Title
Uncovering the Molecular Networks of Metabolic Diseases Using Systems Biology

Author
Blencowe, Montgomery

Publication Date
2022

Peer reviewed|Thesis/dissertation

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

Montgomery Charles Thomas Blencowe

2022

ABSTRACT OF THE DISSERTATION

Utilizing a systems biology approach to
uncover the molecular networks of
metabolic diseases

by

Montgomery Charles Thomas Blencowe
Doctor of Philosophy in Molecular, Cellular, and Integrative
Physiology
University of California, Los Angeles, 2022
Professor Xia Yang, Chair

Common complex metabolic diseases (MetDs) such as obesity, type 2 diabetes (T2D), coronary artery disease (CAD) and non-alcoholic fatty liver disease (NAFLD), impose an unprecedented burden on public health worldwide and demonstrate sex differences. Our general hypothesis is that genetic risk factors perturb set of genes in the form of functional gene networks, which subsequently induces the initiation and progression of MetDs. Following this hypothesis, our research focuses on dissecting the molecular networks that are perturbed by genetic risk factors of MetDs utilizing multiomics systems biology approaches. To address this challenge, I embarked interdisciplinary systems biology studies encompassing the development of an accessible multi-omics integration webserver, elucidation of genetically perturbed tissue networks in numerous MetDs, and uncovering the relative contribution of three sex factors in gene regulation in tissues relevant to MetDs. First, I contributed to the development of a user-friendly webserver for

multiomics integration, network modeling, and network-based drug repositioning for complex diseases such as MetDs. Second, I investigated the genetically perturbed gene networks that underly various MetDs, namely, lipid traits, diabetes, CAD, and NAFLD. Third, I employed systems biology approaches to uncover the individual and interactive contribution of three sex factors (sex chromosomes, gonads, and sex hormones) in gene regulation in tissues relevant to MetDs. Completion of these projects offer a user-friendly bioinformatic tool, molecular insights, and drug candidates for MetDs.

The dissertation of Montgomery Charles Thomas Blencowe is approved.

Aldons J Lusis

Matteo Pellegrini

Arthur P. Arnold

Xia Yang, Committee Chair

University of California, Los Angeles

2022

DEDICATION

This dissertation is dedicated to my family.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENT

My dissertation work would not be possible without the guidance and support from my advisor, Dr. Xia Yang, whose research area inspired me in the first place to pursue a PhD. She spared no effort in lighting the path of my doctoral training and future career choices. I could not have asked for a better PhD experience. I also would like to thank my committee for all their expert advice and guidance throughout my studies, I am particularly grateful to all the wonderful collaborations that we have been able to pursue together. I am also grateful for all the support received from the lab members, fellow researchers, and students at UCLA, with whom I have worked closely. Lastly, I would like to express my gratitude to my family who have always supported throughout all my academic endeavors.

Chapter 2 is a version of Ding J[+], Blencowe M[+], Nghiem TX[+], Ha SM, Chen Y-W, Li G, Yang X. "Mergeomics 2.0: A Web Server for Multi-omics Data Integration to Elucidate Disease Networks and Predict Therapeutics." *Nucleic Acids Research* 49: W375-387, 2021. The work was supported NIH R01 grants NS117148, NS111378, DK117850, HL145708, HL147883, HD100298 to XY; American Heart Association Predoctoral Fellowship (839009) to MB; UCLA QCBio Collaboratory Postdoc Fellowship to SMH.

Chapter 3 is a version of Blencowe M[+], Ahn IS[+], Saleem Z, Luk H, Cely I, Makinen VP, Zhao Y, Yang X. "Gene networks and pathways for plasma lipid traits via multi-tissue multi-omics systems analysis." *Journal of Lipid Research*, 62: 100019, 2021. The study was supported by NIH grants R01 DK104363 and R01 DK117850.

Chapter 4 is a version of Blencowe M, Furterer A, Wang Q, Gao F, Rosenberger M, Pei L, Nomoto H, Mawla AM, Huising MO, Coppola G, Yang X, Butler PC, Gurlo T. "IAPP-induced beta cell stress recapitulates the islet transcriptome in type 2 diabetes" Diabetologia, 65, 173-187, 2022. These studies were supported by the United States Public Health Services National Institute of Health grant (DK059579) and the Larry L. Hillblom Foundation (2014-D-001-NET). The study sponsor was not involved in the design of the study; the collection, analysis, and interpretation of data; writing the report; and did not impose any restrictions regarding the publication of the report. The work in Chapter 6 is supported by the American Heart Association Predoctoral Fellowship (829009), UCLA IBP Edith Hyde Fellowship and National Institutes of Health (R01DK117850)

Chapter 5 is supported by the American Heart Association Predoctoral Fellowship (829009), UCLA IBP Edith Hyde Fellowship and National Institutes of Health (R01DK117850)

Chapter 6 is a version of Blencowe M, Chen X, Zhao Y, Itoh Y, McQuillen CN, Han Y, Shou BL, McClusky R, Reue K, Arnold AP, Yang X. "Relative contributions of sex hormones, sex chromosomes, and gonads to sex differences in tissue gene regulation." *Genome Research,* Apr 8, 2022. The study is partially supported by NIH grants NS043196, HD076125, and HD100298 (AA). NIH grants DK104363, DK117850, and HD100298 (XY). MB is supported by the American Heart Association Predoctoral Fellowship (829009) and UCLA IBP Edith Hyde Fellowship.

VITA

EDUCATION

2022	PhD Candidate, Molecular, Cellular, and Integrative Physiology, University of California, Los Angeles
2019	Master of Science, Physiological Science, University of California, Los Angeles
2015	Bachelor of Science, Biomedical Science, King's College London, United Kingdom

RESEARCH EXPERIENCE

2019 – Present	PhD trainee Department of Integrative Biology and Physiology University of California, Los Angeles
2017- 2019	Graduate Student Researcher Department of Integrative Biology and Physiology University of California, Los Angeles

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

Blencowe M, Chen X, Zhao Y, Itoh Y, McQuillen C, Han Y, Shou B, McClusky R, Reue K, Arnold AP, Yang X. Relative contributions of sex hormones, sex chromosomes, and gonads to sex differences in tissue gene regulation. *Genome Research,* May;32(5):807-824. **2022**

Chen YW, Diamanate G, Ding J, Nghiem T, Yang J, Ha S, Cohn P, Arneson D, **Blencowe M**, Garcia J, Zaghari N, Patel P, Yang X. PharmOmics: A Species- and Tissue Specific Drug Signature Database and Online Tool for Drug Repurposing. *iScience,* Apr;25(4):104052. **2022**

Blencowe M, Furterer A, Wang Q, Gao F, Rosenberger M, Pei L, Nomoto H, Mawla A, Huising MO, Coppola G, Yang X, Butler PC, Gurlo T. IAPP induced beta cell stress recapitulates the islet transcriptome in type 2 diabetes. *Diabetologia*, Jan;65(1):173-187, **2022**

Ding J*, **Blencowe M***, Nghiem T*, Ha S, Chen YW, Li G, Yang X. Mergeomics 2.0: A Web Server for Multiomics Data Integration to Uncover Disease Networks and Therapeutics. *Nucleic Acids Research, 49: W375-387,* **2021**

Blencowe M*, Ahn IS*, Saleem Z, Luk H, Cely I, Makinen VP, Zhao Y, Yang X. Genes and Pathways for Plasma Lipid Traits via Multi-tissue Multi-omics Systems Analysis. *Journal of Lipid Research, 62, 100019,* **2021**

Diamante G, Cely I, Zamora Z, Ding J, **Blencowe M**, Lang J, Bline A, Singh M, Lusis AJ, Yang X. Systems toxicogenomics of prenatal low-dose BPA exposure on liver metabolic pathways, gut microbiota, and metabolic health in mice. *Environment International*, 146, 106260, **2021**

Zhang G, Meng Q, **Blencowe M**, Agrawal R, Gomez-Pinilla F, Yang X. Multi-tissue Multi-Omics

Nutrigenomics indicates context-specific effects of docosahexaenoic acid on rat brain. *Molecular Nutrition & Food Research,* 64 (23), 2000788, **2020**

Blencowe M, Karunanayake T, Wier J, Hsu N, Yang X. Network Modeling Approaches and Applications to Unravelling Non-Alcoholic Fatty Liver Disease. *Genes*, *10*, 966, **2019**

Zhang G, Byun HR, Ying Z, **Blencowe M**, Zhao Y, Hong J, Shu L, Gomez-Pinilla F*, Yang X*. Differential Metabolic and Multi-tissue Transcriptomic Responses to Fructose Consumption among Genetically Diverse Mice. *BBA – Molecular Basis of Disease,* 1866 (1), 165569, **2019**

Blencowe M, Arneson D, Ding J, Chen YW, Saleem Z, Yang X. Network modeling of single-cell omics data: challenges, opportunities, and progresses. *Emerging Topics in Life Sciences*, 16;3(4):379-98, **2019**

Zhao Y, **Blencowe M**, Shi X, Shu L, Levian C, Ahn IS, ICBP consortium, Kim SK, Huan T, Levy D, Yang X. Integrative Genomics Analysis Unravels Tissue-Specific Pathways, Networks, and Key Regulators of Blood Pressure Regulation. *Frontiers in Cardiovascular Medicine,* 6:21, **2019**.

Martin LJ, Meng Q, **Blencowe M**, Lagarrigue S, Xiao S, Pan C, Wier J, Temple WC, Devaskar SU, Lusis AJ, Yang X. Maternal High-protein and Low-protein Diets Perturb Hypothalamus and Liver Transcriptome and Metabolic Homeostasis in Adult Mouse Offspring. *Frontiers in Genetics* 9:642, **2018**

Shu L, **Blencowe M**, Yang X. Translating GWAS Findings to Novel Therapeutic Targets for Coronary Artery Disease. *Frontiers in Cardiovascular Medicine,* 5:56, **2018**.

INVITED PREVIEW

Blencowe M, Yang X. Found in Translation – core network preservation across liver diseases and species. *Cell Reports Medicine, 2:7,* **2021**

SELECTED CONFERENCE PRESENTATIONS

Blencowe M, Feng, Hsu N, Yang X. Multiomics Network Analysis of Three Liver Diseases using UK Biobank data. **Oral Presentation** at the Systems Biology of Human Disease Conference (Berlin) July 2021.

Blencowe M, Norheim F, Saleem Z, Hsu N, Hui S, Krishnan KC, Edilor C, Yang X, Lusis AJ. 1173-P: Sex differences in NASH pathways informed by multi-omics. **Poster Presentation** at the American Diabetes Association Session June 2021

Blencowe M, Saleem Z, Wier J, Karunanayake T, Yang X. 1698-P: Integrative Genomics Analysis Reveals Tissue-Specific Pathways and Gene Networks for Type 1 Diabetes. **Poster Presentation** at the American Diabetes Association Session (San Francisco) June 2019.

AWARDS

American Heart Association Predoctoral Fellowship (2021 – 2023)

Integrative Biology & Physiology Edith Hyde Fellowship (2021 – 2022)

Chapter 1. Introduction

Metabolic diseases (MetDs) are an ever-increasing burden on the general population including diseases such as NAFLD, CAD, obesity and diabetes. The 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], CAD patients to be as high as 6.7% of the population [2] and NAFLD patients to be over 25% of the population [3]. Although the exact etiology of MetDs remains elusive, evidence supports the importance of genetics in disease development, progression and heritability [4], e.g. CAD and NAFLD have heritability estimates of 50-60% [5] and 20-70% [6], respectively. Particularly with the large number of GWAS studies available we have been successful in elucidating novel disease variants. However, with highly complex diseases such as MetDs, standard approaches such as GWAS alone have limited capacity to fully dissect the complexity. Integrating other omics areas, which highlights other critical information such as EWAS (epigenetic contribution), PWAS (protein contribution), TWAS (transcript contribution), or MWAS (metabolome contribution), will help provide a more complete image. Thus, using a systems biology approach becomes vital to understand the molecular mechanisms underlying the actions of genetics for which we can target to elucidate effective therapeutic and preventive strategies.

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 [7-9]. Previous studies by us and others 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 [9-15]. Therefore, rather than focusing on the investigation of isolated genes, it is more appealing to elucidate the gene-gene interactions to dissect the molecular mechanisms

of 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 [16]. 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 i) sufficient power to adequately explain heritability and gene-by-environment interaction and ii) mechanistic connection between the risk loci and downstream events. By integrating different layers of biological information, we 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 processes and gene interactions underlying complex diseases [17-20]. Nevertheless, well-defined high-throughput systems biology tools that effectively convert large-scale biological data and network resources into meaningful outputs are still lacking.

Additionally, the incidence, progression, clinical manifestation, and genetic risks of many diseases, such as MetDs differ between females and males, which indicates that one sex may have endogenous protective or risk factors that could become targets for therapeutic interventions. Current sexual differentiation theory suggests that three major classes of factors cause sex differences. First, some sex differences are caused by different circulating levels of ovarian and testicular hormones, known as “activational effects”. These differences are reversible because they are eliminated by gonadectomy of adults. Second, certain sex differences persist after gonadectomy in adulthood and represent the effects of permanent or differentiating effects of gonadal hormones, known as “organizational effects,” that form during development. A third class

of sex differences are caused by the inequality of action of genes on the X and Y chromosomes in male (XY) and female (XX) cells, and are called “sex chromosome effects”. To date, few studies have systematically evaluated the relative size and importance of each of these three classes of factors acting on phenotypic or global gene regulation systems. Therefore, we used the Four Core Genotypes (FCG) mice on a C57BL/6J B6 background. In FCG mice, the Y chromosome (from strain 129) has sustained a spontaneous deletion of *Sry* (testis determining), and a *Sry* transgene is inserted onto chromosome 3. Where, we defined “male” (M) as a mouse with testes, and “female” (F) as a mouse with ovaries. FCG mice include XX males (XXM) and females (XXF), and XY males (XYM) and females (XYF).With this approach we could study the tissue-specific contributions and the interactions of activational, organizational, and sex chromosome effects on gene regulation to better understand the specific sex factors and genes contributing to complex diseases such as MetDs.

Aiming to address these areas, my work was centered on 1) the development of a user friendly and accessible webserver for multiomics integration “Mergeomics”, a pipeline that integrates genetic associations, tissue-specific functional genomics such as eQTLs, canonical pathways and gene-gene interaction networks; 2) utilizing Mergeomics, to pinpoint key regulatory hubs of MetD related subnetworks. These key drivers shall have the potential to normalize the entire gene network spectrum and serve as novel therapeutic targets [21-23]; 3) Using the FCG model, we can assess the role of the three sex-biasing factors on gene expression, molecular pathways, and gene network organization in key metabolic tissues. We can uncover how the three factors influence gene regulation involved in critical processes and their tissue specific contribution. We can further

integrate sex-biased genes and networks influenced by each sex-biasing factor with diverse diseases in human populations to predict the phenotypic consequences of each sex-biasing factor.

The details of the updated and user-friendly webserver for Mergeomics are described in **Chapter 2**. Compared to other tools, Mergeomics not only accommodates diverse data types (GWAS, EWAS, TWAS, PWAS, MWAS) from different sources, or species for a given disease, but also considers relationships between omics layers through functional genomics such as expression quantitative trait loci (eQTLs), molecular pathways, and tissue-specific gene regulatory networks to derive disease networks and predict therapeutics. Mergeomics also uses full summary statistics, not raw data, as input, thereby reducing the need for raw data processing and harmonization and for pre-determining a specific cutoff to call for significant markers. Mergeomics has the ability to conduct pathway analysis and model gene regulatory networks, protein-protein interaction networks, and transcription factor networks in order to predict and visualize network regulators of disease. These unique features help maximize the utility of existing datasets and overcome limitations of other tools which utilize a narrower range of multi-omics data sources, do not provide mechanistic interpretations, or require programming skills with no intuitive webserver for ease of use.

In **Chapter 3**, we utilized an integrative genomics approach leveraging diverse genomic data from human populations to investigate whether genetic variants associated with various plasma lipid traits, namely, total cholesterol, high and low density lipoprotein cholesterol (HDL and LDL), and triglycerides, from GWASs were concentrated on specific parts of tissue-specific gene regulatory networks. In addition to the expected lipid metabolism pathways, gene subnetworks involved in

"interferon signaling," "autoimmune/immune activation," "visual transduction," and "protein catabolism" were significantly associated with all lipid traits. In addition, we detected trait-specific subnetworks, including cadherin-associated subnetworks for LDL; glutathione metabolism for HDL; valine, leucine, and isoleucine biosynthesis for total cholesterol; and insulin signaling and complement pathways for triglyceride. Finally, by using gene-gene relations revealed by tissue-specific gene regulatory networks, we detected both known (e.g., *APOH*, *APOA4*, and *ABCA1*) and novel (e.g., *F2* in adipose tissue) key regulator genes in these lipid-associated subnetworks. Knockdown of the *F2* gene (coagulation factor II, thrombin) in 3T3-L1 and C3H10T1/2 adipocytes altered gene expression of *Abcb11*, *Apoa5*, *Apof*, *Fabp1*, *Lipc*, and *Cd36*; reduced intracellular adipocyte lipid content; and increased extracellular lipid content, supporting a link between adipose thrombin and lipid regulation. Our results shed light on the complex mechanisms underlying lipid metabolism and highlight potential novel targets for lipid regulation and lipid-associated diseases.

In **Chapter 4,** we utilized a systems biology approach to further elucidate the contributing factors to type 2 diabetes development. The islet in type 2 diabetes is characterized by islet amyloid derived from islet amyloid polypeptide (IAPP), a protein co-expressed with insulin by beta cells that when misfolded and in aggregate form may contribute to beta cell failure. Human IAPP (hIAPP) toxicity is most potently mediated by small intracellular membrane permeant oligomers. Species with amyloidogenic IAPP, such as humans, non-human primates and cats, share vulnerability to type 2 diabetes, while those with non-amyloidogenic IAPP, such as mice and rats, do not. While numerous hypotheses have been put forward to explain the wide-ranging changes in islets in type 2 diabetes, there is a consensus that misfolded protein stress induced by toxic

oligomers of amyloidogenic proteins initiate these changes in neurodegenerative diseases. Given the known proximal role of misfolded protein stress in neurodegenerative diseases, and the connection of the risk factors for type 2 diabetes to misfolded protein stress, we hypothesised that hIAPP misfolded protein stress may be a proximal cause of the wide-ranging changes in islets in individuals with type 2 diabetes. In the study we evaluated the islet transcriptome from a mouse model of beta cell hIAPP toxicity before diabetes onset in order to avoid the confounding effects of hyperglycaemia. To control for the increased burden of IAPP expression, we evaluated the transcriptome from mice overexpressing rodent IAPP (rIAPP). We then compared the changes in the transcriptome of hIAPP or rIAPP islets to those in humans with prediabetes or type 2 diabetes to establish if the changes in the islet in type 2 diabetes are potentially attributable in part to hIAPP protein misfolding stress. Overall, the study suggests that much of the islet transcriptome in type 2 diabetes is adaptive to the increased beta cell burden of protein synthesis and folding. Beta cell hIAPP toxicity induces a prominent islet inflammatory response, consistent with that observed in type 2 diabetes, implying protein misfolding stress may serve to initiate or contribute to beta cell injury in type 2 diabetes. There are also shared pro-survival gene networks in hIAPP and type 2 diabetes islets. Strategies to suppress IAPP expression warrant further investigation due to the mounting evidence to suggest its role in type 2 diabetes pathogenesis.

In **Chapter 5,** we utilized an integrative and systems biology approach through genetic and transcriptomic data in an attempt to holistically differentiate liver fibrosis pathogenesis between male and female mice. Our study overall highlights a greater immune response in males along with more protein and lipid metabolism abnormalities; for females, we found more carbohydrate metabolism related abnormalities contributing to liver fibrosis. Through our key driver analysis,

novel key drivers were found. More research into these genes can help identify plausible targets and create sex-specific therapeutic treatments for NASH.

