentration with selected indices of glucose homeostasis among U.S. adults.* Ann Epidemiol, 2014. **24**(2): p. 90-7.

132. Teppala, S., S. Madhavan, and A. Shankar, *Bisphenol A and Metabolic Syndrome: Results from NHANES.* Int J Endocrinol, 2012. **2012**: p. 598180.

133. Mouneimne, Y., et al., *Bisphenol A urinary level, its correlates, and association with cardiometabolic risks in Lebanese urban adults.* Environ Monit Assess, 2017. **189**(10): p. 517.

134. Wassenaar, P.N.H., L. Trasande, and J. Legler, *Systematic Review and Meta-Analysis of Early-Life Exposure to Bisphenol A and Obesity-Related Outcomes in Rodents.* Environ Health Perspect, 2017. **125**(10): p. 106001.

135. Han, C. and Y.C. Hong, *Bisphenol A, Hypertension, and Cardiovascular Diseases: Epidemiological, Laboratory, and Clinical Trial Evidence.* Curr Hypertens Rep, 2016. **18**(2): p. 11.

136. Ranciere, F., et al., *Bisphenol A and the risk of cardiometabolic disorders: a systematic review with meta-analysis of the epidemiological evidence.* Environ Health, 2015. **14**: p. 46.

137. Liu, J., et al., *Perinatal bisphenol A exposure and adult glucose homeostasis: identifying critical windows of exposure.* PLoS One, 2013. **8**(5): p. e64143.

138. Ryan, K.K., et al., *Perinatal exposure to bisphenol-a and the development of metabolic syndrome in CD-1 mice.* Endocrinology, 2010. **151**(6): p. 2603-12.

139. Miyawaki, J., et al., *Perinatal and postnatal exposure to bisphenol a increases adipose tissue mass and serum cholesterol level in mice.* J Atheroscler Thromb, 2007. **14**(5): p. 245-52.

140. Rubin, B.S. and A.M. Soto, *Bisphenol A: Perinatal exposure and body weight.* Mol Cell Endocrinol, 2009. **304**(1-2): p. 55-62.

141. Garcia-Arevalo, M., et al., *Exposure to bisphenol-A during pregnancy partially mimics the effects of a high-fat diet altering glucose homeostasis and gene expression in adult male mice.* PLoS One, 2014. **9**(6): p. e100214.

142. Manikkam, M., et al., *Plastics derived endocrine disruptors (BPA, DEHP and DBP) induce epigenetic transgenerational inheritance of obesity, reproductive disease and sperm epimutations.* PLoS One, 2013. **8**(1): p. e55387.

143. Susiarjo, M., et al., *Bisphenol a exposure disrupts metabolic health across multiple generations in the mouse.* Endocrinology, 2015. **156**(6): p. 2049-58.

144. Bansal, A., et al., *Sex- and Dose-Specific Effects of Maternal Bisphenol A Exposure on Pancreatic Islets of First- and Second-Generation Adult Mice Offspring.* Environ Health Perspect, 2017. **125**(9): p. 097022.

145. Baillie-Hamilton, P.F., *Chemical toxins: a hypothesis to explain the global obesity epidemic.* J Altern Complement Med, 2002. **8**(2): p. 185-92.

146. Heindel, J.J., *Endocrine disruptors and the obesity epidemic.* Toxicol Sci, 2003. **76**(2): p. 247-9.

147. Newbold, R.R., et al., *Effects of endocrine disruptors on obesity.* Int J Androl, 2008. **31**(2): p. 201-8.

148. EFS Authority, *Scientific opinion on the risks to public health related to the presence of bisphenol A (BPA) in foodstuffs.* EFSA Journal, 2015. **13**(1).

149. Beronius, A., et al., *The influence of study design and sex-differences on results from developmental neurotoxicity studies of bisphenol A: implications for toxicity testing.* Toxicology, 2013. **311**(1-2): p. 13-26.