In **Chapter 6**, we focused on teasing apart the relative contribution of sex hormones, sex chromosomes and gonads in gene regulation in tissues critical to MetDs. The design allowed detection of differences caused by three factors contributing to sex differences in traits. (1) "Sex chromosome effects" were evaluated by comparing XX and XY groups. (2) "Gonadal sex effects" were determined by comparing mice born with ovaries vs. testes. Since mice were analyzed as adults after removal of gonads, the gonadal sex effects represent organizational (long-lasting) effects of gonadal hormones, such as those occurring prenatally, postnatally, or during puberty. This group also includes effects of the *Sry* gene, which is present in all mice with testes and absent in those with ovaries. Any direct effects of *Sry* on non-gonadal target tissues would be grouped with effects of gonadal sex. (3) "Hormone treatment effects" refers to the effects of circulating gonadal hormones (activational effects) and were evaluated by comparing E vs. B groups for estradiol effects, and T vs. B groups for testosterone effects. We found that the activational hormone levels have the strongest influence on gene expression, followed by the organizational gonadal sex effect, and last, sex chromosomal effect, along with interactions among the three factors. Tissue specificity was prominent, with a major impact of estradiol on adipose tissue gene regulation and of testosterone on the liver transcriptome. The networks affected by the three sex-biasing factors include development, immunity and metabolism, and tissue-specific regulators were identified for these networks. Furthermore, the genes affected by individual sex-biasing factors and interactions among factors are associated with human disease traits such as coronary artery disease, diabetes, and inflammatory bowel disease. Our study offers a tissue-specific

account of the individual and interactive contributions of major sex-biasing factors to gene regulation that have broad impact on systemic metabolic, endocrine, and immune functions.

Chapter 7 is a concluding summary of the PhD work completed and covers the future directions of the research topics

Chapter 2. Mergeomics 2.0: A Web Server for Multiomics Data Integration to Elucidate Disease Networks and Predict Therapeutics

2.1 Introduction

The advent of omics technologies has made significant strides in unveiling various disease associated genetic and epigenetic variants, genes, proteins, and metabolites. The ever-growing source of multiomics datasets available including genomics, epigenomics, transcriptomics, proteomics, and metabolomics now presents a new challenge of integrating these different data types for more meaningful and holistic interpretation of complex diseases. To conduct a comprehensive investigation of disease pathogenesis, we must consider multiple omics layers that contribute to biological complexity [24]. The computational pipeline Mergeomics was developed to meet the need for multiomics integration and functional interpretation to obtain mechanistic understanding. Mergeomics provides flexibility to incorporate the full spectrum of summary statistics (not just top hits) of individual layers of omics or multiomics data simultaneously along with diverse functional genomics data across data types, studies, and species. As such, genome-wide association studies (GWAS) as well as epigenome- (EWAS), transcriptome- (TWAS), proteome- (PWAS), and metabolome-wide association studies (MWAS) can all be accommodated. The development of our Mergeomics tool follows the philosophy of utilizing a systems biology approach to unravel the complex interactions across molecular domains as well as cell types, tissues, and organ systems that occur in disease. In particular, we are guided by the omnigenic disease model [25], which states that a large proportion of the genome likely contributes to disease pathogenesis through molecular interactions both within and between tissues. Utilizing this data-driven analysis considering the interactions among different omics layers and tissue contexts will uncover global maps to identify critical targets in disease pathogenesis, which can be followed by

experimental approaches to investigate the detailed events that occur through the predicted molecules or pathways.

With the abundance of omics data available, it is unsurprising that various tools or methods have been developed to help better integrate and interpret these datasets [26-28]. These tools can be broadly categorized into two application categories: multiomics biomarker predictions of diseases or subtypes (i.e., uncovering correlative or predictive but not necessarily disease-causing features) or mechanistic understanding of disease pathogenesis (i.e., regulators, molecular interactions, and processes involved in disease development). Mergeomics focuses on mechanistic modeling but not predictive modeling. In terms of approaches, fusion (such as PFA [29], SNF [30], PSDF [31]), Bayesian (e.g., iCluster [32], PSDF [31], BCC [33]), correlation, multivariate (e.g., MFA [34], IntegrOmics [35], MixOmics [36]), pathway and network methods (PARADIGM [37], SNF [30], iOmicsPASS [38], MiBiOmics [39], Lemon-Tree [40], PaintOmics [41], NetICS [42], Metascape [43]) have been implemented [26, 27, 44]. Mergeomics falls within the network method category that mainly focuses on understanding disease pathogenesis through uncovering multiple molecular targets within biological processes important to disease. The benefit of a network approach over other integrative options is in its ability to provide biological interpretability, which is reliant not on the identification of latent structures through mathematical deconvolution but on the utilization of prior information based on molecular interactions, which can help provide clear targetable options (e.g., genes) in disease. Compared to other tools, Mergeomics not only accommodates diverse data types (GWAS, EWAS, TWAS, PWAS, MWAS) from different sources, studies, or species for a given disease, but also considers relationships between omics layers through functional genomics such as expression quantitative trait loci (eQTLs), molecular pathways, and tissue-specific gene regulatory networks to derive disease networks and predict therapeutics.

Mergeomics also uses full summary statistics, not raw data or lists of top associations, as input, thereby reducing the need for raw data processing and harmonization and for pre-determining a specific cutoff to call for significant markers. Mergeomics has the ability to conduct pathway analysis and model gene regulatory networks, protein-protein interaction networks, and transcription factor networks in order to predict and visualize network regulators of disease. These unique features help maximize the utility of existing datasets and overcome limitations of other tools which utilize a narrower range of multiomics data sources, do not provide mechanistic interpretations, or require programming skills with no intuitive web server for ease of use.

Since the release of the open source Mergeomics R package (https://bioconductor.org/packages/release/bioc/html/Mergeomics.html) [45] and web server in 2016 [46], this tool has been used to model a diverse set of diseases including cardiometabolic disorders such as non-alcoholic fatty liver disease [47], cardiovascular disease [48-50], and type 2 diabetes [51], autoimmunity including psoriasis [52] and rheumatoid arthritis [53], alcohol dependence [54], brain injury [55], Sjogren's syndrome [56], and environmental contributions to disease [57-59]. Importantly, multiple validations of molecular predictions from Mergeomics with *in silico*, *in vitro,* and *in vivo* studies highlight the validity and causal nature of the disease network predictions [20, 47, 51, 52, 55, 59-63]. Due to increasing demand for multiomics integration and interpretation from scientists with different areas of expertise, we have implemented major revisions and improvement on the Mergeomics web server. Specifically, we have redesigned the user interface, simplified workflows, offered detailed tutorials and case studies, and provided more datasets and network models for utilization. The Mergeomics 2.0 web server offers the scientific community much-improved accessibility to our pipeline, caters to each user's specific goals in multi-omics studies, and addresses a broad range of biological questions, particularly emphasizing

a mechanistic understanding of disease pathogenesis and prediction of potential therapeutics based on mechanistic understanding.

2.2 Results and discussions

Overview and Updates on the Core Functions of Mergeomics**:** *Overview of core functions:* Mergeomics 2.0 features four core functions as previously implemented in version 1.0 with an addition of a new function. First, we provide a preprocessing tool, Marker Dependency Filtering (MDF) to remove omics marker redundancies such as linkage disequilibrium (LD) between single nucleotide polymorphisms (SNPs). Second, Marker Set Enrichment Analysis (MSEA) is used to identify omics-informed disease processes through the integrations of omics markers such as SNPs with functional genomics, canonical pathways, or co-expression networks. Third, Meta-MSEA runs MSEA on multiple datasets and conducts pathway/network level meta-analysis to retrieve consistent disease processes informed across datasets. Fourth, Key Driver Analysis (KDA) pinpoints network regulators of disease processes based on the topology of biological networks. In Mergeomics 2.0, we added a new functional module called PharmOmics, which takes as input multiomics-informed disease pathways or networks from Mergeomics to match with drug signatures to predict potential therapeutic drugs.

Introduction of PharmOmics into Mergeomics 2.0: We have recently developed a novel species- and tissue-specific network-based drug repositioning tool, PharmOmics, which is based on *in vivo* molecular studies of drugs [64]. PharmOmics is a complementary drug repositioning tool to other existing tools, such as CMap [65] and LINCS L1000 [66], which are mostly based on *in vitro* cell line data. We provide two drug repositioning methods: network-based drug repositioning and gene overlap-based drug repositioning. Network-based drug repositioning ranks drugs based on the

degree of connectivity of genes influenced by drug treatments to disease gene signatures in a given gene network model [67]. Gene overlap-based drug repositioning is based on the degree of direct overlap between drug genes and disease genes. Users can directly input their disease pathway results from MSEA (genes from disease pathways are used as input) or KDA (genes from the disease network or significant key drivers (KDs) are used as input). For both MSEA and KDA, specific gene sets can be input into drug repositioning for a more refined analysis. As PharmOmics is based on gene expression studies, inputs are limited to genes or proteins. Users can also input their genes of interest into PharmOmics for drug repositioning analysis without running any other functions in Mergeomics.

Flexible workflows using the core functions: Each of the main functions of Mergeomics described above can be utilized as a standalone analysis tool or a multi-step workflow with several different cases as portrayed in **Figure 2.1**. There are four cases or starting points that a user has the option to select. In case one, the user has one GWAS dataset and is prompted first to run MDF where they provide their association dataset, mapping data (e.g., SNP to gene), and marker dependency data (linkage disequilibrium or LD in the case of GWAS) to retrieve corrected SNP associations and mapping files. The MDF step is optional if the user does not wish to correct for LD, although we highly recommend this correction to avoid statistical artefacts due to LD. These results along with a gene set are fed into MSEA to uncover disease-associated pathways, which can be further analyzed in KDA to identify key regulators or PharmOmics for drug repositioning. In case two, the user has EWAS, TWAS, PWAS, or MWAS data, and they are led to MSEA, where MDF and marker mapping are optional. As in the GWAS path, results from MSEA can be carried to KDA or PharmOmics. In case three, the user has multiple omics datasets and utilizes Meta-MSEA,

which will run MSEA on each dataset and then conduct a meta-analysis across datasets to retrieve consistent biological processes, which can be input into PharmOmics or KDA. Finally, in case four, the user has a gene set and network of interest and can directly run KDA, which will provide KD genes and a subnetwork visualization of the top KDs, and the KDs or subnetwork can be input into PharmOmics to predict drugs.

Update on Marker Dependency Filtering (MDF): MDF prepares input files for MSEA by correcting for dependency between omics markers and is an optional function. This preprocessing step is most commonly used for GWAS data to correct for LD between SNPs and filter out redundant SNPs, which is critical for removing redundant association signals that can result in statistical and biological artefacts in downstream analysis. Another purpose of MDF is to link the SNPs to potential downstream genes based on functional evidence, such as tissue-specific eQTLs. Correcting for dependency between other omics markers is currently seldom used. However, this feature can be utilized to correct for dependency between other types of markers (methylation sites, transcripts, etc.), if desired. MDF uses as input an association file which details markers (e.g., SNPs) and their disease association strengths (e.g., $-\log_{10}$ p-values or effect size, note that p values are prohibited as MDF ranks larger values as stronger association strength, which is opposite of p values), a mapping file used for marker to gene mapping (e.g., SNPs are mapped to genes to be enriched for gene sets), and a marker dependency file indicating the dependency between markers (e.g., LD between SNPs, to remove redundant markers) (**Figure 2.2**). The resulting corrected association and mapping files are then used as input to MSEA. MDF also allows for the selection of a top percentage of markers (50% or 25% recommended) to be considered in the analysis which reduces noise from low signal markers.

Updates to MDF include an increased number of marker to gene mapping options such as the addition of all available tissue-specific Genotype-Tissue Expression project (GTEx) [68] cis-eQTLs and splicing QTLs (cis-sQTLs) (**Table 2.1**), the ability to combine up to five multiple mapping options, and the inclusion of LD files for all 26 populations from 1000 Genomes (1000G) [69] and methylation disequilibrium from EWAS software 2.0 [70]. For analysis starting from GWAS data, MDF is a default preprocessing step, but we have included the option to skip MDF. For analytical paths starting from other omics data, users have the option to add MDF if needed.

Update on Marker Set Enrichment Analysis (MSEA): In MSEA, full summary statistics of omics markers such as SNPs from GWAS, epigenetic sites from EWAS, genes from TWAS, proteins from PWAS, or metabolites from MWAS and their disease association values are taken as input and are integrated with functional genomics, canonical pathways, or co-expression networks to retrieve disease-associated pathways and networks. MSEA calculates and summarizes enrichment of disease/trait omics markers in sets of functionally related genes, such as canonical pathways and co-expression networks, across a range of statistical cutoffs in the full summary statistics file using a chi-square like statistic, and then uses permutation to determine statistical p values for the enrichment. We emphasize the importance to provide the association strength of the given marker wherein a larger number reflects greater association such as $-\log_{10}$ p-values or effect size to avoid incorrect downstream analysis and interpretability.

MSEA is able to analyze diverse data types, and each has different considerations of inputs which was partly described in the above MDF section (**Figure 2.2**). The output from MSEA can be interpreted as omics-informed disease pathways or networks. If GWAS is used, MSEA results can imply causal disease processes since GWAS carries causal inference. For other omics data,

the MSEA results can only be interpreted as disease-associated processes but may or may not be causal. Considering GWAS along with other omics data, in our opinion, is a useful way to identify causal genes and processes. We also advise the user to take care in their interpretation of the names or annotations of pathways deemed to be significant (FDR < 0.05) as some can be misleading. Attention to the genes enriched in a given pathway derived from the input dataset should be checked in the gene details output file to confirm whether the pathway name is indeed appropriate as the genes may be more suitable or representative of another biological process. A user can conclude the analysis with results from MSEA or use the MSEA results as input to KDA with a user-defined statistical cutoff to identify network KDs of the disease processes based on molecular network topology.

In Mergeomics 2.0, we added the ability to use disease-associated gene sets derived from MSEA as input to PharmOmics for drug repositioning analysis, selecting either specific gene sets or by false discovery fate (FDR) or P-value threshold, to pinpoint drugs whose gene signatures align with those of the disease-associated gene sets identified by MSEA.

Update on Meta-MSEA: Meta-MSEA allows for integration of multiple of the same omics type (e.g. two or more GWAS datasets) or multiple omics types (e.g., GWAS, EWAS, TWAS) by running MSEA for each omics type followed by a meta-analysis. This integration reveals consistencies and differences in biological perturbation across different omics types or different studies of the same omics type.

In Mergeomics 2.0, we improved the guidance of running Meta-MSEA in regard to the differences in preprocessing of the different types of omics data. In addition, we have increased the flexibility of this analysis to allow for specific inputs and parameters for each association data.

After each individual omics dataset is added, the user will be able to review which datasets have been successfully uploaded and their individual MSEA parameters with the option to add additional datasets or delete certain datasets, providing an easy way to track all the different inputs. As in results from individual MSEA, significantly associated gene sets from Meta-MSEA can be used as input to KDA or PharmOmics drug repositioning. We have also implemented user-defined individual MSEA FDR cutoffs to KDA in that the disease-associated pathways must pass all individual MSEA FDR cutoffs as well as the meta-FDR to be used in KDA, allowing the user to focus on the most consistent and robust disease processes across different datasets. In addition, we now provide heterogeneity q and p values to indicate the variability between datasets.

Update on Key Driver Analysis (KDA): KDA identifies essential regulators of disease-associated pathways and networks, which are then visualized in the web browser using Cytoscape.js (**Figure 2.3**). KDA results can also be downloaded as network files ready to be used on Cytoscape Desktop for further customization of network visualization. A Chi-square like statistic, , is used to identify genes (KDs) that are connected to a significantly larger number of disease-associated genes than what is expected by random chance. O and E represent the observed and expected numbers of disease-associated genes in a hub subnetwork, and E is estimated by where N_p is the disease gene set size, N_k is the hub degree, and N is the full network order. KDs represent prioritized disease regulatory genes based on network topology. In numerous recent applications of Mergeomics, top KDs have been shown to be causal for diseases based on experimental evidence [47, 51, 60], thereby supporting their importance. KDA can be utilized as a follow up analysis to MSEA or Meta-MSEA, and it can also be used as an independent analysis using a gene list of interest and a given network as inputs. For instance, the user can upload a list of curated disease genes and choose

or upload a relevant network to run KDA to identify how the disease genes interact in the network and whether there are key hub nodes in the network that regulate the disease genes.

In Mergeomics 2.0, we added the ability to visualize input gene overlap with a given network, if any, in the case that no KDs were found. The user can therefore be better informed on the reason for the lack of KD hits based on the distribution and connectivity of the input genes in the network. If few input genes are in the network or the input genes are widely dispersed in the network, that will explain the lack of KD identification. We have additionally increased the number of sample tissue-specific networks (**Table 2.1**). As we have done similarly with MSEA and Meta-MSEA, disease subnetworks or significant KDs from KDA can be used directly for PharmOmics drug repositioning, and users can further customize which processes in the subnetwork are used in drug repositioning for a more focused analysis.

Data and sample input updates: We have significantly augmented the amount of Mergeomics-ready sample files with commonly used datasets and will continue to actively update sample files to enrich data resources useful for users on a monthly basis.

In Mergeomics 2.0, we include over 20 GWAS datasets from a broader range of diseases from metabolic syndrome to psychiatric disorders (**Table 2.1**). For omics dependency filtering options, we have added the full array of LD data from 26 human populations studied in 1000G [69] with LD above 0.5 and 0.7 for SNP filtering to remove redundant SNPs in high LD and have also provided an example methylation disequilibrium data file for correction of EWAS data. For SNP to gene mapping options, we have added all tissue-specific cis-eQTL and cis-sQTL mapping files from the GTEx version 8 (q-value < 0.05) [68], which inform on the SNPs associated with gene expression level changes (eQTL) or differential splicing (sQTL). In addition, we offer

ENCODE regulatory gene mapping [71] and various chromosomal location-based mapping options (**Table 2.1**). Moreover, we have increased the number of curated pathways from version 1 to include all gene sets from Molecular Signatures Database (MSigDB) [72] such as KEGG [73], Reactome [74], Biocarta [75] canonical pathways, chemical and genetic perturbation, microRNA and transcription factor targets, and cell type marker signatures, Gene Ontology [76], Wikipathways [77], and Bioplanet [78], among others (**Table 2.1)**. To complement knowledge-based pathways, we include our data-driven tissue-specific co-expression network modules utilizing GTEx transcriptome datasets and co-expression network construction tools MEGENA [79] and WGCNA [80] (**Table 2.1)**. Finally, we have constructed tissue-specific Bayesian gene regulatory networks [81] and include them as sample networks on the web server. We also provide human protein-protein interaction networks [82], transcription factor networks [83], and GIANT networks [84] (**Table 2.1**). Sample files are available to download from our sample resources page (http://mergeomics.research.idre.ucla.edu/samplefiles.php), and further clarification on correct formatting of input data is detailed on the webserver and in **Figure 2.2**.

***General Updates*:** We have completely redesigned the user interface to make it much more intuitive in guiding the use of the pipeline for different omics data types. To start the pipeline, users are presented with four workflow options in regard to the data that they have: (1) GWAS, (2) EWAS, TWAS, PWAS, or MWAS, (3) multiple of the same or different types of omics data, and (4) a gene set list (user can run KDA or PharmOmics). The separation of GWAS from other omics datasets is for the additional need to correct for LD and link SNPs to candidate genes through MDF, which is not required or is optional for other omics datasets. For EWAS, a marker to gene mapping file is required if the user uploads epigenetic markers such as CpG probes. For MWAS,

a metabolite to gene mapping file is optional but not required if the user uses metabolite sets as the marker sets to be tested. Marker mapping is not needed for TWAS and PWAS as the markers (genes and proteins) match the gene sets. This workflow design clearly delineates what is needed for each specific data type, which is more intuitive for the user. We have also improved the fluidity and presentation of the pipeline workflow as each collapsible step appears below the previous in a vertical format so that the user can revisit input files, parameters, and results of previous steps in the pipeline and choose to rerun a step at any point in the pipeline. A workflow map with navigation links is also generated on the left sidebar to help visualize the steps taken and downstream paths.

We have improved the system that allows users to return to their session where results of analyses can be revisited or continued onto the next step using a unique tracking ID number that is valid for up to 48 hours after the start of their session. The user can also choose to have their results emailed upon completion of the analysis, which is not mandatory but is recommended because the tracking ID allows the user to reload their session and retrieve completed jobs in case a crash occurs. Because later steps of the pipeline, KDA and PharmOmics, can be run independently, downloadable result files from MSEA and KDA can be uploaded directly to the desired next step in the analysis (e.g., MSEA to KDA/PharmOmics or KDA to PharmOmics).

In addition, we have improved case-specific responsiveness of the web server to better inform the user such as error-checking of user uploaded files to ensure the file is formatted correctly and providing feedback on user results such as whether the results are substantial to be used in the next step of the analysis. Across all applications of Mergeomics 2.0 we have provided an improved review of analysis inputs and parameters and new interactive tables with pagination, sorting, and search features (**Figure 2.5**). We also implemented real-time runtime analysis output

and progress updates, and this job log including any errors that occurred is available for download at the conclusion of the analysis. Finally, we have improved multi-device usage including on tablets and phones such that it can be appropriately viewed on different screen sizes. We further improved the tutorial to explain how to prepare input files and the underlying methods and various parameters underlying each computational function and provide a tutorial video to demonstrate the different pipeline options.

***Use Case: Identifying pathogenic pathways and networks for psoriasis based on multiomics data*:** The use case described here utilizes publicly available GWAS and EWAS data to perform Meta-MSEA and subsequently KDA to find pathogenic pathways and regulators of psoriasis (**Figure 2.4)**. All data used in this example are provided as sample data on the web server which can be downloaded (http://mergeomics.research.idre.ucla.edu/samplefiles.php). GWAS of psoriasis was obtained from dbGAP database (www.ncbi.nlm.nih.gov/gap) with accession phs000019.v1.p1, and two EWAS of psoriasis was obtained from GEO (GSE31835 and GSE63315) [85, 86]. For preprocessing of the GWAS data, we use the top 50% of SNPs ranked by $-\log_{10}$ p-value and correct for LD between SNPs using MDF with the psoriasis GWAS summary statistics as the marker associations, combined skin and blood eQTLs as the SNP to gene mapping, and the 1000G CEU LD structure containing SNPs with $r^2 > 0.7$ as the marker dependency file. For the EWAS data, CpG sites are mapped to adjacent genes within 5 kb. Next, we chose canonical pathways from the KEGG database and a positive control gene set from the NHGRI-EBI GWAS catalog [87] for psoriasis as the pathways or marker sets to be examined. We ran Meta-MSEA across GWAS and the two EWAS datasets. At the conclusion of Meta-MSEA, a set of results files

and a summary table display are generated on the webpage detailing the pathways ranked by meta p-value and their top mapped markers and corresponding genes **(Figure 2.5A)**.

As shown in **Figure 2.5A**, "Cytokine cytokine receptor interaction", "Graft versus host disease", and "Natural killer cell mediated cytotoxicity" were three of the top pathways identified among others. Following Meta-MSEA, KDA was run with default parameters using nonredundant supersets (pathways that were merged due to significant overlap in genes enriched) significantly associated with psoriasis from Meta-MSEA and a blood GIANT Bayesian gene regulatory network [84] (chosen due to the relevance of the immune system to psoriasis) to identify KDs of the disease related gene sets. At the conclusion of KDA, a table is produced on the webpage listing the KDs and significance of enrichment of psoriasis associated gene sets in their network neighborhood (**Figure 2.5B**). For example, *ICAM2* is identified as the KD for the viral myocarditis/tight junction/autoimmune pathway, and *CD2* is identified as a KD for the Autoimmune Disease Superset. By default, the top five KDs and their local subnetworks from each gene set is included in the interactive subnetwork visualization in the browser (**Figure 2.3**).

With addition of the PharmOmics pipeline to the Mergeomics webserver, we ran two drug repositioning analyses: one directly from the MSEA results and the other considering the whole subnetwork derived from the KDA (**Figure 2.5C**). In this case study, we do not consider gene expression direction changes (upregulation or downregulation) in psoriasis and therefore will simply be utilizing genes involved in disease without considering if they are protective or pathogenic; thus, our predicted drug list will contain drugs that can induce as well as drugs that can potentially treat psoriasis. In addition, PharmOmics interrogates all drug signatures regardless of the tissue or species, and the user can choose to focus on the relevant drug studies for their given dataset. For example, we mainly focused on drugs that were studied in integument tissue, due to

its relevance to psoriasis. In the top 10 repositioned drugs derived from psoriasis associated gene sets from Meta-MSEA, we find 8/10 to have prior association with a role in psoriasis pathogenesis (Imiquimod [88]) or treatment including broad options suggesting classes of drugs such as anti-inflammatory, immunosuppressant, JAK inhibitors, and anti-rheumatic drugs and more specific options such as Baricitinib [89], Ingenol [90], and Etinostat [91] (**Figure 2.5C**). Similarly, using the psoriasis subnetwork from KDA highlights Imiquimod and Ingenol within the top 10 drugs, and the remainder of the results are broad categories such as JAK inhibitors, anti-inflammatory drugs, and anti-rheumatic drugs, each of which are actively being investigated in the treatment of psoriasis [92, 93]. The predicted drugs can form new hypotheses for experimental testing.

***Future Directions*:** The web server will continue to actively incorporate the most up-to-date public resources including multiomics association data, functional genomics data such as eQTLs or protein QTLs (pQTLs), knowledge-based pathways, gene co-expression networks, and gene regulatory networks on a monthly basis. We will also include single cell networks when available to understand the gene regulatory connections within a given cell type or between cell types rather than across a whole tissue, which will offer higher resolution molecular mechanisms of disease pathogenesis. Cell type level association data derived from single cell omics studies can be used in the current platform. We will also continue incorporating additional analytical functions into the web server such as different forms of meta-analysis that can be conducted within the Meta-MSEA tool as well as adding new features to better accommodate analysis of data types that are currently not considered or well tested, such as gut microbiome and spatial transcriptomics data.

2.3 Conclusions

Thanks to advancements in technologies, the number of multiomics data (GWAS, EWAS, TWAS, PWAS and MWAS) increases exponentially. The systems biology approach to interrogate multi-tissue multi-omics data has become a promising method to understand biology in a data-driven way and sheds light on the hidden mechanisms. However, the computational knowledge and skills required to perform such integrative analysis are often considered as a hurdle to many biologists. Therefore, the Mergeomics web server was developed to lower this barrier to enable fellow researchers to dive into multiomics systems biology. The current update, Mergeomics 2.0, is a versatile web-based tool that provides multiomics data integration using a pathway- and network-based approach. The improvements we made support a wide range of precalculated networks and data for all steps of the pipeline to fulfill a variety of needs and research purposes. In addition, the new user interface presents a more intuitive and flexible environment that greatly improves its ease of use. In addition to a detailed tutorial, each step of the pipeline contains embedded guidance to facilitate the user experience. We believe that the Mergeomics 2.0 and systematics approach applied here will accelerate our understanding of complex diseases and guide therapeutics.

2.4 Tables

Table 2.1 Sample resources on Mergeomics web server

General Data Category	Data Type	Specifics	Citation
Association Data	GWAS	Alzheimer's Disease	[94]
		Attention Deficit Hyperactivity Disorder	[95]
		Alcohol Dependence	[96]
		Body Mass Index	[97]
		Breast Cancer	[98]
		Coronary Artery Disease	[99]
		Fasting Glucose	[100]
		Heart Failure	[101]
		High Density Lipoproteins (HDL)	[102]
		Low Density Lipoproteins (LDL)	[102]
		Major Depressive Disorder	[103]
		Parental Lifespan	[104]
		Parkinson's Disease	[105]
		Psoriasis	[106]
		Severe illness in Covid-19	[107]
		Schizophrenia	[108]
		Stroke	[109]
		Systemic Lupus Erythematosus	[110]
		Type 2 Diabetes	[111]
		Total Cholesterol	[102]
		Triglycerides	[102]
	EWAS	Birth Weight	[112]
		Maternal Anxiety	[113]
		Social Communication	[114]
		Psoriasis	[85, 86]
Marker Mapping	Chromosomal Distance	10kb, 20kb, 50kb	[69]
	Regulome	RegulomeDB (ENCODE)	[115]
	eQTL	49 tissue types	[68]
	sQTL	49 tissue types	[68]
Marker Dependency	Linkage Disequilibrium	26 populations at $r^2 > 0.5$ and > 0.7	[69]
	Methylation Disequilibrium	$r^2 > 0.5$	[70]
Marker Sets	Canonical (knowledge based)	KEGG	[73]
		Reactome	[74]
		BioCarta	[75]
		MSigDB	[72]
		GO	[76]
		BioPlanet	[78]
		WikiPathways	[77]
	Data-driven (Co-expression)	24 tissue specific modules (WGCNA/MEGENA)	[68, 79, 80]
Networks	Gene Regulatory Human and Mouse Composite (Bayesian)	Adipose, Blood, Brain, Kidney, Liver, Muscle	[12, 81, 116-120]
	Gene Regulatory (GIANT)	Adipose, Blood, Brain, Kidney, Liver, Muscle	[84]
	Protein-Protein Interaction	STRING	[82]
	Transcription Factor-Target (FANTOM5)	Adipose, Blood, Brain, Kidney, Liver, Muscle	[83]

2.5 Figures

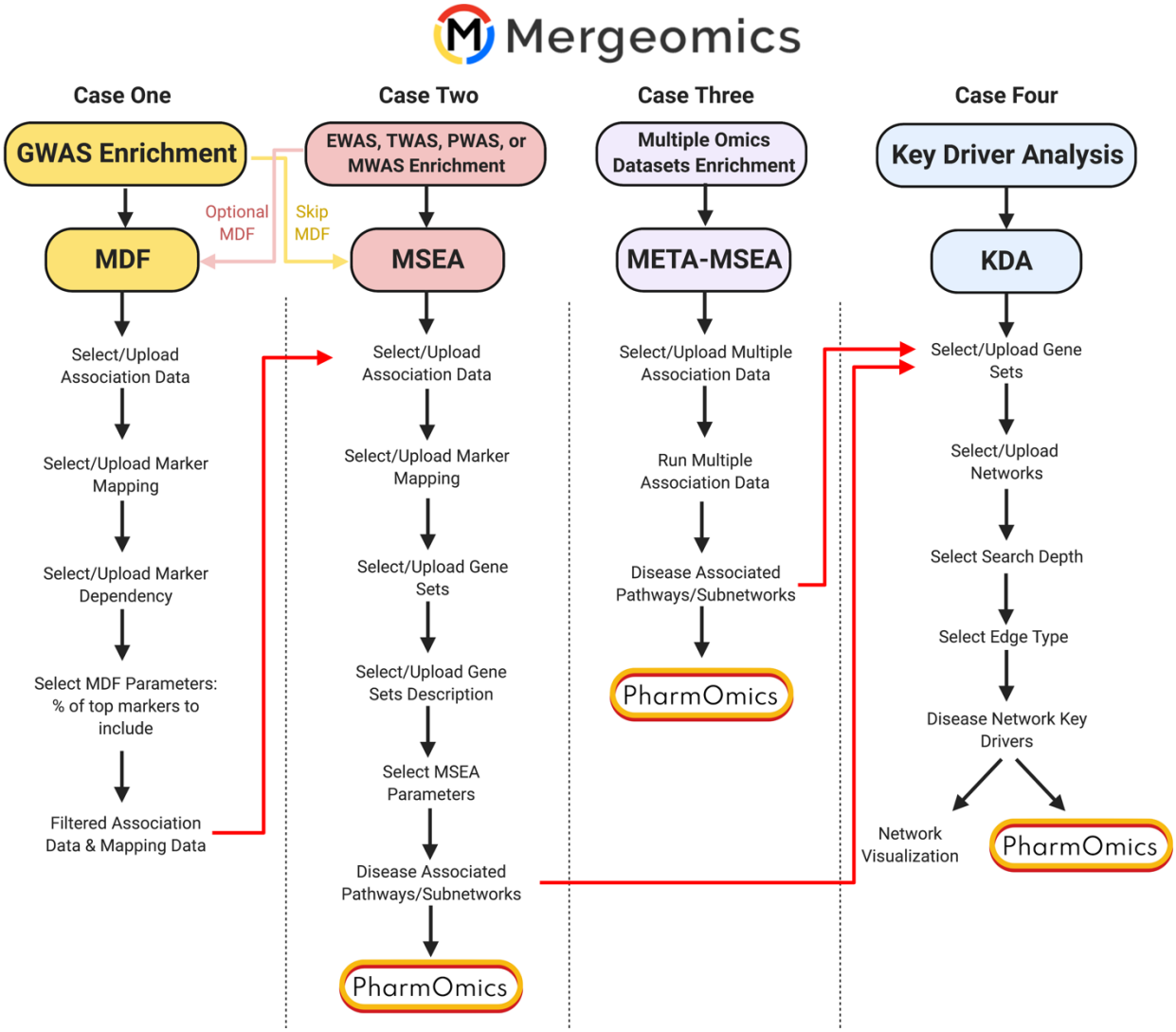

Figure 2.1. Workflow of Mergeomics. We provide four options on the web server to tailor to the user's data type. Case One: Individual GWAS analysis. For GWAS datasets we advise utilizing the MDF function; however, we also provide the ability to skip MDF and go directly into MSEA and follow the workflow to PharmOmics or KDA. Case Two: Individual EWAS, TWAS, PWAS, or MWAS analysis. In this case, we can directly start at MSEA without MDF; however, we also provide the ability to utilize the MDF function if needed. From here the user can feed the MSEA results into PharmOmics or to KDA. Case Three: Multiomics analysis. If the user has multiple omics of the same type (e.g., two GWAS) or different types (e.g., TWAS and EWAS), they can utilize the Meta-MSEA function to derive disease-associated pathways and can input their results into PharmOmics or KDA. Case Four: A gene list(s) to run KDA. The user in this case can upload their gene sets of interest and upload or select a network to derive KD genes and visualize top KD subnetworks. The disease subnetwork or significant KDs can be fed into PharmOmics for drug repositioning.

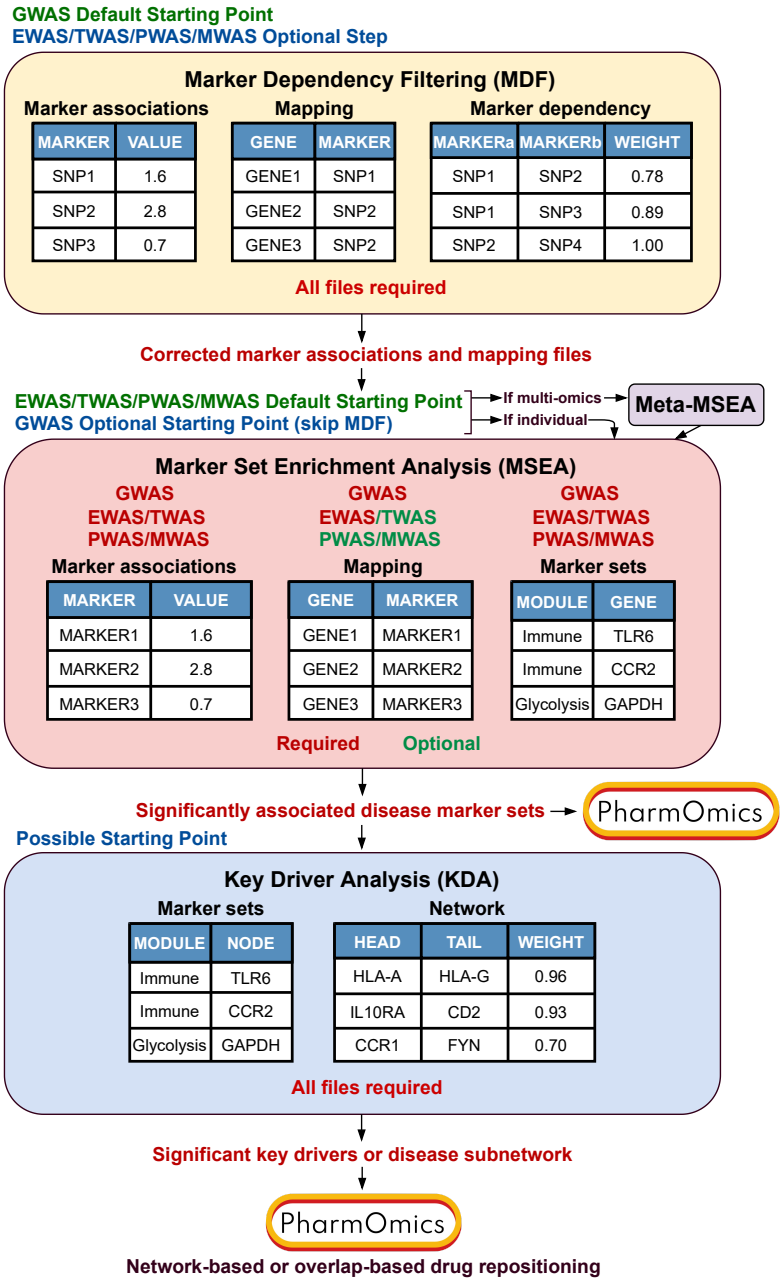

Figure 2.2. Mergeomics pipeline inputs. MDF is the default starting point for GWAS analysis and is an optional step for EWAS/TWAS/PWAS/MWAS. MDF requires marker-disease associations, a marker-gene mapping file, and a marker dependency file. Users with GWAS data can also skip MDF and run MSEA directly. MDF produces corrected marker-disease associations and marker-gene mapping files containing independent markers that are used for MSEA. For MSEA, required files for all datasets are the marker-disease associations and marker sets (pathway/modules). The marker to gene mapping file is required for GWAS and EWAS and optional for MWAS, TWAS, and PWAS. Disease-associated marker sets from MSEA can be fed into KDA which requires gene sets and a network. KDA can also be a starting point of analysis. Disease-associated gene sets from MSEA or KDs and disease subnetwork from KDA can be fed into PharmOmics drug repositioning.

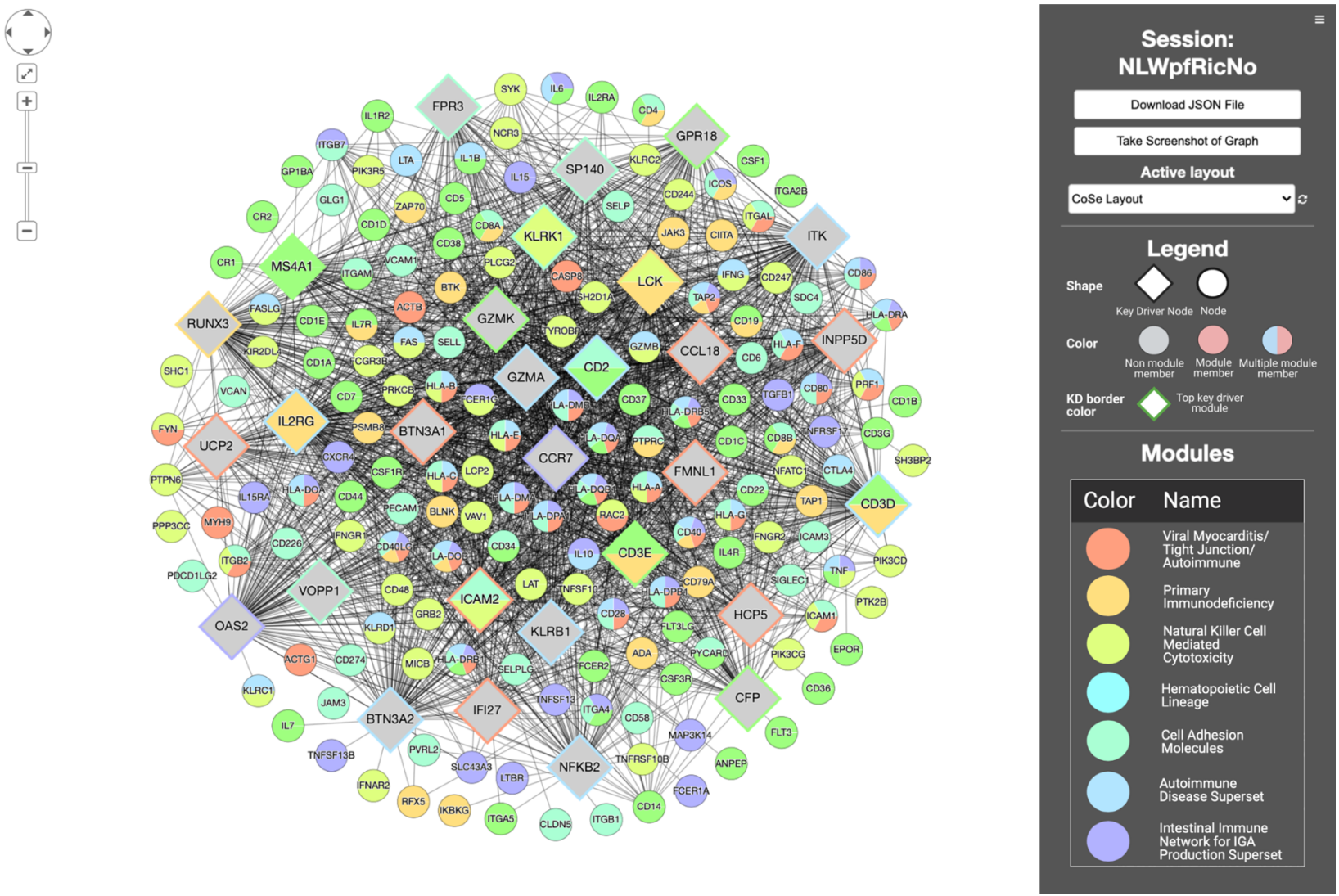

Figure 2.3. Top KDs network visualization. Screenshot of the in-browser interactive network visualization (using Cytoscape.js) directed from the KDA results page. The colors of the nodes represent member genes of a disease-associated pathway. The diamond shaped nodes represent KD genes, where the border color represents the top pathway that is regulated by the KD. If a node has multiple colors, it is part of two or more disease-associated pathways, and if a node is grey, it does not belong to the disease pathways (non-member genes) but is present in the input network.