150. Ariemma, F., et al., *Low-Dose Bisphenol-A Impairs Adipogenesis and Generates Dysfunctional 3T3-L1 Adipocytes.* PLoS One, 2016. **11**(3): p. e0150762.

151. Ben-Jonathan, N., E.R. Hugo, and T.D. Brandebourg, *Effects of bisphenol A on adipokine release from human adipose tissue: Implications for the metabolic syndrome.* Mol Cell Endocrinol, 2009. **304**(1-2): p. 49-54.

152. Olsvik, P.A., K.H. Skjaerven, and L. Softeland, *Metabolic signatures of bisphenol A and genistein in Atlantic salmon liver cells.* Chemosphere, 2017. **189**: p. 730-743.

153. Lejonklou, M.H., et al., *Effects of Low-Dose Developmental Bisphenol A Exposure on Metabolic Parameters and Gene Expression in Male and Female Fischer 344 Rat Offspring.* Environ Health Perspect, 2017. **125**(6): p. 067018.

154. Anderson, O.S., et al., *Novel Epigenetic Biomarkers Mediating Bisphenol A Exposure and Metabolic Phenotypes in Female Mice.* Endocrinology, 2017. **158**(1): p. 31-40.

155. Ma, Y., et al., *Hepatic DNA methylation modifications in early development of rats resulting from perinatal BPA exposure contribute to insulin resistance in adulthood.* Diabetologia, 2013. **56**(9): p. 2059-67.

156. Taylor, J.A., et al., *Prenatal Exposure to Bisphenol A Disrupts Naturally Occurring Bimodal DNA Methylation at Proximal Promoter of fggy, an Obesity-relevant Gene Encoding a Carbohydrate Kinase, in Gonadal White Adipose Tissues of CD-1 Mice.* Endocrinology, 2017.

157. Faulk, C., et al., *Bisphenol A-associated alterations in genome-wide DNA methylation and gene expression patterns reveal sequence-dependent and non-monotonic effects in human fetal liver.* Environ Epigenet, 2015. **1**(1).

158. Nahar, M.S., et al., *Bisphenol A-associated alterations in the expression and epigenetic regulation of genes encoding xenobiotic metabolizing enzymes in human fetal liver.* Environ Mol Mutagen, 2014. **55**(3): p. 184-95.

159. Messerlian, C., et al., *'Omics' and endocrine-disrupting chemicals—new paths forward.* Nature Reviews Endocrinology, 2017. **13**(12): p. 740.

160. Meng, Q., et al., *Systems Nutrigenomics Reveals Brain Gene Networks Linking Metabolic and Brain Disorders.* EBioMedicine, 2016. **7**: p. 157-66.

161. Shu, L., et al., *Shared genetic regulatory networks for cardiovascular disease and type 2 diabetes in multiple populations of diverse ethnicities in the United States.* PLoS Genet, 2017. **13**(9): p. e1007040.

162. Krishnan, K.C., et al., *Integration of Multi-omics Data from Mouse Diversity Panel Highlights Mitochondrial Dysfunction in Non-alcoholic Fatty Liver Disease.* Cell Systems, 2018.

163. Dolinoy, D.C., D. Huang, and R.L. Jirtle, *Maternal nutrient supplementation counteracts bisphenol A-induced DNA hypomethylation in early development.* Proc Natl Acad Sci U S A, 2007. **104**(32): p. 13056-61.

164. Susiarjo, M., et al., *Bisphenol a exposure disrupts genomic imprinting in the mouse.* PLoS Genet, 2013. **9**(4): p. e1003401.

165. Bromer, J.G., et al., *Bisphenol-A exposure in utero leads to epigenetic alterations in the developmental programming of uterine estrogen response.* FASEB J, 2010. **24**(7): p. 2273-80.

166. Lewinska, M., et al., *Hidden disease susceptibility and sexual dimorphism in the heterozygous knockout of Cyp51 from cholesterol synthesis.* PLoS One, 2014. **9**(11): p. e112787.