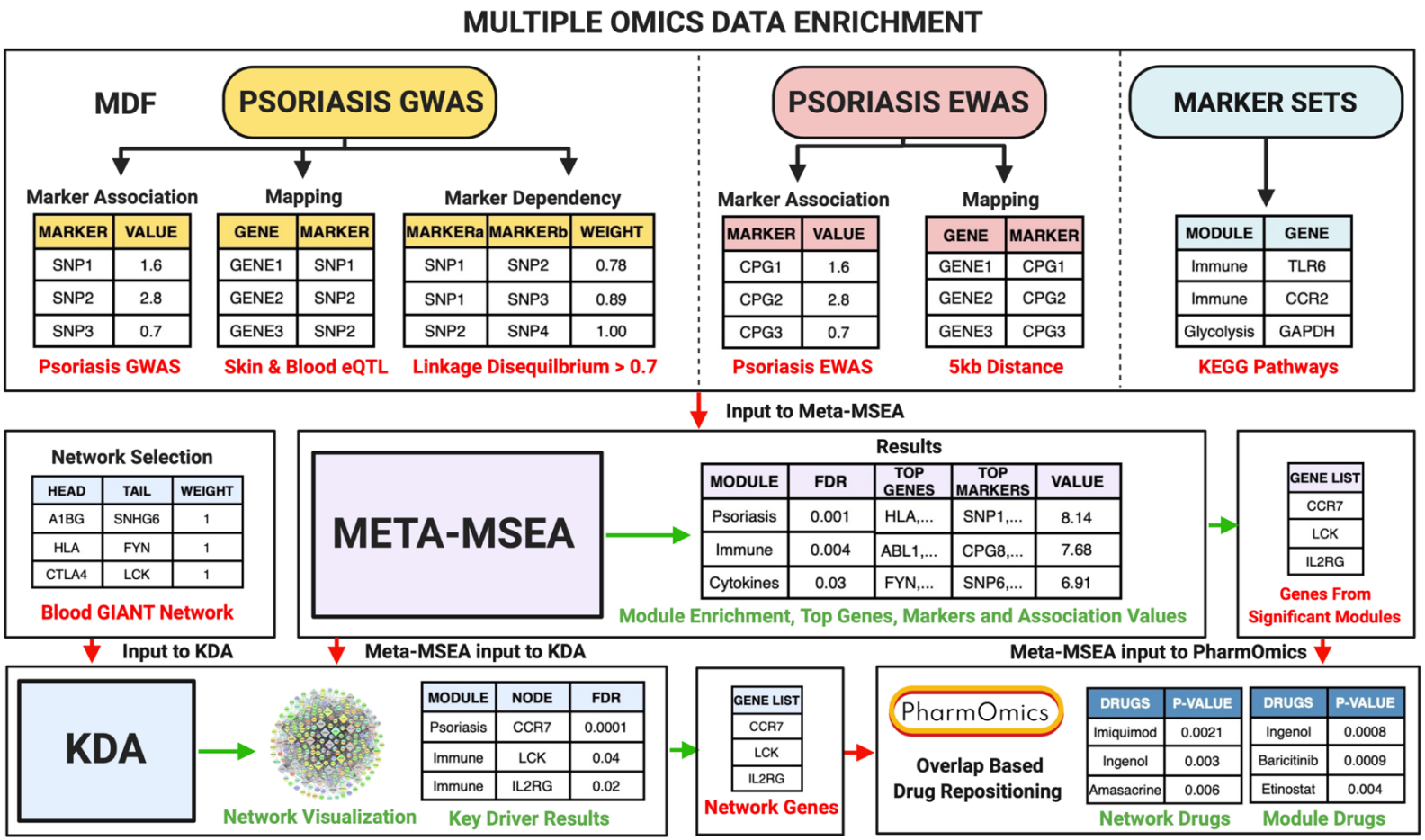

Figure 2.4. Meta-MSEA use case study overview. To showcase the function and output of the web server, we utilized multiple human psoriasis GWAS and EWAS data and ran the multiple omics data workflow (Case 3 in Figure 1, Meta-MSEA). Firstly, we uploaded the psoriasis GWAS data, mapped the SNPs to genes using a combined skin and blood eQTL file, and filtered for LD > 0.7 to remove redundant SNPs in LD. Next, we uploaded our psoriasis EWAS association datasets and mapped the CpG sites to genes based on a 5kb distance. Finally, we uploaded KEGG pathways with a psoriasis control set. Pathway enrichment results are produced, and each pathway's top genes, markers, and corresponding association values are displayed. Psoriasis associated pathways are used as input into KDA as well as PharmOmics drug repositioning (using genes from significant pathways/modules). In the KDA, along with the Meta-MSEA input, we chose the blood GIANT network option and ran the KDA providing KD results and visualization (Figure 3) and additionally utilized the network genes as an input into PharmOmics. Finally, two sets of drug repositioning results were produced using gene overlap-based drug repositioning in PharmOmics: one based on the genes of significant pathways from the Meta-MSEA results and the other based on the KDA network genes.

A

Module Results | Merge Module Results | Combined Results

SHOW 5 ENTRIES SEARCH:

Module ID	MSEA: P-Value	MSEA: FDR	Module Top Gene	Module Top Marker	Module Top Association Score
Psoriasis GWAS Catalog Control Set	2.87e-11	0.00%	|CDSN,HLA-B|PSORS1C1|MICA|MICB|N	|rs12191877|rs12191877|rs	|50.69|50.69|50.69|50.69|5(
Cytokine Cytokine Receptor Interaction	1.06e-10	0.00%	|CSF3|CXCR1|TGFBR2|EGFI	|cg21432842|cg21004129|c	|13.39|9.94|8.24|8.21|8.14|
Graft Versus Host Disease	2.21e-10	0.00%	|CDSN,HLA-B|HCG27,HLA-C,POU5F1|HLA-DRB5|HLA-DMA|AGER,HLA-DQA1,HLA-DQB1,HLA-DQB1-AS1,HLA-DQB2,HLA-DRB1,HLA-DRB6|	|rs12191877|rs2524163|rs8	|50.69|19.27|16.96|16.96|1:
Natural Killer Cell Mediated Cytotoxicity	2.27e-10	0.00%	|CDSN,HLA-B|MICB|MICA|HCG27,HLA-C,POU5F1|CD48|	|rs12191877|rs12191877|rs	|50.69|50.69|50.69|19.27|11
Hematopoietic Cell Lineage	7.31e-10	0.00%	|HLA-DRB5|CSF3|AGER,HLA-DQA1,HLA-DQB1,HLA-DQB1-AS1,HLA-DQB2,HLA-DRB1,HLA-DRB6|HLA-DRA|CD55|	|rs8365|cg21432842|rs926(	|16.96|13.39|12.05|12.05|1(

Showing 1 to 5 of 187 entries

Previous 1 2 3 4 5 ... 38 Next

B

SHOW 10 ENTRIES SEARCH:

Merge Module ID	Key Driver Node	P-Value	FDR	Module Genes	KD Subnetwork Genes	Module and Subnetwork Overlap	Fold Enrichment
Autoimmune Disease Superset	CD2	2.32e-59	6.98e-56	52	461	27	12.23
Viral Myocarditis/Tight Junction/Autoimmune	ICAM2	3.58e-59	6.98e-56	65	393	24	10.20
Autoimmune Disease Superset	NFKB2	2.17e-51	2.82e-48	52	392	21	11.19
Viral Myocarditis/Tight Junction/Autoimmune	BTN3A1	4.59e-50	4.47e-47	65	316	23	12.16
Autoimmune Disease Superset	GZMA	1.93e-49	1.50e-46	52	356	25	14.66
Viral Myocarditis/Tight Junction/Autoimmune	INPP5D	2.32e-48	1.51e-45	65	329	22	11.17
Autoimmune Disease Superset	IL2RG	3.23e-47	1.26e-44	52	464	24	10.80
Intestinal Immune Network for IGA Production Superset	OAS2	5.80e-45	1.26e-42	48	489	24	11.10
Intestinal Immune Network for IGA Production Superset	IL2RG	1.92e-44	3.56e-42	48	464	24	11.70
Hematopoietic Cell Lineage	GZMK	3.59e-36	1.70e-34	84	284	27	12.29

Showing 1 to 10 of 3,894 entries

Previous 1 2 3 4 5 ... 334 Next

C

Excel SEARCH:

Database	Method	Drug	Species	Tissue or Cell Line	Study	Dose	Treatment Duration	Jaccard Score	Odds Ratio	P value	Within Species Rank
GSE61551, GSE61552, GSE61554, GSE61555, GSE45514, GSE45551	Characteristic direction	Janus kinase inhibitor	Mus musculus	integument	PharmOmics meta	PharmOmics meta	PharmOmics meta	0.043	4.682	2.1400E-20	1.00000
L1000_CPC006_HA1E_6H	Characteristic direction	Ingenol	Homo sapiens	HA1E	In Vitro	10 uM	6hr	0.042	10.901	1.0300E-20	1.00000
GSE61551, GSE61552, GSE61554, GSE61555	Characteristic direction	Baricitinib	Mus musculus	integument	PharmOmics meta	PharmOmics meta	PharmOmics meta	0.041	4.122	1.1500E-20	0.99893
GSE80028	Characteristic direction	Imiquimod	Homo sapiens	integument	PharmOmics meta	PharmOmics meta	PharmOmics meta	0.041	14.631	1.6700E-20	0.99992
GSE80028	Characteristic direction	Toll-like receptor agonist	Homo sapiens	integument	PharmOmics meta	PharmOmics meta	PharmOmics meta	0.041	14.631	1.6700E-20	0.99992
GSE61551, GSE61552, GSE61554, GSE61555, GSE51194	Characteristic direction	Immunosuppressant	Mus musculus	integument	PharmOmics meta	PharmOmics meta	PharmOmics meta	0.040	4.055	2.2600E-18	0.99786
L1000_CPC011_PC3_24H	Characteristic direction	Pizotifen	Homo sapiens	PC3	In Vitro	10 uM	24hr	0.039	11.017	2.1800E-18	0.99984
GSE49872, E-GEOD-6196, GSE61551, GSE61552, GSE61554, GSE61555, GSE49872	Characteristic direction	Anti-inflammatory	Mus musculus	integument	PharmOmics meta	PharmOmics meta	PharmOmics meta	0.038	3.894	1.2000E-16	0.99678
drugMatrix_Affy	limma	Indomethacin	Rattus norvegicus	kidney	In Vivo	9.6 mg/kg	5d	0.038	3.216	1.2000E-08	0.99993
GSE61551, GSE61552, GSE61554, GSE61555, GSE45514, GSE45551, GSE49872	Characteristic direction	Antirheumatic	Mus musculus	integument	PharmOmics meta	PharmOmics meta	PharmOmics meta	0.038	3.838	2.2100E-16	0.99571

Showing 1 to 10 of 47,410 entries

Previous 1 2 3 4 5 ... 4741 Next

Figure 2.5. Output files from Meta-MSEA, KDA, and PharmOmics based on the case study of psoriasis outlined in Figure 2.4. Tables are interactive with pagination, search, and sort functions. Result files are downloadable from links on the webpage above the output tables (not shown). **(A)** Example Meta-MSEA output from the psoriasis use case. The table shown details the significance of association of each pathway/module and the top markers and corresponding association strengths that contributed to the module association. There are two additional tables which can be displayed by clicking on the tabs to the right of 'Module Results' at the top. The second table shows the significance and details of merged modules after merging redundant pathways (termed "Supersets"), and these nonoverlapping gene sets are used as input to KDA. The third table shows the individual significance values for each omics dataset included in this Meta-MSEA analysis of one GWAS and two EWAS of psoriasis. **(B)** Example KDA output from the psoriasis use case. The table shown records the significance of KDs, the pathways/modules that they regulate based on network topology, and details of the local subnetwork such as the number of KD subnetwork genes and number of pathway/module gene overlap with the KD subnetwork. Merged pathways/modules are represented by the term "Superset", which means they are comprised of multiple redundant (significant gene overlap) pathways. **(C)** Example PharmOmics drug repositioning output using a gene overlap-based analysis between disease pathways and drug signatures. Gene overlap-based drug repositioning queries all tissue- and species-specific meta-analyzed and dose/time segregated gene signatures of drugs in our PharmOmics database as well as all L1000 drug signatures. The table shown gives the dataset source of the drug signature, the method of differential gene expression analysis, details of the drug study including species, tissue or cell line, whether the study was done in vitro or in vivo, the dose and time regimen, the Jaccard score, and statistical significance of the gene overlap between the input psoriasis related genes from Meta-MSEA and the drug signatures.

Chapter 3. Gene Networks and Pathways for Plasma Lipid Traits via Multi-tissue Multi-omics Systems Analysis

3.1 Introduction

Lipid metabolism is vital for organisms as it provides energy as well as essential materials such as membrane components and signaling molecules for basic cellular functions. Lipid dysregulation is closely related to many complex human diseases, such as atherosclerotic cardiovascular disease [121], Alzheimer's disease [122, 123], type 2 diabetes (T2D) [124], and cancers [125]. The notion of targeting lipid metabolism to treat human diseases has been reinforced by the fact that many disease-associated genes and drug targets (e.g., *HMGCR* as the target of statins and *PPARA* as the target of fibrates) are involved in lipid metabolic pathways [126-128].

Accumulating evidence supports that plasma lipids are complex phenotypes influenced by both environmental and genetic factors [129, 130]. Heritability estimates for main plasma lipids are high (e.g. ~70% for low density lipoprotein cholesterol [LDL] and ~55% for high density lipoprotein cholesterol [HDL]) [131], indicating that DNA sequence variation plays an important role in explaining the inter-individual variability in plasma lipid levels. Indeed, genome-wide association studies (GWAS) have pinpointed a total of 386 genetic loci, captured in the form of single nucleotide polymorphisms (SNPs) associated with lipid phenotypes [132-136]. For example, the most recent GWAS on lipid levels identified 118 loci that had not previously been associated with lipid levels in humans, revealing a daunting genetic complexity of blood lipid traits [136].

However, there are several critical issues that cannot be easily addressed by traditional GWAS analysis. First, even very large GWAS may lack statistical power to identify SNPs with small effect sizes and as a result the most significant loci only explain a limited proportion of the

genetic heritability, for example, 17.2 – 27.1% for lipid traits [137]. Second, the functional consequences of the genetic variants and the causal genes underlying the significant genetic loci are often unclear and await elucidation. To facilitate functional characterization of the genetic variants, genetics of gene expression studies [17, 138] and the ENCODE efforts [139] have documented tissue- or cell-specific expression quantitative trait loci (eQTLs) and functional elements of the human genome. These studies provide the much-needed bridge between genetic polymorphisms and their potential molecular targets. Third, the molecular mechanisms that transmit the genetic perturbations to complex traits or diseases, that is, the cascades of molecular events through which numerous genetic loci exert their effects on a given phenotype, remain elusive. Biological pathways that capture functionally related genes involved in molecular signaling cascades and metabolic reactions, and gene regulatory networks formed by regulators and their downstream genes can elucidate the functional organization of an organism and provide mechanistic insights [140]. Indeed, various pathway- and network-based approaches to analyzing GWAS datasets have been developed [17, 141-143] and demonstrated to be powerful to capture both the missing heritability and the molecular mechanisms of many human diseases or quantitative phenotypes [17, 142, 144, 145]. For these reasons, integrating genetic signals of blood lipids with multi-tissue multi-omics datasets that carry important functional information may provide a better understanding of the molecular mechanisms responsible for lipid regulation as well as the associated human diseases.

In this study, we apply an integrative genomics framework to identify important regulatory genes, biological pathways, and gene subnetworks in relevant tissues that contribute to the regulation of four critical blood lipid traits, namely TC, HDL, LDL, and TG. We combine the GWAS results from the Global Lipids Genetics Consortium (GLGC) with functional genomics

data from a number of tissue-specific eQTLs and the ENCODE project, and gene-gene relationship information from biological pathways and data-driven gene network studies. The integrative framework is comprised of four main parts (**Figure 4.1**): 1) Marker Set Enrichment Analysis (MSEA) where GWAS, functional genome, and pathways or co-regulated genes are integrated to identify lipid-related functional units of genes, 2) merging and trimming of identified lipid gene sets, 3) key driver analysis (KDA) to pinpoint important regulatory genes by further integrating gene regulatory networks, and 4) validation of key regulators using genetic perturbation experiments and *in silico* evidence. This integrated systems biology approach enables us to derive a comprehensive view of the complex and novel mechanisms underlying plasma lipid metabolism.

3.2 Results

Identification of pathways and gene co-expression modules associated with lipid traits

To asses biological pathway enrichment for the four lipid traits with GLGC GWAS, we curated a total of 4532 gene sets including 2705 tissue-specific co-expression modules (i.e., highly co-regulated genes based on tissue gene expression data) and 1827 canonical pathways from Reactome, Biocarta and the Kyoto Encyclopedia of Genes and Genomes (KEGG). These gene sets were constructed as data- and knowledge-driven functional units of genes. Four predefined positive control gene sets for HDL, LDL, TC, and TG were also created based on candidate genes curated from the GWAS catalog [146]. To map potential functional SNPs to genes in each gene set, tissue-specific eQTLs, ENCODE functional genomics information, and chromosomal distance-based mapping were used (details in Methods). Tissue-specific eQTL sets were obtained from the GTEx database from studies on human adipose tissue, liver, brain, blood, and human aortic endothelial cells (HAEC), and a total of nine SNP-gene mapping methods were created. The

liver and adipose tissues have established roles in lipid regulation, whereas the other tissues are included for comparison.

Integrating the datasets mentioned above using MSEA, we identified 65, 86, 90, and 92 gene sets whose functional genetic polymorphisms showed significant association with HDL, LDL, TC, and TG, respectively, in GLGC GWAS (FDR < 10%; **Supplemental Table S4.1)**. The predefined positive controls for the four lipid traits were among the top signals for their corresponding traits (**Table 4.1**), indicating that our MSEA method is sensitive in detecting true lipid trait-associated processes. Compared with other tissues, more pathways were captured when using liver and adipose eSNPs to map GWAS SNPs to genes (**Supplemental Table S4.1)**. For example, 56 out of the 86 LDL-associated pathways were found when liver and adipose eSNPs were used in our analysis. These results confirmed the general notion that liver and adipose tissue play critical roles in regulating plasma lipids, leading us to focus the bulk of our analysis on these two tissues, with the remaining tissues serving as a supplement.

Among the significant gene sets, 39 were shared across the four lipid traits. These gene sets represented the expected lipid metabolic pathways as well as those that are less known to be associated with lipids, such as 'antigen processing and presentation', 'cell adhesion molecules (CAMs)', 'visual phototransduction', and 'IL-5 signaling pathway' (summary in **Table 4.1**; details in **Supplemental Table S4.1)**. We broadly classified the common gene sets detected into 'positive controls', 'lipid metabolism', 'interferon signaling', 'autoimmune/immune activation', 'visual transduction', and 'protein catabolism' (**Table 4.1**).

Beside the common gene sets described above, we also detected 18, 5, 6, and 17 trait-specific pathways/modules for HDL, LDL, TC, and TG, respectively (**Table 4.2**; **Supplement Table S4.1**), suggesting trait-specific regulatory mechanisms. Among the 18 pathways for HDL

were ‘cation-coupled chloride transporters’, ‘glycerolipid metabolism’ and ‘negative regulators of RIG-I/MDA5 signaling’ across analyses using different tissue eSNP mapping methods, ‘alcohol metabolism’ from brain-based analysis, ‘packaging of telomere ends’ in adipose tissue, ‘glutathione metabolism’ in liver, and ‘cobalamin metabolism’ and ‘taurine and hypotaurine metabolism’ in both adipose and liver-based analyses. LDL-specific pathways included the ‘platelet sensitization by LDL’ pathway and a liver co-expression module related to cadherin. TC-specific pathways included ‘valine, leucine and isoleucine biosynthesis’ across tissues and ‘wound healing’ in the brain-based analysis. When looking at the TG-specific pathways, gene sets associated with ‘cellular junctions’ were consistent across tissues whereas ‘insulin signaling’ and complement pathways were exclusively seen in adipose tissue-based analysis.

Replication of lipid-associated pathways using additional dataset and method

To replicate our results from the analysis of GLGC GWAS datasets, we utilized an additional lipid genetic association dataset based on a MetaboChip lipid association study [135] which involved individuals independent of those included in GLGC. The gene sets detected using this independent dataset highly overlapped with those from the GLGC dataset (**Table 4.1**; **Figure 4.2**; overlapping p values $< 10^{-20}$ by Fisher’s exact test). We also utilized a different pathway analysis method iGSEA [147] and again many of the gene sets were found to be reproducible (**Table 4.1**; **Figure 4.3**; overlapping p values $< 10^{-20}$).

Construction of non-overlapping gene supersets for lipid traits

As the knowledge-based pathways and data-driven co-expression modules used in our analysis can converge on similar functional gene units, some of the lipid-associated gene sets have redundancies. We therefore merged overlapping pathways to derive independent, non-overlapping gene sets associated lipid traits. For the 39 shared pathways/co-expression modules across the four

lipid traits described earlier, we merged and functionally categorized them into five independent supersets (**Table 4.1; Table 4.3**). For the significant gene sets for each lipid trait, we merged them into 17, 16, 18, and 14 supersets for HDL, LDL, TC, and TG, respectively (**Table 4.3; Supplemental Table S4.2**), and confirmed that the merged supersets still showed significant association with the corresponding lipid traits in a second round of MSEA ($p < 0.05$ after Bonferroni correction for the number of supersets tested; **Table 4.3**).

Identification of central regulatory genes in the lipid-associated supersets

Subsequently, we performed a key driver analysis (KDA; **Figure 4.1**) to identify potential regulatory genes or key drivers (KDs) that may regulate genes associated with each lipid trait using Bayesian networks constructed from genetic and gene expression datasets of multiple tissues (detailed in Methods; full KD list in **Supplemental Table S4.3**). The top adipose and liver KDs for the shared supersets of all four lipid traits and the representative Bayesian subnetworks are shown in **Figure 4.2**.

In adipose tissue (**Figure 4.2A**), the top KDs for the 'lipid metabolism' subnetwork include well-known lipoproteins and ATP-binding cassette (ABC) family members that are responsible for lipid transport, such as *APOF*, *APOA5* and *ABCB11*. We also found several KDs that are less known to be associated with lipid metabolism, particularly *F2* (coagulation Factor II or thrombin). For the 'autoimmune/immune activation' subnetwork, *CD86*, *HCK*, and *HLA-DMB* were identified as KDs. *PSMB9* was a KD for the 'protein catabolism' subnetwork, whereas *NUP210* is central for the 'interferon signaling' subnetwork. Moreover, the *SYK* gene is a shared KD between 'lipid metabolism' and 'autoimmune/immune activation'.

In the liver (**Figure 4.2B**), the top KDs for the 'lipid metabolism' subnetwork are enzymes involved in lipid and cholesterol biosynthesis and metabolism, such as *FADS1* (fatty acid

desaturase 1), *FDFT1* (farnesyl-diphosphate farnesyltransferase 1), *HMGCS1* (3-hydroxy-3-methylglutaryl-CoA synthase 1), and *DHCR7* (7-dehydrocholesterol reductase). We also identified more KDs for the 'interferon signaling' subnetwork in the liver compared to the adipose tissue, with *MX1*, *MX2*, *ISG15*, *IFI44*, and *EPSTI1* being central to the subnetwork. Similar to the adipose network, *PSMB9* and *HLA-DMB* were also identified as KDs for 'protein catabolism' and 'autoimmune/immune activation' subnetworks in liver, respectively. We did not detect key driver genes for the 'visual transduction' subnetwork in either tissue, possibly as a result that the networks of liver and adipose tissues did not capture gene-gene interactions important for this subnetwork.

In addition to the KDs for the subnetworks shared across lipid traits as discussed above, we identified tissue-specific KDs for individual lipid traits (**Supplemental Table S4.3**). In adipose, *PANK1* and H2B histone family members were specific for the HDL subnetworks (**Figure 4.3A**); *HIPK2* and *FAU* were top KDs for the LDL subnetworks (**Figure 4.3B**); genes associated with blood coagulation such as *KNG1* and *FGL1* were KDs for the TC and TG subnetworks (**Figure 4.3C-4.3D**). Interestingly, genes related to insulin resistance, *PPARG* and *FASN*, were KDs for both HDL and TG subnetworks. Similarly, trait specific KDs and subnetworks were also detected in the liver; 37 KDs were identified for the TG subnetwork including *ALDH3B1* and *ORM2*, whereas *AHSG*, *FETUB*, *ITIH1*, *HP*, and *SERPINC1* were KDs found in the LDL subnetwork. We note that most of the KDs are themselves not necessarily GWAS hits but are surrounded by significant GWAS genes. For example, gene *F2* is centered by many GWAS hits in the adipose subnetwork (*APOA4*, *APOC3*, *APOA5*, *LIPC*, etc.; **Figure 4.2; Figure 4.4**). The observation of GWAS hits being peripheral nodes in the network is consistent with previous findings from our group and others [148-153], and again supports that important regulators may not necessarily harbor common variations due to evolutionary constraints.

Experimental validation of *F2* KD subnetworks in 3T3-L1 and C3H10T1/2 adipocytes

Taking into account that the *F2* gene is surrounded by various significant GWAS hits within its subnetwork, we aimed to validate the role of the *F2* gene subnetwork in lipid regulation through siRNA-mediated knockdown experiments in two adipocyte cell lines (3T3-L1 and C3H10T1/2) to ensure reproducibility and robustness of our results. We found that *F2* gene expression was low in preadipocytes for both cells lines, but gradually increased during adipogenesis. In fully differentiated adipocytes between day 8 and day 10, *F2* gene expression level was higher than preadipocytes by 12-fold and 6-fold for 3T3-L1 and C3H10T1/2 lines, respectively (**Figure 4.5A; 4.5B)**. When treated with *F2* siRNA, both adipocyte cell lines showed a significant decrease ($p < 0.01$) in lipid accumulation based on Oil red O staining, as compared with controls treated with scrambled siRNA (**Figure 4.5C; 4.5D**). Subsequently, we tested the effect of *F2* gene siRNA knockdown on ten neighbors of the *F2* gene in the adipose network (selected from **Figure 4.2A**). With 60% knockdown efficiency of *F2* siRNA in the 3T3-L1 adipocytes, seven *F2* network neighbors (*Abcb11, Apoa5, Apof, Fabp1, Lipc, Gc* and *Proc*) exhibited significant changes in expression levels (**Figure 4.5E**). With 74 % knockdown efficiency of *F2* in C3H10T1/2 adipocytes, six *F2* network neighbors (*Abcb11, Apoa5, Apof, Fabp1, Lipc,* and *Plg*) showed significant changes in expression levels (**Figure 4.5F**). Several of these genes are involved in lipoprotein transport and fatty acid uptake. In contrast, none of the four negative controls (random genes not in F2 network neighborhood) showed significant changes in their expression levels for 3T3-L1 cell line. However, one negative control gene (*Snrpb2*) did change in the C3H10T1/2 cell line. These results overall support our computational predictions on the structures of *F2* gene subnetworks.

Next, we measured expression levels of genes related with adipogenesis (*Pparg*, *Cepba*, *Srepb1*, *Fasn*), lipolysis (*Lipe*), fatty acid transport (*Cd36*, *Fabp4*), and other adipokines following *F2* siRNA treatment. We found no change in the expression of most of the tested genes, with the exception of *Fasn* (in C3H10T1/2), important in the formation of long chain fatty acids, and *Cd36* (in both 3T3-L1 and C3H10T1/2), which encodes fatty acid translocase facilitating fatty acid uptake. *Cd36* expression was decreased by 15 % in 3T3-L1 cells (**Figure 4.5G**) and 35% in C3H10T1/2 cells (**Figure 4.5H**) ($p < 0.05$) and *Fasn* expression was decreased by 25% (**Figure 4.5H**) ($p < 0.01$) in C3H10T1/2 cells compared to control. The decreases in *Cd36* and *Fasn* after *F2* knockdown suggest that fatty acid synthesis and uptake by adipocytes are compromised, which could contribute to alterations in circulating lipid levels.

We subsequently measured the lipid contents within the cells and in the media of C3H10T1/2 adipocytes. Following *F2* siRNA treatment, we found significant decreases in total intracellular lipid levels (cTotal Lipid), total cholesterol (cTC), and unesterified cholesterol (cUC), as well as a non-significant trend for decreased triglycerides (cTG) (**Figure 4.5I**). By contrast, in the culture media, there were significant increases in the total lipid levels (mTotal Lipid) and triglycerides (mTG) following *F2* siRNA treatment (**Figure 4.5J**). These results support that *F2* knockdown led to decreased intracellular lipids and increased extracellular lipids, agreeing with the overall decreased expression of *F2* network neighbor genes involved in lipid transport and uptake.

The association between the lipid subnetworks and human diseases

Epidemiological studies consistently show that plasma lipids are closely associated with human complex diseases. For example, high TC and LDL levels are associated with increased risk of cardiovascular disease (CVD). Here, we examined the association between the lipid subnetworks identified in our study and four human complex diseases, namely, Alzheimer's

disease, CVD, T2D, and cancer (Materials and Methods). We found that the gene supersets identified for each lipid traits were significantly enriched for GWAS candidate genes reported by GWAS catalog for the four diseases at Bonferroni-corrected $p < 0.05$ (**Figure 4.6; Supplemental Table S4.4**). The superset 'lipid metabolism', which was shared across lipid traits, was associated with Alzheimer's disease and CVD. When trait-specific subnetworks were considered, those associated with TC, LDL, and TG had more supersets associated with CVD compared to those associated with HDL, a finding consistent with recent reports [135, 154, 155]. In addition, supersets of each lipid trait, except HDL, were also found to be significantly associated with cancer, whereas supersets associated with HDL, LDL, and TG but not TC, were linked to T2D.

3.3 Discussion

To gain comprehensive insights into the molecular mechanisms of lipid traits that are important for numerous common complex diseases, we leveraged the large volume of genomic datasets and performed a data-driven multi-omics study combining genetic association signals from large lipid GWAS, tissue-specific eQTLs, ENCODE functional data, known biological pathways, and gene regulatory networks. We identified diverse sets of biological processes, guided by their tissue-specific gene-gene interactions, to be associated with individual lipid traits or shared across lipid traits. Many of the lipid associated gene sets were significantly linked to multiple complex diseases including CVD, T2D, cancer, and Alzheimer's disease. More importantly, we elucidated tissue-specific gene-gene interactions among the gene sets and identified both well characterized and novel KDs for these lipid-associated processes. We further experimentally validated a novel adipose lipid regulator, *F2*, in two different adipocyte cell lines. Our results offer new insight into the molecular regulation of lipid metabolism, which would not have been possible without the systematic integration of diverse genetic and genomic datasets.

We identified shared pathways associated with all four lipid traits, including 'lipid metabolism' and 'autoimmune/immune activation', which have been consistently linked to lipid phenotypes, as well as additional pathways such as 'interferon signaling', 'protein catabolism', and 'visual transduction'. Interferon factors have previously been linked to lipid storage attenuation and differentiation in human adipocytes [156]. Protein catabolism has only recently been identified to be important in regulating lipid metabolism through the PSMD9 protein, which had no previously known function but was shown to cause significant alterations in lipid abundance in both a gain of function and loss of function study in mice [157]. The 'visual transduction' superset contains retinol-binding proteins, which are carrier proteins involved retinol transport, and play key roles in gene expression regulation and developmental processes [158]. 'visual transduction' also shares lipoprotein genes with 'lipid metabolism', suggesting that retinol-related signal transduction is intimately linked to lipoprotein transport and hence plasma lipid levels.

Furthermore, our results indicate that the trait-specific supersets are tissue-specific. For example, most TG-specific pathways were found to be significant when adipose eSNPs were used, and complement and insulin signaling pathways in the adipose tissue were specific for TG. This is in line with adipose tissue functioning as the major storage site for TG and the regulatory role of immune system and insulin signaling in adipocyte functions and fat storage [159]. We also found five HDL-specific pathways, most of which are associated with glucose, lipid, and amino acid metabolism, and were signals derived from liver eSNPs. As HDL acts as the major vehicle for transporting cholesterol to the liver for excretion and catabolism, the critical role of the liver as well as the connections between major metabolic pathways in HDL regulation is recapitulated by our analysis. Interestingly, the TC-specific pathways can be only found when brain eSNPs are

used. While the brain accounts for 2% of body weight, it contains 23% of TC in the body [160] and deregulated cholesterol trafficking appears to be involved in the pathogenesis of neurodegenerative diseases, such as Parkinson's and Alzheimer's disease [161]. These tissue- and trait-specific pathways or processes support the unique features of each lipid species and point to tissue-specific targeting strategies to modulate levels of individual lipid traits and the associated diseases.

In addition to detecting trait- and tissue-specific causal pathways for the lipid traits, our study attempted to delineate the interactions between lipid genes and pathways through gene network analysis. Indeed, the tissue-specific gene networks revealed in our study highlight the regulatory connections between lipid genes and pathways, and thus put individual genes in a broader context. The identification of KDs in a network is essential for uncovering key regulatory components and for identifying drug targets and biomarkers for complex diseases [143, 162]. Here, we adopted data-driven Bayesian gene regulatory networks that combine various genomic data [163] to detect the central genes in plasma lipid regulation. The power of this data-driven objective approach has been demonstrated recently [9, 143, 151, 152, 164, 165] and is again supported in this study by the fact that many KDs detected are known regulators for lipids or have served as effective drug targets based on the DrugBank database [166]. For instance, for the shared 'lipid metabolism' subnetwork, four top KDs (*ACAT2*, *ACSS2*, *DHCR7*, and *FADS1*), are targeted by at least one FDA approved anti-cholesteremic drug. Another KD, *HMGCS1*, is a rate-limiting enzyme of cholesterol synthesis, and is considered a promising drug target in lipid-associated metabolic disorders [167]. These lines of evidence lead us to speculate the other less-studied KDs are also important for lipid regulation.

Among the top network KDs predicted, several including *F2, KLKB1* and *ANXA4* are involved in blood coagulation. A previous study revealed polymorphisms in the anticoagulation genes modify the efficacy of statins in reducing risk of cardiovascular events [168], which in itself is not surprising. However, the intimate relationship between a coagulation gene *F2* and lipid regulation predicted by our analysis is intriguing (**Figure 4.5)**. We found that the partner genes in the adipose *F2* subnetwork tend to be differentially expressed after *F2* knockdown in both 3T3-L1 and C3H10T1/2 adipocytes, with several of the altered genes (*Apoa5*, *Apof*, *Abcb11*, *Fabp1, Fasn* and *Cd36*) closely associated with cholesterol and fatty acid transport and uptake. We further observed that *F2* knockdown affects lipid storage in adipocytes, with intracellular lipid content decreasing and extracellular lipid content in the media increasing. Interestingly, *F2* expression level is low in preadipocytes and only increases during the late phase of adipocyte differentiation. Our findings support a largely untapped role of *F2* in lipid transport and storage in adipocytes and provide a novel target in the *F2* gene.

In addition to the shared KDs such as *F2* for different lipids, it may be also of value to focus on the trait-specific KDs as numerous studies have revealed these lipid phenotypes play different roles in many human diseases. For example, LDL and TC are important risk factors for CVD [169] and TG has been linked to T2D [170], while the role of HDL in CVD has been controversial [171]. We detected 37 genes as TG-specific KDs in liver regulatory subnetworks. Among these, *CP* (ceruloplasmin) and *ALDH3B1* (aldehyde dehydrogenase 3 family, member B1) were clinically confirmed to be associated with T2D [172, 173] while most of the other genes such as *DHODH* and *ANXA4* were less known to be associated with TG and thus may serve as novel targets. In adipose tissue, genes important for insulin resistance and diabetes such as *PPARG* and *FASN* were found to be KDs for TG, further supporting the connection between TG and diabetes.

Additionally, *FASN* has been implicated as a KD in numerous studies for non-alcoholic fatty liver disease [153, 165, 174], again highlighting the importance of this gene in common metabolic disorders.

We acknowledge some potential limitations to our study. First, the GWAS datasets utilized are not the most recently conducted and therefore provides the possibility of not capturing the full array of unknown biology. However, despite this our results are consistent with the biology found more recently including overlapping signals in pathways for chylomicron-mediated lipid transport and lipoprotein metabolism [175] as well as more novel findings such as visual transduction pathways. In addition, one of our key drivers *KLKB1*, which was not found to be a GWAS hit in the dataset we utilized, has since been found to pass the genome wide significance threshold in more recent larger GWAS and is a hit on apolipoprotein A-IV concentrations, which is a major component of HDL and chylomicron particles important in reverse cholesterol transport [176]. This further exemplifies the robustness of our integrative network approach to find key genes important to disease pathogenesis even when smaller GWAS were utilized.

3.4 Conclusion

In summary, we used an integrative genomics framework to leverage a multitude of genetic and genomic datasets from human studies to unravel the underlying regulatory processes involved in lipid phenotypes. We not only detected shared processes and gene regulatory networks among different lipid traits, but also provide comprehensive insight into trait-specific pathways and networks. The results suggest there are both shared and distinct mechanisms underlying very closely related lipid phenotypes. The tissue-specific networks and KDs identified in our study shed light on molecular mechanisms involved in lipid homeostasis. If validated in additional population genetic and mechanistic studies, these molecular processes and genes can be used as novel targets

for the treatment of lipid-associated disorders such as CVD, T2D, Alzheimer's disease and cancers.

3.5 Methods

GWAS of lipid traits

The experimental design, genotyping, and association analyses of HDL, LDL, TC, and TG were described previously [132]. The dataset used in this study is comprised of > 100,000 individuals of European descent (sample size 100,184 for TC, 95,454 for LDLC, 99,900 for HDLC and 96,598 for TG), ascertained in the United States, Europe, or Australia. More than 906,600 SNPs were genotyped using Affymetrix Genome-Wide Human SNP Array 6.0. Imputation was further carried out to obtain information for up to 2.6 million SNPs using the HapMap CEU (Utah residents with ancestry from northern and western Europe) panel. SNPs with minor allele frequency (MAF) < 1% were removed. Finally, a total of ~ 2.6 million SNPs tested for association with each of the four lipid traits were used in our study.

Genetic association study of lipid traits using MetaboChip

The experimental design, genotyping, and association analyses of the lipid MetaboChip study were described previously [177]. The study examined subjects of European ancestry, including 93,982 individuals from 37 studies genotyped with the MetaboChip array, comprised of 196,710 SNPs representing candidate loci for cardiometabolic diseases. There was limited overlap between the individuals involved in GWAS and those in MetaboChip.

Knowledge-based biological pathways

We included canonical pathways from the Reactome (version 45), Biocarta, and the Kyoto Encyclopedia of Genes and Genomes (KEGG) databases [178, 179]. In addition to the curated pathways, we constructed four positive control pathways based on known lipid-associated loci (p

$< 5.0 \times 10^{-8}$) and candidate genes from the GWAS Catalog [180]. These gene sets were based on 4, 11, 13, and 13 studies for TC, TG, LDL, and HDL, respectively (full lists of genes in each positive control sets are in **Supplemental Table S4.5**) and serve as positive controls to validate our computational method.

Data-driven modules of co-expressed genes

Beside the canonical pathways, we used co-expression modules that were derived from a collection of genomics studies (**Supplement Table S4.6**) of liver, adipose tissue, aortic endothelial cells (HAEC), brain, blood, kidney, and muscle [181-190]. A total of 2706 co-expression modules were used in this study. Although liver and adipose tissue are likely the most important tissues for lipid regulation, we included the other tissue networks to confirm whether known tissue types for lipids could be objectively detected and whether any additional tissue types are also important for lipids.

Mapping SNPs to genes

Three different mapping methods were used in this study to link SNPs to their potential target genes.

Chromosomal distance-based mapping

First, we used a standard distance-based approach where a SNP was mapped to a gene if within 50 kb of the respective gene region. The use of ± 50 kb to define gene boundaries is commonly used in GWAS.

eQTL-based mapping

The expression levels of genes can be seen also as quantitative traits in GWAS. Hence, it is possible to determine eQTLs and the expression SNPs (eSNPs) within the eQTLs that provide a functionally motivated mapping from SNPs to genes. Moreover, the eSNPs within the eQTL are specific to the tissue where the gene expression was measured and can therefore provide

mechanistic clues regarding the tissue of action when intersected with lipid-associated SNPs. Results from eQTL studies in human adipose tissue, liver, brain, blood, and HAEC were used in this study [181, 183, 184, 191-199]. We included both *cis*-eSNPs (within 1 Mb distance from gene region) and *trans*-eSNPs (beyond 1 Mb from gene region), at a false discovery rate < 10%.