167. Keber, R., D. Rozman, and S. Horvat, *Sterols in spermatogenesis and sperm maturation.* J Lipid Res, 2013. **54**(1): p. 20-33.

168. Lou, S., et al., *Whole-genome bisulfite sequencing of multiple individuals reveals complementary roles of promoter and gene body methylation in transcriptional regulation.* Genome Biol, 2014. **15**(7): p. 408.

169. Faulk, C., et al., *Detection of differential DNA methylation in repetitive DNA of mice and humans perinatally exposed to bisphenol A.* Epigenetics, 2016. **11**(7): p. 489-500.

170. Acconcia, F., V. Pallottini, and M. Marino, *Molecular Mechanisms of Action of BPA.* Dose Response, 2015. **13**(4): p. 1559325815610582.

171. MacKay, H. and A. Abizaid, *A plurality of molecular targets: The receptor ecosystem for bisphenol-A (BPA).* Horm Behav, 2017.

172. Wang, J., et al., *The environmental obesogen bisphenol A promotes adipogenesis by increasing the amount of 11beta-hydroxysteroid dehydrogenase type 1 in the adipose tissue of children.* Int J Obes (Lond), 2013. **37**(7): p. 999-1005.

173. Ahmed, S. and E. Atlas, *Bisphenol S- and bisphenol A-induced adipogenesis of murine preadipocytes occurs through direct peroxisome proliferator-activated receptor gamma activation.* Int J Obes (Lond), 2016. **40**(10): p. 1566-1573.

174. Vafeiadi, M., et al., *Association of early life exposure to bisphenol A with obesity and cardiometabolic traits in childhood.* Environ Res, 2016. **146**: p. 379-87.

175. Marbach, D., et al., *Tissue-specific regulatory circuits reveal variable modular perturbations across complex diseases.* Nat Methods, 2016. **13**(4): p. 366-70.

176. Marmugi, A., et al., *Low doses of bisphenol A induce gene expression related to lipid synthesis and trigger triglyceride accumulation in adult mouse liver.* Hepatology, 2012. **55**(2): p. 395-407.

177. Morrison, J.L., et al., *Fetal growth restriction, catch-up growth and the early origins of insulin resistance and visceral obesity.* Pediatr Nephrol, 2010. **25**(4): p. 669-77.

178. Inadera, H., *Neurological Effects of Bisphenol A and its Analogues.* Int J Med Sci, 2015. **12**(12): p. 926-36.

179. Kim, J.H., et al., *Perinatal bisphenol A exposure promotes dose-dependent alterations of the mouse methylome.* BMC Genomics, 2014. **15**: p. 30.

180. Ilagan, Y., et al., *Bisphenol-A exposure in utero programs a sexually dimorphic estrogenic state of hepatic metabolic gene expression.* Reprod Toxicol, 2017. **71**: p. 84-94.

181. Ke, Z., et al., *Bisphenol A exposure may induce hepatic lipid accumulation via reprogramming the DNA methylation patterns of genes involved in lipid metabolism.* Scientific reports, 2016. **6**.

182. Yang, S., et al., *Dysregulated Autophagy in Hepatocytes Promotes Bisphenol A-Induced Hepatic Lipid Accumulation in Male Mice.* Endocrinology, 2017. **158**(9): p. 2799-2812.

183. Shimpi, P.C., et al., *Hepatic Lipid Accumulation and Nrf2 Expression following Perinatal and Peripubertal Exposure to Bisphenol A in a Mouse Model of Nonalcoholic Liver Disease.* Environ Health Perspect, 2017. **125**(8): p. 087005.

184. Karlsson, O. and A.A. Baccarelli, *Environmental Health and Long Non-coding RNAs.* Curr Environ Health Rep, 2016. **3**(3): p. 178-87.