ENCODE-based mapping

In addition to eQTLs and distance-based SNP-gene mapping approaches, we integrated functional data sets from the Regulome database [139], which annotates SNPs in regulatory elements in the *Homo sapiens* genome based on the results from the ENCODE studies [200].

Nine unique combinations of SNP-gene mapping

Using the above three mapping approaches, we derived nine unique sets of SNP-gene mapping. These are: eSNP adipose, eSNP liver, eSNP blood, eSNP brain, eSNP HAEC, eSNP all (i.e., combining all the tissue-specific eSNPs above), Distance (chromosomal distance-based mapping), Regulome (ENCODE-based mapping), and Combined (combining all the above methods).

Removal of SNPs in linkage disequilibrium

We observed a high degree of linkage disequilibrium (LD) in the eQTL, Regulome, and distance-based SNPs, and this LD structure may cause artifacts and biases in the downstream analysis. For this reason, we devised an algorithm to remove SNPs in LD while preferentially keeping those with a strong statistical association with lipid traits. Technical details are available in Supplementary Methods. We chose a LD cutoff ($R^2 < 0.5$) to remove redundant SNPs in high LD.

Marker Set Enrichment Analysis (MSEA)

We applied a modified MSEA method [143, 201] to find pathways/co-expressed modules associated with lipid traits (**Supplemental Methods**). False discovery rates (FDR) were estimated with the method by Benjamini and Hochberg [202]. Pathways or co-expression modules with FDR

< 10% were considered statistically significant. MSEA were applied to both the GLGC GWAS dataset and the MetaboChip dataset. The combined FDR from these two datasets was expected to be < 1% (10% * 10% = 1%).

Comparison between MSEA and other computational method

To ensure that the pathway results from MSEA are reproducible, we used the improved gene-set-enrichment analysis approach (iGSEA) [147]. In the iGSEA analysis, we generated gene sets using the same canonical pathways and co-expression modules in MSEA. The SNPs were mapped to genes using the default settings of iGSEA. For each given gene set, significance proportion-based enrichment score was calculated to estimate the enrichment of genotype–phenotype association. Then, iGSEA performed label permutations to calculate nominal P-values to assess the significance of the pathway-based enrichment score and FDR to correct multiple testing, with FDR < 0.25 (default setting) regarded as significant pathways. Considering that MSEA and iGSEA were independent, the combined FDR from these two methods of analysis was expected to be < 5% (10% x 25% = 2.5%).

Construction of independent supersets and confirmation of lipid association

Because the pathways or co-expression modules were collected from multiple sources, there were overlapping or nested structures among the gene sets. To make the results more meaningful, we constructed relatively independent supersets that captured the core genes from groups of redundant pathways and co-expression modules (**Supplemental Methods**). After merging, we annotated each superset based on function enrichment analysis of the known pathways from the Gene Ontology and KEGG databases (P < 0.05 in Fisher's exact test after Bonferroni correction). The

supersets were given a second round of MSEA to confirm their significant associated with lipids using $P < 0.05$ after Bonferroni correction as the cutoff.

Key driver analysis (KDA)

We adopted a previously developed KDA algorithm [163, 164, 203] of gene-gene interaction networks to the lipid-associated supersets in order to identify the key regulatory genes (**Figure 4.1**). In the study, we included Bayesian gene regulatory networks from diverse tissues, including adipose tissue, liver, blood, brain, kidney and muscle [181-189]. A key driver was defined as a gene that is directionally connected to a large number of genes from a lipid superset, compared to the expected number for a randomly selected gene within the Bayesian network (details in **Supplemental Methods**). The MSEA, merging, and KDA were performed using R.

Enrichment analysis of lipid-associated subnetworks in human complex diseases

We collected disease susceptibility genes from GWAS Catalog with GWAS P<10E-5 for four human complex diseases, including cardiovascular diseases ('myocardial infarction', 'myocardial infarction (early onset)', 'coronary artery calcification', and 'coronary heart disease'), Alzheimer's disease, type 2 diabetes, and cancer ('colon cancer', 'breast cancer', 'pancreas cancer', 'prostate cancer', and 'chronic lymphocytic leukemia'). Fisher's exact test was used to explore the enrichment of genes in the lipid-associated subnetworks in the disease gene sets. Bonferroni-corrected $p < 0.05$ was considered significant.

Validation of *F2* in adipocyte functions via *F2* siRNA transfection in 3T3-L1 and C3H10T1/2 adipocyte cell lines

The mouse preadipocytes 3T3-L1 and C3H10T1/2 cells were obtained from ATCC and maintained and differentiated to adipocytes according to the manufacturer's instruction. For knockdown experiments, 3 predesigned siRNAs targeting *F2* gene (sequences in **Supplemental Table S3**;

GenePharma, Paramount, CA) were tested and the most effective one was selected for the experiment (**Supplemental Figure S1**). We first measured *F2* expression during adipocyte differentiation and found increased *F2* expression on days 8-10 in 3T3-L1 and days 6-10 in C3H10T1/2 during differentiation, which helped inform on the timing of siRNA transfection in these cell lines. 3T3-L1 adipocytes were transfected with 50 nM of *F2* siRNA using Lipofectamin 2000 on day 7 (D7) of differentiation, a day before *F2* increase. Followed by 72 hrs of siRNA treatment, adipocytes were processed for Oil red O staining of lipids and Real-time qPCR for select genes. C3H10T1/2 adipocytes were transfected with 50 nM of *F2* siRNA using Lipofectamin 2000 on day 5 (D5) and day 7 (D7), and adipocytes were processed on day 9 (D9) for Oil red O staining of lipids, Real-time qPCR for select genes, and quantitative lipid assays. As control, 50 nM of scrambled siRNA (GenePharma, Paramount, CA) was transfected at the same time points as the *F2* siRNA in the two cell lines. To determine changes in lipid accumulation, adipocytes were stained by Oil red O stain solution. After obtaining images, Oil red O was eluted in isopropyl alcohol and optical density (OD) values were measured at 490 nm.

RNA extraction and Real-time qPCR

Total RNA was extracted from the adipocytes (Zymo Research, Irvine, CA), and RNA was reverse transcribed using cDNA Reverse Transcription Kit (Thermo Scientific, Madison, WI, USA), Real-time qPCR for select network and non-network genes was performed using the primers shown in **Supplemental Table S3.** Each reaction mixture (20 ul) is composed of PowerUp SYBR Green Master Mix (Applied Biosystems), 0.5 uM each primer, and cDNA (150 ng for *F2* gene, 20-50 ng for the other genes), Each sample was tested in duplicate under the following amplification conditions: 95°C for 2 min, and then 40 cycles of 95°C for 1 s and 60°C for 30 s in QuantStudio 3 Real-Time PCR System (Applied Biosystems, Foster City, CA, USA). PCR primers were designed

using the Primer-BLAST tool available from the NCBI web site [204]. Melt curve was checked to confirm the specificity of the amplified product. Relative quantification was calculated using the 2 ^ (-ΔΔ CT) method [205]. Beta actin was used as an endogenous control gene to evaluate the gene expression levels. All data are presented as the mean ± s.e.m of n = 4/group. Statistical significance was determined by Two-tailed Student's t test and values were considered statistically significant at $P < 0.05$.

Extraction and quantification of lipids in cells and media

Lipids were extracted from C3H10T1/2 cells and culture media using the Folch method [206] with minor modifications. Briefly, whole culture medium (1 mL) from each well of 12-well plate was collected in a separate tube. Cells were washed with phosphate buffered saline (PBS), and collected in 1 mL PBS and homogenized. Media or cell homogenate was mixed in 5 ml of chloroform: methanol (2:1, vol/vol) by shaking vigorously several times, and centrifuged at 2,500 x g for 15 min. Bottom organic layer was transferred to a new glass tube. The remaining aqueous phase and interphase including soluble protein were mixed with 5 mL chloroform by vigorous shaking, followed by centrifugation at 2,500 x g for 15 min. Bottom organic layer was combined with the first collected organic layer. The combined organic phase was evaporated using nitrogen, and then the dried lipids were resuspended in 0.5 % Triton X-100 in water. Samples were stored in - 80°C until lipid analysis. Triglyceride (TG), total cholesterol (TC), unesterified cholesterol (UC), and phospholipid (PL) levels in lipid extractions from cells and from culture media were measured separately using a colorimetric assay at the UCLA GTM Mouse Transfer Core [207]. Intracellular lipids were normalized to the cellular protein amount measured by BCA protein assay kit (Pierce, Rockford, IL, USA). Extracellular lipids are presented as lipid quantity in 1 mL of collected media.

3.6 Tables

Table 3.1. Common pathways shared by the four lipid traits in SNP set enrichment analysis.

Categories	Descriptions	Traits*				MetaboChip	iGSEA
		HDL	LDL	TC	TG		
Positive Controls	Positive control gene set for TG	1,2,3,5,6,7,8,9	2,3,5,6,7,8,9	2,3,5,6,7,8,9	1,2,3,5,6,7,8,9	YES	YES
	Positive control gene set for LDL	5,6,7,8,9	1,2,3,4,5,6,7,8,9	1,2,3,4,5,6,7,8,9	1,2,3,5,6,7,8,9	YES	YES
	Positive control gene set for TC	3,5,6,7,8,9	1,2,3,4,5,6,7,8,9	1,2,3,4,5,6,7,8,9	1,2,3,5,6,7,8,9	YES	YES
	Positive control gene set for HDL	1,2,3,4,5,6,7,8,9	2,6,7,8,9	2,5,6,7,8,9	1,2,5,6,7,8,9	YES	YES
Lipid metabolism	Lipoprotein metabolism	1,2,5,6,7,8,9	5,6,7,8,9	5,6,7,8,9	5,6,7,8,9	YES	YES
	Chylomicron-mediated lipid transport	5,6,7,8,9	7,8,9	5,6,7,8,9	5,6,7,8,9	YES	YES
	LDL-mediated lipid transport	6,7,9	6,7,9	6,7,9	6,7,9	NO	YES
	HDL-mediated lipid transport	1,2,5,6,7,8,9	5,7,8,9	5,7,8,9	5,7,8,9	YES	YES
Protein catabolism	ER-Phagosome pathway	1,5,8,9	1,3,5,6,8,9	1,2,3,5,6,8,9	1,3,5,6,8,9	YES	YES
	Antigen processing and presentation	5,9	1,2,3,5,6,7,8,9	1,2,3,5,6,7,8,9	1,2,3,5,6,7,8,9	YES	YES
Interferon Signaling	Interferon Signaling	7,9	1,3,5,6,8,9	1,2,3,5,6,8,9	1,3,5,8	YES	YES
Autoimmune /Immune activation	Type I diabetes mellitus	1,5	1,2,3,5,6,7,8,9	1,2,3,5,6,7,8,9	1,2,3,5,6,7,8,9	YES	YES
	Scavenging by Class B Receptors	6,7,8,9	7,9	7,9	7,9	NO	YES
	Asthma	6	1,3,5,6,7,8,9	1,2,3,5,6,7,8,9	1,2,3,5,6,7,8,9	YES	YES
	IL 5 Signaling Pathway	5	1,5,6,8,9	1,5,6,8,9	5,6,8	NO	NO
	Th1/Th2 Differentiation	3	1,3,5,6,8	1,3,5,6,8,9	1,3,5,6,8	NO	YES
	Natural killer cell mediated cytotoxicity	5	1,3,5	1,3,5,6,9	1,3,5	YES	YES
	HLA genes	1,3,5,6,7,8,9	1,2,3,5,6,7,8,9	1,2,3,5,6,7,8,9	1,2,3,5,6,7,8,9	YES	YES
	Cell adhesion molecules (CAMs)	5	1,2,3,5,6,7,8,9	1,2,3,5,6,7,8,9	1,3,5,6,8,9	YES	NO
	Autoimmune thyroid disease	1,3,5,6,8,9	1,2,3,5,6,7,8,9	1,2,3,5,6,7,8,9	1,2,3,5,6,7,8,9	YES	YES
Visual transduction	Diseases associated with visual transduction	7	7,8,9	7,8,9	7,9	YES	YES
	Visual phototransduction	7	7,8,9	7,8,9	7,9	YES	YES

6..1 Note: *: The trait columns represent in which methods the MSEA of the pathways is significant with FDR < 10%. Number 1 to 9 represent: adipose eSNP (1), blood eSNP (2), brain eSNP (3), human aortic endothelial cells (HAEC) eSNP (4), liver eSNP (5), all eSNP (6), Distance (7), Regulome (8), and Combined (9), respectively. The Metabochip and iGSEA columns tell whether the gene set can also be detected as statistically significant in the analysis.

Table 3.2. Trait-specific pathways identified in the SNP set enrichment analysis for four lipid traits.

Traits	Modules	Descriptions	Methods*
HDL	rctm0846	Packaging of telomere ends	1
	Haec:M1+	(Cholesterol biosynthesis)	9
	M12882	Taurine and hypotaurine metabolism	1,5
	rctm0060	Activation of Genes by ATF4	9
	rctm0216	Cation-coupled Chloride cotransporters	7,8,9
	rctm0697	Metabolism of water-soluble vitamins and cofactors	5
	Cerebellum:M1+	(Alcohol metabolism)	3
	Cerebellum:M2+		3
	rctm0507	Glutathione synthesis and recycling	5
	Liver:M1+	(Transition metal ion homeostasis)	2,9
	rctm0937	RIG-I/MDA5 mediated induction of IFN-alpha/beta pathways	7,8,9
	rctm0772	Negative regulators of RIG-I/MDA5 signaling	7,8,9
	rctm0255	Cobalamin (Cbl, vitamin B12) transport and metabolism	1,5
	M15902	Glycerolipid metabolism	6,7,9
	rctm1178	Striated muscle contraction	9
	rctm0696	Metabolism of vitamins and cofactors	5
LDL	Haec:M2+	(Positive regulation of cellular metabolism)	3
	Liver:M2+	(Cadherin)	6
	Cerebellum:M3+	(Immunity and defense)	8
	M6831	The citric acid cycle	6
	rctm0876	Platelet sensitization by LDL	7,9
TC	M17946	Valine, leucine and isoleucine biosynthesis	1,6,9
	PC:M1+	(Chaperone)	3
	Cerebellum:M4+	(Response to wounding)	9
	Adipose:M1+		8
	Omental:M1+		3
	rctm1111	Signal transduction by L1	3
TG	rctm1276	Tight junction interactions	1,6,8,9
	rctm0589	Initial triggering of complement	1
	rctm0235	Cholesterol biosynthesis	2
	M18155	Insulin signaling pathway	1
	Blood:M1+	(Carbohydrate metabolism)	1,6
	rctm0225	Cell-cell junction organization	1,6,8
	Blood:M3+	(Transferase activity, transferring glycosyl groups)	1
	M7146	Classical complement pathway	1
	rctm0059	Activation of Gene Expression by SREBP (SREBF)	2
	M917	Complement pathway	1
	M5872	Steroid biosynthesis	2
	Omental:M2+	(hemopoietic or lymphoid organ development)	8
	M2164	Leukocyte transendothelial migration	1

Note: *: The method column represents in which methods the MSEA of the pathways is significant with FDR < 10%. Number 1 to 9 represent: adipose eSNP (1), blood eSNP (2), brain eSNP (3), human aortic endothelial cells (HAEC) eSNP (4), liver eSNP (5), all eSNP (6), Distance (7), Regulome (8), and Combined (9), respectively. $^{+}$: Co-expression modules. The statistically overrepresented Gene Ontologies satisfying $p < 0.01$ in Fisher's exact test after Benjamini-Hochberg correction within the modules are listed in the parentheses. PC: prefrontal cortex. #: The column tells whether the trait-specific pathways can also be detected as trait-specific ones in either Metabochip and/or iGSEA.

Table 3.3. Supersets shared by four lipid traits and key driver genes.

Supersets	No. Genes	Methods[a]				Top Adipose KDs	Top Liver KDs
		HDL	LDL	TC	TG		
Lipid metabolism	793	1,2,3,5	1,2,3,5	1,2,3,5	1,2,3,5	APOH, ABCB11, F2, ALB, APOA5, APOC4, DMGDH, SERPINC1, APOF, HADHB, ETFDH, KLKB1	HMGS1, FDFT1, FADS1, DHCR7, ACAT2, ACSS2
Protein catabolism	253	1,3,4,5,6, 7,8,9	1,3,5,6	1,3,5,6,9	1,3,5,6,8	PSMB9	PSMB9
Interferon Signaling	171	1,3,5,7,8, 9	1,2,3,5,6, 7,8,9	1,2,3,5,6, 7,8,9	1,2,3,5,6, 8,9	NUP210	MX1, ISG15, MX2, IFI44, EPSTI1
Autoimmune/ Immune activation	152	1,3,4,5,6, 7,8,9	1,2,3,4,5, 6,7,8,9	1,2,3,4,5, 6,7,8,9	1,2,3,4,5, 6,7,8,9	HLA-DMB, HCK, SYK, CD86	HLA-DMB, CCL5, HLA-DQA1
Visual transduction	86	7,9	7,8,9	7,8,9	7,8,9	-	-

Note: [a] The method column represents in which methods the MSEA of the pathways is significant with Bonferroni-adjusted P<0.05. Number 1 to 9 represent: adipose eSNP, blood eSNP, brain eSNP, haec eSNP, liver eSNP, all eSNP, Distance, Regulome, and Combined, respectively.

3.6 Figures

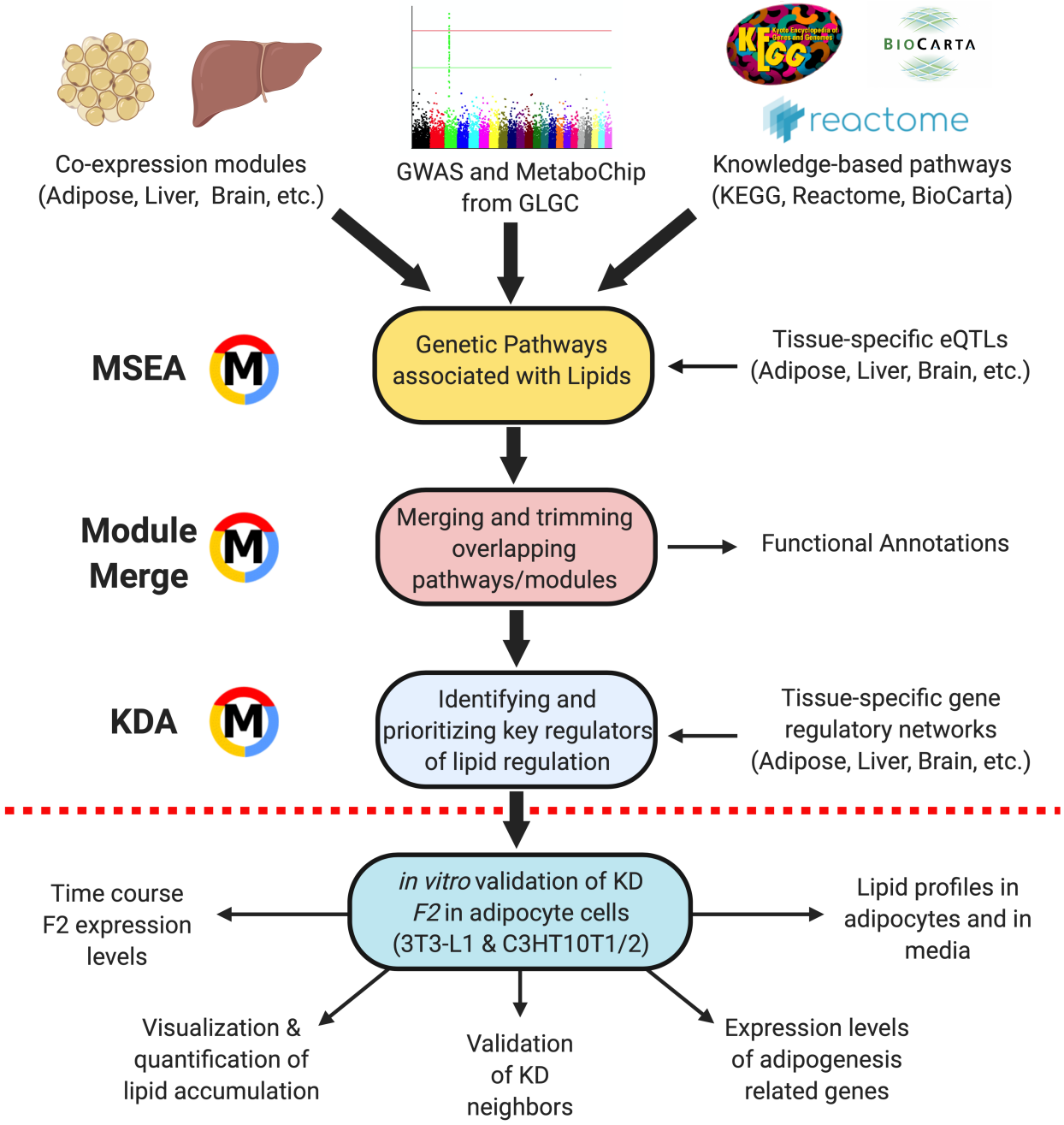

Figure 3.1. Overall design of the study. The statistical framework can be divided into four main parts, including Marker Set Enrichment Analysis (MSEA), merging and trimming of gene sets, Key Driver Analysis (KDA), and validation of the key regulators using *in vitro* testing.

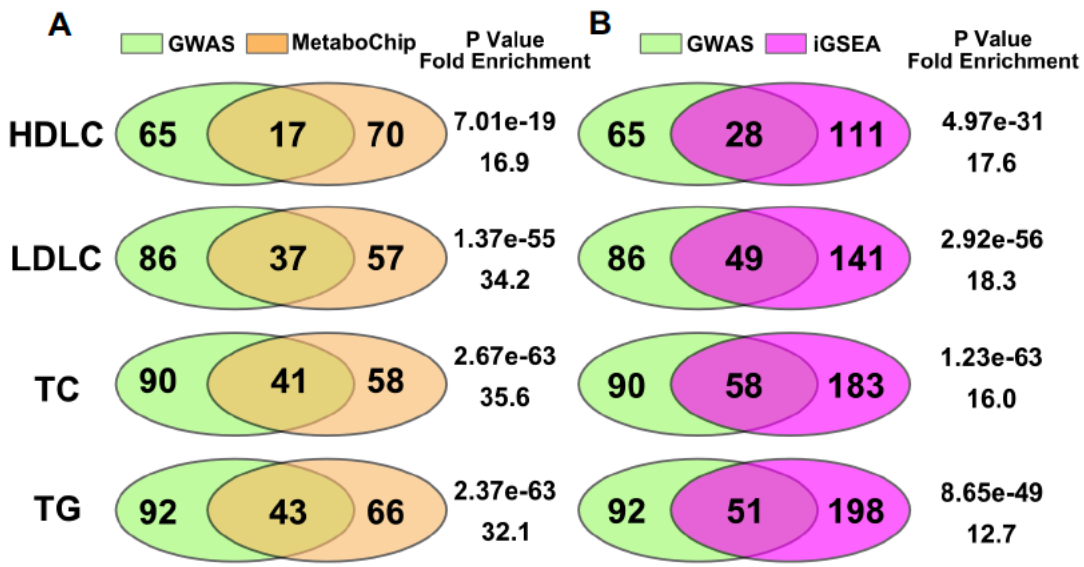

Figure 3.2. Validation of MSEA results from GLGC GWAS using independent genetic association data from MetaboChip and a different method iGSEA. A) Venn diagram of the convergent pathways between GLGC GWAS and MetaboChip dataset using the same MSEA method. B) Venn diagram of the convergent pathways between MSEA and iGSEA for the same GLGC GWAS dataset. Fisher exact test was applied to evaluate the overlap in the pathways detected using different datasets or using different methods, with 4532 pathways in total.

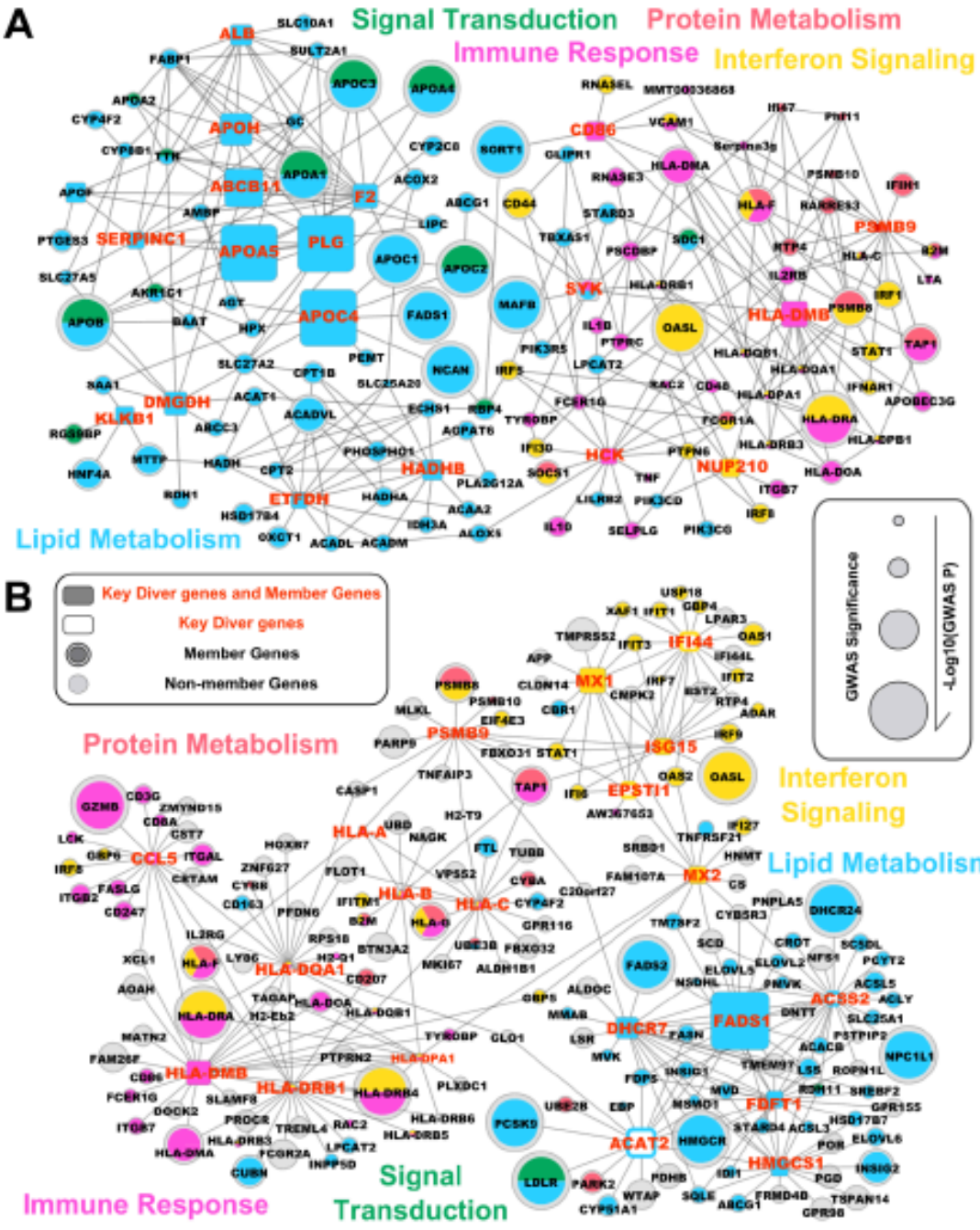

Figure 3.3. Common KDs and their neighboring genes in the shared lipid- associated subnetworks. A) Adipose KDs and subnetworks. B) Liver KDs and subnetworks. The subnetworks shared by HDL, LDL, TC, and TG are depicted by different colors according to the difference in their functional categories. Nodes are the KDs and their adjacent regulatory partner genes, with KDs depicted as larger nodes. Only network edges that were present in at least two independent network studies were included. The node size corresponds to the GWAS significance.

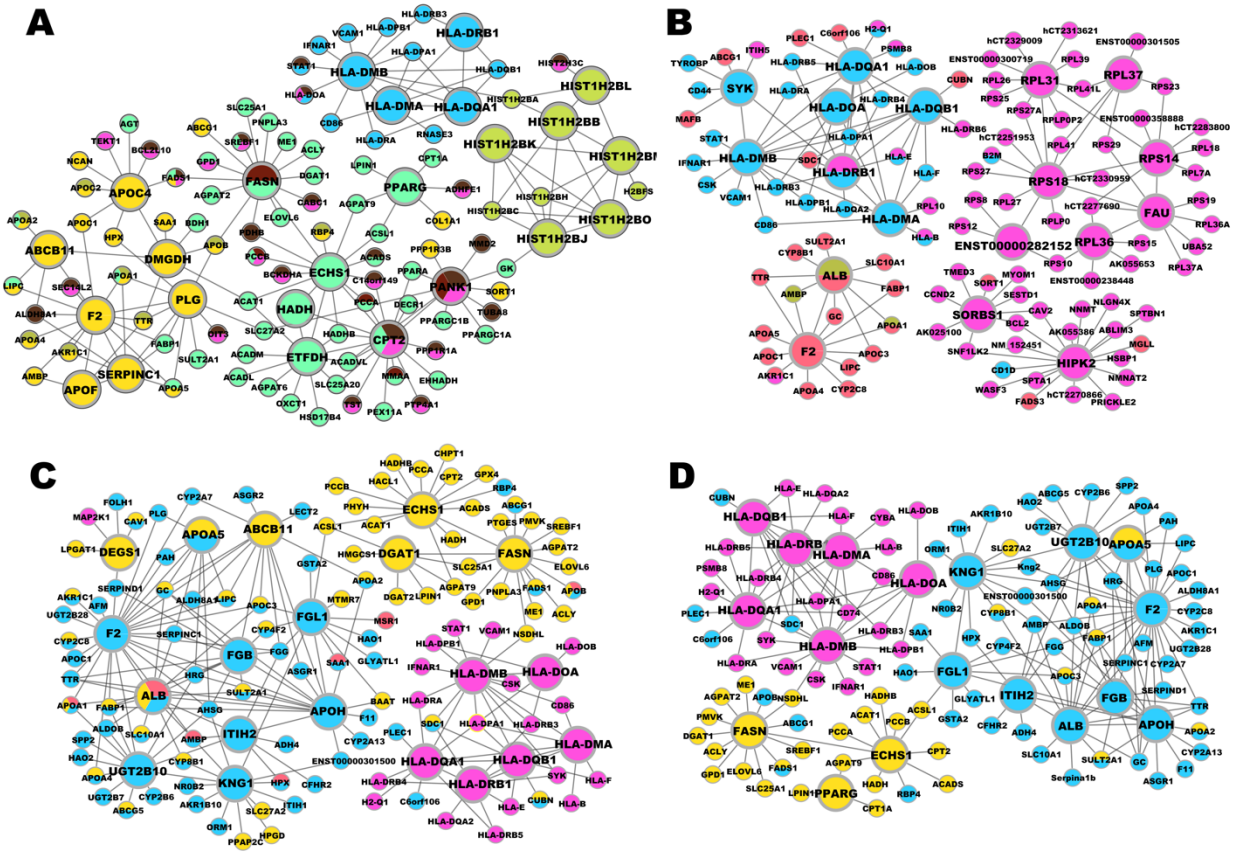

Figure 3.4. Adipose KDs and subnetworks for each lipid trait. Panel (A)-(D) represent HDL, LDL, TC, and TG subnetworks. Nodes are the KDs and their adjacent regulatory partner genes, with KDs depicted as larger nodes. The yellow color signifies networks associated with interferon signaling, blue with lipid metabolism, pink with immune response, green with protein metabolism, red with lipoprotein metabolism and brown with fatty acid oxidation.

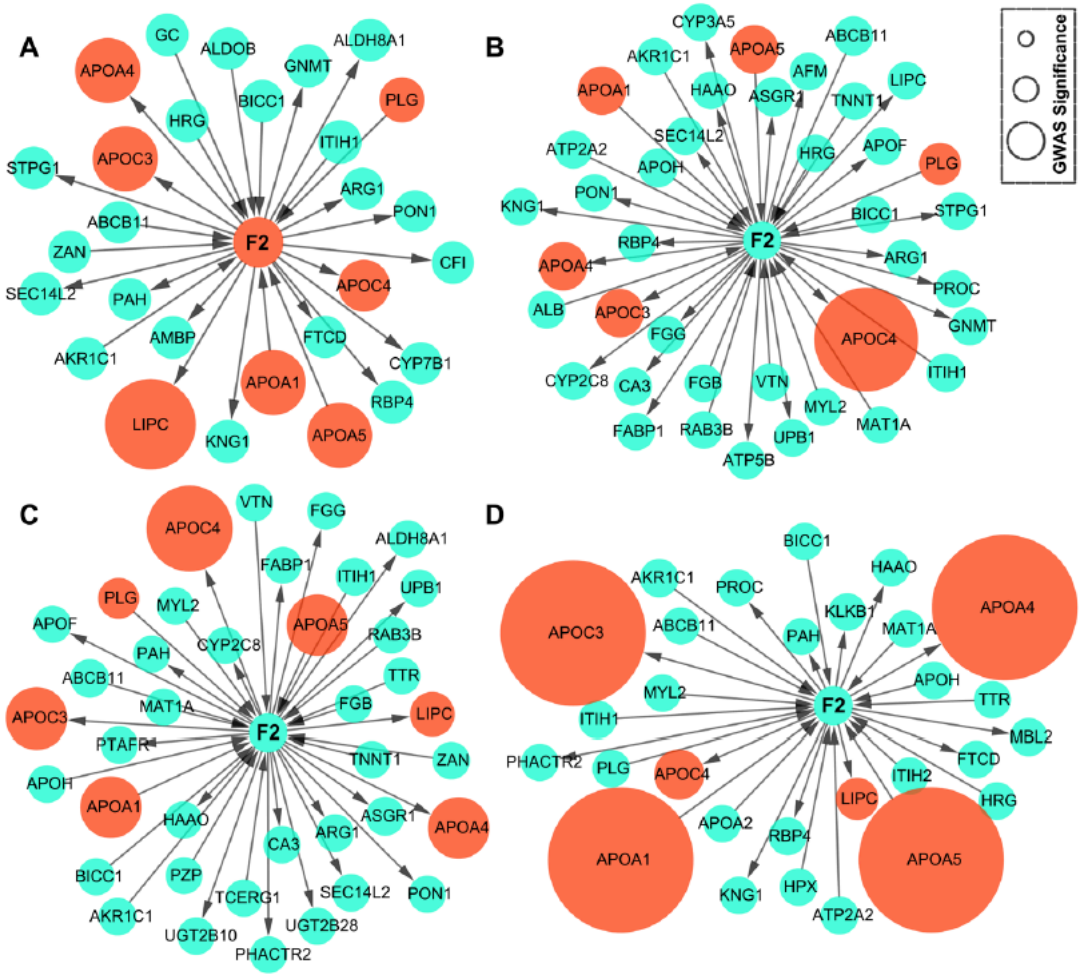

Figure 3.5. GWAS genes in Neighboring genes of Gene F2 in human Bayesian networks. Panel (A-D) represent GWAS susceptibility genes around gene F2 for HDL, LDL, TC, and TG respectively. The interactions come from a combined Bayesian network from different human tissues, including adipose, liver, blood, kidney, muscle, and brain. The node size corresponds to the GWAS significance.

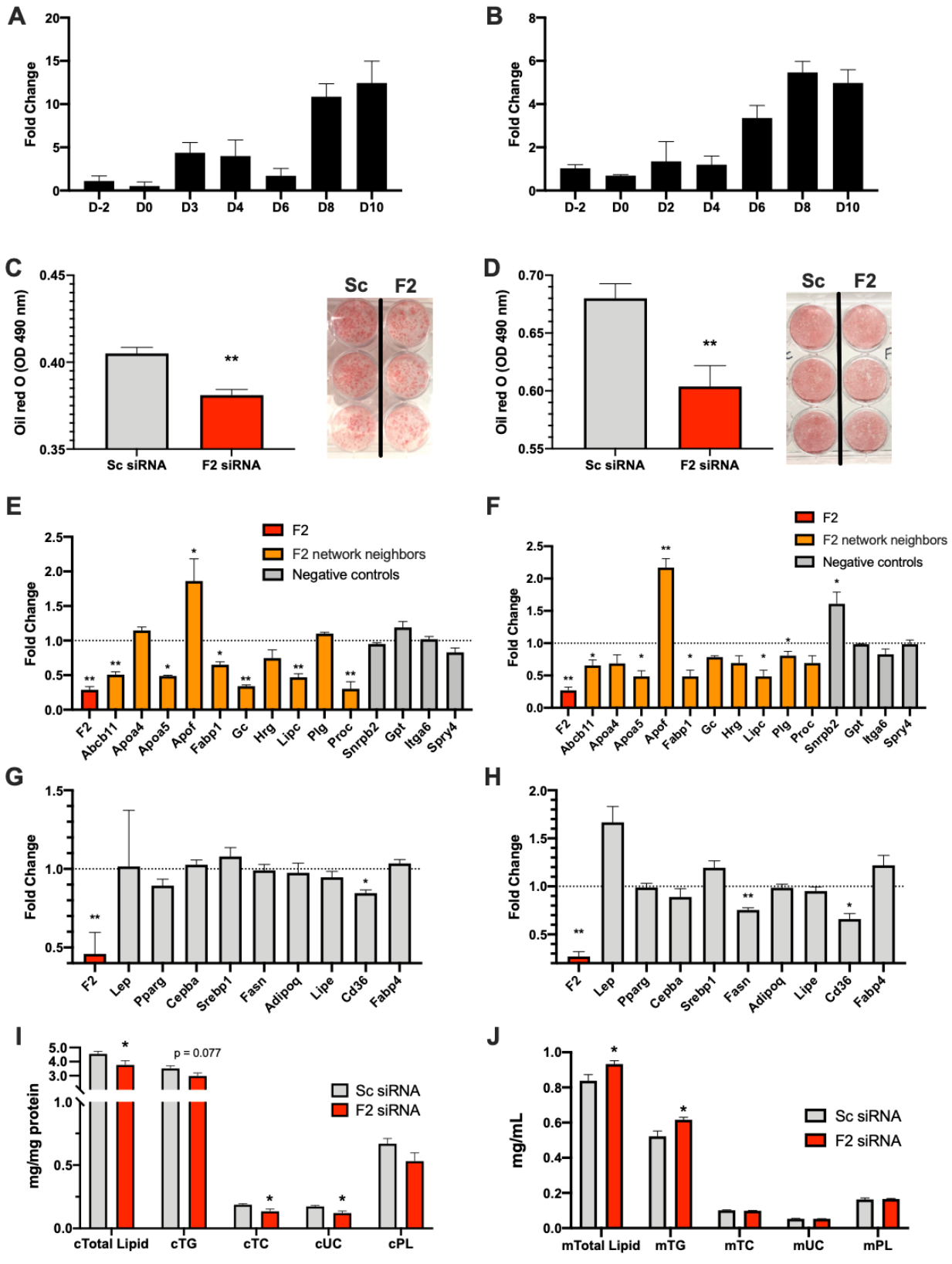
A
Fold Change
D-2
D0
D3
D4
D6
D8
D10
B
Fold Change
D-2
D0
D2
D4
D6
D8
D10
C
Oil red O (OD 490 nm)
Sc siRNA
F2 siRNA
Sc
F2
D
Oil red O (OD 490 nm)
Sc siRNA
F2 siRNA
Sc
F2
E
F2
F2 network neighbors
Negative controls
Fold Change
F2
Abcb11
Apoa4
Apoa5
Apof
Fabp1
Gc
Hrg
Lipc
Plg
Proc
Snrpb2
Gpt
Itga6
Spry4
F
F2
F2 network neighbors
Negative controls
Fold Change
G
Fold Change
F2
Lep
Pparg
Cepba
Srebp1
Fasn
Adipoq
Lipe
Cd36
Fabp4
H
Fold Change
I
mg/mg protein
p = 0.077
Sc siRNA
F2 siRNA
cTotal Lipid
cTG
cTC
cUC
cPL
J
mg/mL
Sc siRNA
F2 siRNA
mTotal Lipid
mTG
mTC
mUC
mPL

Figure 3.6. Validation of *F2*'s predicted subnetwork and regulatory role in adipocytes. A, B) Time course of *F2* expression during adipocyte differentiation in 3T3-L1 cells (A) and C3H10T1/2 cells (B). D-2, D0, D2, D3, D4, D6, D8, D10 indicate two days before initiation of differentiation, day 0, day 2, day 3, day 4, day 6, day 8, and day 10 of differentiation, respectively. Sample size n=2-3/time point. C, D) Visualization and quantification (OD value) of lipid accumulation by Oil red O staining in 3T3-L1 adipocytes (C) and C3H10T1/2 adipocytes (D). Sample size n = 5-8/group for adipocytes. E, F) Fold change of expression level for *F2* adipose subnetwork genes and negative control genes after siRNA knockdown. At day 7 of differentiation of 3T3-L1 and day 5 and day 7 of differentiation of C3H10T1/2, adipocytes were transfected with *F2* siRNA for the knockdown experiments. Ten *F2* neighbors were randomly selected from the first and second level neighboring genes of *F2* in adipose network. Four negative controls were randomly selected from the genes not directly connected to *F2* in adipose network. G, H) The fold changes of adipokine/adipogenesis-related genes in 3T3-L1 (G) and in C3H10T1/2 (H). Gene expression levels were determined by RT-qPCR, normalized to Beta actin. The fold changes were relative to scrambled siRNA control. Sample size n=4/group. I, J) Lipid profiles: Total Lipid, Triglyceride (TG), Total Cholesterol (TC), Unesterified Cholesterol (UC) and Phospholipid (PL) in C3H10T1/2 cells (I) and in media (J). Total Lipid was estimated using the sum of the four lipids (TG, TC, UC, PL). Intracellular lipids plotted in (I) were normalized to total cellular protein quantity. Extracellular lipids plotted in (J) are presented as lipid quantity in 1 mL of collected media. Sample size n = 6/group. Results represent mean ± s.e.m. Statistical significance was determined by two-sided Student's t-test (*$p < 0.05$ and **$p < 0.01$).