185. Somm, E., et al., *Perinatal exposure to bisphenol a alters early adipogenesis in the rat.* Environ Health Perspect, 2009. **117**(10): p. 1549-55.

186. Vom Saal, F.S., et al., *The estrogenic endocrine disrupting chemical bisphenol A (BPA) and obesity.* Mol Cell Endocrinol, 2012. **354**(1-2): p. 74-84.

187. Angle, B.M., et al., *Metabolic disruption in male mice due to fetal exposure to low but not high doses of bisphenol A (BPA): evidence for effects on body weight, food intake, adipocytes, leptin, adiponectin, insulin and glucose regulation.* Reprod Toxicol, 2013. **42**: p. 256-68.

188. Vercruysse, P., et al., *Hypothalamic alterations in neurodegenerative diseases and their relation to abnormal energy metabolism.* Frontiers in Molecular Neuroscience, 2018. **11**: p. 2.

189. MacKay, H., Z.R. Patterson, and A. Abizaid, *Perinatal Exposure to Low-Dose Bisphenol-A Disrupts the Structural and Functional Development of the Hypothalamic Feeding Circuitry.* Endocrinology, 2017. **158**(4): p. 768-777.

190. Jones, P.A., *Functions of DNA methylation: islands, start sites, gene bodies and beyond.* Nat Rev Genet, 2012. **13**(7): p. 484-92.

191. Patil, V., R.L. Ward, and L.B. Hesson, *The evidence for functional non-CpG methylation in mammalian cells.* Epigenetics, 2014. **9**(6): p. 823-8.

192. Gracia, A., et al., *Fatty acid synthase methylation levels in adipose tissue: effects of an obesogenic diet and phenol compounds.* Genes Nutr, 2014. **9**(4): p. 411.

193. Hui, S.T., et al., *The genetic architecture of NAFLD among inbred strains of mice.* Elife, 2015. **4**.

194. Dallio, M., et al., *Role of bisphenol A as environmental factor in the promotion of non‐alcoholic fatty liver disease: in vitro and clinical study.* Alimentary pharmacology & therapeutics, 2018.

195. Veiga-Lopez, A., et al., *Gender-Specific Effects on Gestational Length and Birth Weight by Early Pregnancy BPA Exposure.* J Clin Endocrinol Metab, 2015. **100**(11): p. E1394-403.

196. Andrews, S., *FastQC: a quality control tool for high throughput sequence data*. 2010.

197. Pertea, M., et al., *Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown.* Nat Protoc, 2016. **11**(9): p. 1650-67.

198. Love, M.I., W. Huber, and S. Anders, *Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.* Genome Biol, 2014. **15**(12): p. 550.

199. Storey, J.D. and R. Tibshirani, *Statistical significance for genomewide studies.* Proc Natl Acad Sci U S A, 2003. **100**(16): p. 9440-5.

200. Xi, Y. and W. Li, *BSMAP: whole genome bisulfite sequence MAPping program.* BMC Bioinformatics, 2009. **10**: p. 232.

201. Sun, D., et al., *MOABS: model based analysis of bisulfite sequencing data.* Genome Biol, 2014. **15**(2): p. R38.

202. Cavalcante, R.G. and M.A. Sartor, *annotatr: genomic regions in context.* Bioinformatics, 2017. **33**(15): p. 2381-2383.

203. Meng, Q., et al., *Traumatic Brain Injury Induces Genome-Wide Transcriptomic, Methylomic, and Network Perturbations in Brain and Blood Predicting Neurological Disorders.* EBioMedicine, 2017. **16**: p. 184-194.

204. Madigan, D. and J. York, *Bayesian Graphical Models for Discrete-Data.* International Statistical Review, 1995. **63**(2): p. 215-232.

205. Zhu, J., et al., *An integrative genomics approach to the reconstruction of gene networks in segregating populations.* Cytogenetic and Genome Research, 2004. **105**(2-4): p. 363-74.