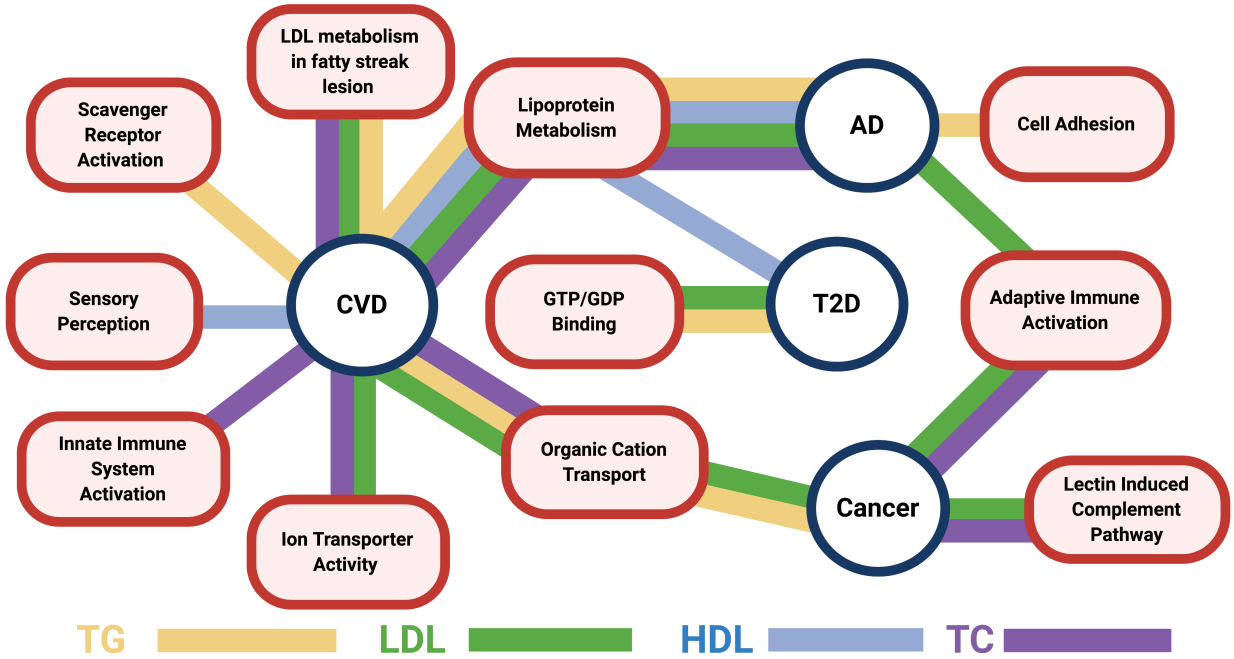

Figure 3.7. The associations between lipid-associated supersets and human complex diseases. The edges represent the associations between supersets for the specific lipid classes matched by color and diseases (p value < 0.05; Fisher exact test with Bonferroni correction). AD: Alzheimer's disease; CVD: cardiovascular diseases; T2D: type 2 diabetes.

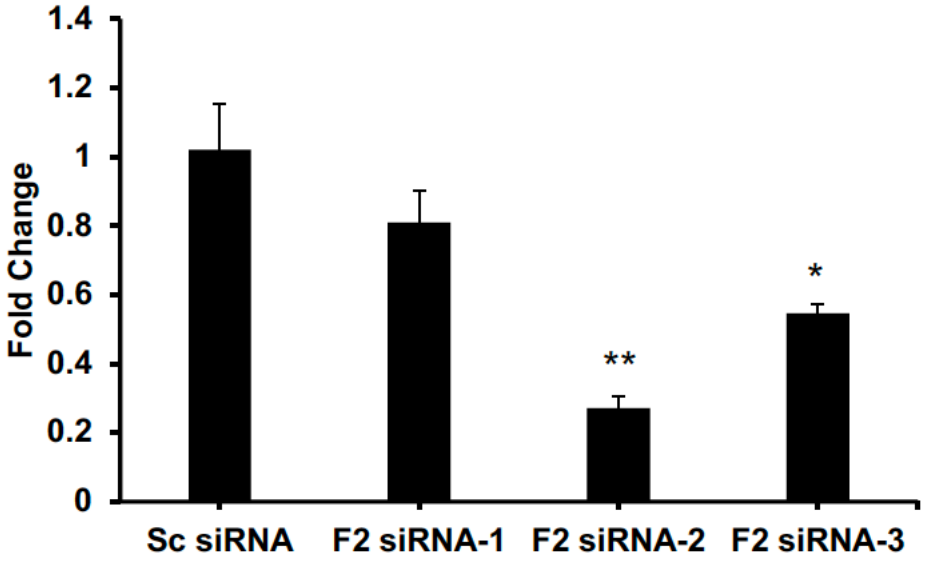

Figure 3.8. Gene knockdown efficiencies of three F2 siRNAs. Three F2 siRNAs were tested to select for the most efficient knockdown of the target gene. 3T3-L1 adipocytes were transfected with F2 siRNA at day 7 of differentiation (D7). Scrambled (Sc) siRNA was used as the control for normalization. After 48 h, F2 gene expression was analyzed by real-time qPCR. Beta actin was used as a housekeeping gene. Result represents the mean ± s.e.m. n = 3/group. Statistical significance was determined by Student's t-test between each F2 siRNA and the Sc siRNA (*$p < 0.05$ and **$p < 0.01$).

Chapter 4. IAPP-induced beta cell stress recapitulates the islet transcriptome in type 2 diabetes

4.1 Introduction

The islet in type 2 diabetes is characterised by islet amyloid derived from islet amyloid polypeptide (IAPP), a protein co-expressed with insulin by beta cells that when misfolded and in aggregate form may contribute to beta cell failure [208-211]. Human IAPP (hIAPP) toxicity is most potently mediated by small intracellular membrane permeant oligomers [212]. Species with amyloidogenic IAPP, such as humans, non-human primates and cats, share vulnerability to type 2 diabetes, while those with non-amyloidogenic IAPP, such as mice and rats, do not [213]. While numerous hypotheses have been put forward to explain the wide-ranging changes in islets in type 2 diabetes [214], there is a consensus that misfolded protein stress induced by toxic oligomers of amyloidogenic proteins initiate these changes in neurodegenerative diseases.

Given the known proximal role of misfolded protein stress in neurodegenerative diseases, and the connection of the risk factors for type 2 diabetes to misfolded protein stress, we hypothesised that hIAPP misfolded protein stress may be a proximal cause of the wide-ranging changes in islets in individuals with type 2 diabetes. Risk factors for type 2 diabetes include insulin resistance [215] and low birthweight [216]. Low birthweight may lead to low adult beta cell mass [217], which together with insulin resistance, predicts a high insulin and IAPP expression rate per beta cell [218]. Beta cell misfolded protein stress is induced when expression of hIAPP per cell exceeds the cellular threshold for clearing misfolded proteins [219]. This threshold declines with ageing [220], a risk factor for both type 2 diabetes and neurodegenerative diseases. hIAPP overexpression in isolated mouse islets in vitro can modify islet gene expression with relevance to type 2 diabetes [221].

In the present study we evaluated the islet transcriptome from a mouse model of beta cell hIAPP toxicity [219] before diabetes onset in order to avoid the confounding effects of hyperglycaemia. To control for the increased burden of IAPP expression, we evaluated the transcriptome from mice overexpressing rodent IAPP (rIAPP) [222]. We then compared the changes in the transcriptome of hIAPP or rIAPP islets to those in humans with prediabetes or type 2 diabetes to establish if the changes in the islet in type 2 diabetes are potentially attributable in part to hIAPP protein misfolding stress.

4.2 Results

Similarity in islet transcriptome in prediabetes or type 2 diabetes and IAPP overexpressing mice

There is striking concordance of the islet transcriptome between individuals with prediabetes and those with type 2 diabetes (**Figure 5.1a**). Since ~80% of individuals with prediabetes do not develop diabetes [223], this finding implies that a high proportion of the changes in islets in type 2 diabetes are adaptive rather than causal of diabetes. There was also a close concordance of changes in islet gene expression in type 2 diabetes and mouse islets with increased hIAPP or rIAPP expression (**Figure 5.1b–d; Figure 5.2**).

Since by design neither the hIAPP or rIAPP mice had diabetes when islets were sampled, we further compared the changes in islet transcriptome in the mouse islets with those in human islets with prediabetes. There was again close concordance in transcriptome in both rIAPP and hIAPP islets with those from individuals with prediabetes (Figure 5.1e, f). These findings imply that a more detailed analysis of the transcriptome in response to increased expression of rIAPP and hIAPP might shed light on adaptive vs disease causal changes in type 2 diabetes.

Islet transcriptome in response to increased beta cell rIAPP or hIAPP expression

To investigate beta cell adaptation to an increased workload of non-amyloidogenic rIAPP expression, we compared the transcriptome of islets from rIAPP vs WT mice. Gene expression was increased in WGCNA modules M8 (Mitogen activated protein kinase (MAPK) signalling), M9 (TGF-β signalling), M13 (cell migration) and decreased in M7 (cell cycle) in rIAPP islets (**Figure 5.3, Table 5.1, Figure. 5.4**). Differential expression analysis identified 2731 DEGs (1306 up- and 1425 downregulated in rIAPP islets; FDR<0.05) (**Figure 5.5**). Of the 2731 DEGs, 14 are implicated in monogenetic diabetes and 39 are located in genomic regions associated with type 2 diabetes by genome-wide association studies (GWAS), consistent with the overlap between rIAPP and type 2 diabetes by RRHO analysis (**Figure 5.1c, Figure 5.6**). Prominently upregulated genes in rIAPP islets include those required for protein synthesis and those that adapt cells to an increased burden of protein folding and quality control, collectively referred to as the unfolded protein response (UPR) (**Figure 5.5c, Figure 5.7, 5.8**).

Since there is a considerable overlap in the changes in transcriptome in hIAPP and rIAPP mice (Fig. 1d), and hIAPP but not rIAPP mice develop diabetes, we next compared these transcriptomes to discern changes related to oligomer toxicity vs adaptation to an increased burden of IAPP expression. Five co-expression modules were differentially expressed between hIAPP and rIAPP islets: M1 (inflammation), M4 (oxidative stress) and M6 (cell cycle; cell signalling) expression was increased while M5 (RNA processing) and M8 (MAPK signalling; cell adhesion) were downregulated in hIAPP (**Figure 5.3b, c, Table 5.1**). This pattern of changes is consistent with islet inflammation reported in type 2 diabetes [224]. We identified 2011 DEGs between hIAPP and rIAPP islets, with 1031 upregulated and 980 downregulated in hIAPP islets (**Figure 5.5b**). Among these, 13 DEGs have been implicated in monogenetic diabetes and 194 are located in genomic regions associated with type 2 diabetes by GWAS (**Figure 5.6**). A number of interesting trends were found in genes

involved in critical pathways associated with type 2 diabetes development (**Figure 5.5, Figure 5.6**).

UPR Consistent with the comparable increase in IAPP expression [222], islets in hIAPP mice share a comparable increase in the UPR to rIAPP mice (**Figure 5.5c**). The evaluation of changes in expression of the same genes in humans with type 2 diabetes or prediabetes compared with non-diabetics show consistently upregulated *BHLHA15* (*MIST1*), a potent endoplasmic reticulum (ER) stress-inducible transcriptional regulator upregulated by beta cell Ca^{2+} overload [225].

Inflammation The pronounced signal for inflammation in hIAPP, but not rIAPP islets is also present in human prediabetes and type 2 diabetes (**Figure 5.5**). Consistent with a proinflammatory state, several key cell surface antigens (*Cd4*, *Cd68*, *Cd84*), immune sensors (*Cx3cr1*, *Clec7a*), and macrophage genes (*Axl*, *Slc11a1)* were upregulated in hIAPP islets and in type 2 diabetes. The top hub gene in the M1 module that is most highly activated in hIAPP islets is *Cx3cr1* [226]. This prominence of islet inflammation in hIAPP as compared with rIAPP is further shown by the lack of difference between rIAPP and WT in module M1 (immune response) (**Figure 5.3b, c**).

Cell cycle Activation of cell replication is a key component of injury repair programmes. Cell cycle related genes are increased in hIAPP islets but decreased in rIAPP islets (**Figure 5.5**), likely reflecting the injury-mediated signalling in hIAPP islets vs adaptive UPR in rIAPP islets [227]. Prominent amongst upregulated genes in hIAPP islets are those that enhance cell replication directly (*Cdk1, Cep55)* or indirectly (*Hmmr, Anln*). In contrast, these genes are downregulated in islets from prediabetes and type 2 diabetes, which is perhaps consistent with epigenetic silencing of cell cycle genes in beta cells in adult humans compared with 9-week-old mice [228].

Beta cell dedifferentiation Genes important for maintaining beta cell differentiation were decreased in islets of both rIAPP and hIAPP mice compared with islets from WT mice (**Figure 5.5**), a pattern reproduced in islets from humans with prediabetes and type 2 diabetes. In mice, beta cell dedifferentiation was confirmed by mRNA and protein analysis (**Figure 5.8a–c**). To establish if this partial dedifferentiation impacts glucose tolerance, we performed IPGTTs. As expected, mice transgenic for hIAPP were glucose intolerant compared with WT mice. Although to a lesser extent, mice transgenic for rIAPP were also glucose intolerant, consistent with the observed partial beta cell dedifferentiation and known action of IAPP to inhibit insulin secretion [229] (**Figure 5.9d**).

Contribution of calpain hyperactivation to beta cell hIAPP toxicity

Calpain hyperactivation has been widely reported as a mediator of amyloidogenic protein induced cytotoxicity, presumably as a consequence of aberrant Ca^{2+} signalling that results from nonselective ion channel activity of toxic oligomers [230]. Concurrent beta cell-specific overexpression of human calpastatin (*CAST*) (hIAPP:hCAST), which inhibits calpain, delays or prevents diabetes in hIAPP transgenic mice [231].

Functional enrichment analysis revealed the sets of genes partially rescued by *CAST* overexpression in hIAPP mice were those that mediate UPR and inflammation (**Figure 5.9**). Sustained calpain hyperactivation may activate proinflammatory signalling pathways mediated by NF-κB and eNOS (endothelial nitric oxide synthase) [232].

To further evaluate the role of calpain activation in type 2 diabetes and prediabetes, we correlated the differential expression patterns and found that the type 2 diabetes islet profile is best reflected by the hIAPP islet profile, outperforming the rIAPP and hIAPP:hCAST profiles (**Figure 5.9e**).

Similarly, RRHO analysis showed a marked decrease in shared transcriptome between hIAPP and type 2 diabetes islets after introduction of human calpastatin (**Figure 5.9f vs Figure 5.1b, e**). These results imply that calpain hyperactivation may play a prominent role in the shared transcriptome between hIAPP and type 2 diabetes.

Cell type analysis for IAPP misfolded protein stress

To explore the cell types that contribute to the transcriptomic signals in the bulk RNA-seq analysis, we conducted cell type marker enrichment analysis of the DEGs and modules as well as cell proportion deconvolution of bulk islet RNA-seq (**Figure 5.10**). These analyses emphasised the impact of beta cell IAPP misfolded protein stress on multiple islet cell types including beta, alpha, endothelial, macrophage and stellate. Both the DEG and module enrichment analysis for cell type markers (**Figure 5.10a, b**) point to a role of stellate cells with marker enrichment for module M13 (cell migration) and downregulated DEGs in hIAPP:hCAST vs hIAPP mice, possibly implicating the role of calpastatin in reducing stellate cell population. The deconvolution results showcase a reduction in beta cell and a rise in alpha cell populations in both rIAPP and hIAPP (**Figure 5.10c**), which alludes to partial beta cell dedifferentiation (**Figure 5.10d**). The increase in endothelial cell populations exemplifies an increase in vascularisation within the islet consistent with response to injury and dedifferentiation. Similarly, the stellate cell populations also follow this trend, and rescue by hIAPP:hCAST shows a reduction in both endothelial and stellate cell (**Figure 5.10d**), which is consistent with reduced inflammation (**Figure 5.9c**). In addition, there is a subtle increase in macrophages in hIAPP (**Figure 5.10d**), matched by macrophage marker enrichment in module M1 (immune response) and in DEGs between hIAPP and rIAPP as well as between hIAPP and hIAPP:hCAST (Fig. 2b, c). These results may explain the differences in inflammation between

groups. Complementing this, we found that module M7, associated with beta cells (**Figure 5.10b**), was also enriched for type 2 diabetes GWAS candidate genes (**Figure 5.6b**).

Prominent regulatory factors shared between hIAPP islets and type 2 diabetes islets

Having established that hIAPP-induced beta cell toxicity in mouse islets results in an islet transcriptome mimicking that of human islets in type 2 diabetes, we evaluated regulatory cascades in hIAPP and type 2 diabetes-associated transcriptomic alterations by TF network and non-TF gene network analysis. TF network analysis uncovered four upstream hub transcriptional factors *NF-κB, ESR1, STAT3 and CTNNB1* active in both type 2 diabetes and hIAPP islets (**Figure 5.11**). NF-κB activation has been implicated as a core transcriptional mediator of neuronal cell inflammatory responses to amyloidogenic misfolded stress in neurons and can be protective or contribute to toxicity of injured beta cells [233] [234] and is activated by aberrant Ca^{2+} signalling and calpain hyperactivation [235]. NF-κB1 was assigned as an upstream regulator to co-expression module M7 that is enriched for beta cell markers (**Figure 5.10b**) and type 2 diabetes GWAS candidate genes (**Figure 5.10b**). NF-κB is a known target of calpain, consistent with calpain hyperactivation in beta cells in hIAPP mouse islets and type 2 diabetes. STAT3 is activated by aberrant Ca^{2+} signalling and has been reported as a key regulator of inflammation in neurodegenerative diseases [236]. STAT3 and NF-κB cooperate as transcriptional regulators to induce angiogenesis, cellular proliferation and pro-survival metabolic remodelling, the latter through activation of HIF1α (hypoxia-inducible factor 1 alpha) [237]. CTNNB1 encodes β-catenin that has been implicated in tissue repair and regeneration responses in gut and beta cells, inducing cell proliferation, cell migration repair of cytoskeleton and regulation of intracellular Ca^{2+} dynamics [227]. ESR1 signalling is protective of beta cells in response to injury in both human and mouse islets [238].

Complementary to the TF network analysis, which is not tissue specific, we utilised an islet gene regulatory network constructed using population-based genetic and transcriptomic datasets, which captures network regulators that are not necessarily TFs. Here, we uncover the interconnectivity between modules/processes for their role in islet pathogenesis and their associated regulatory genes (**Figure 5.12; Figure 5.13**). With stellate and endothelial cell markers enriched for M13 (**Figure 5.10b**), hub genes such as *COL3A1* (extracellular matrix), *NID1* (wound healing) and *CXCL12* (leucocyte trafficking, angiogenesis and vascular repair) are plausible regulatory genes underlying stellate and endothelial cell contribution to islet pathogenesis. Moreover, module M7, which is enriched for beta cell markers (**Figure 5.10b**) and type 2 diabetes GWAS candidates (**Figure 5.12; Figure 5.10**), highlights hub genes such as *ZNF800*, which is closely associated with *PAX4* [239], important in the development and differentiation of beta cells.

4.3 Discussion

The synthesis, folding and processing of insulin is close to the limit of the biosynthetic capacity of beta cells [240]. hIAPP is highly oligomer prone and readily assembles into membrane permeant toxic oligomers if the rate of expression exceeds the capacity of the cell to fold and traffic newly expressed protein. The propensity of IAPP to form toxic oligomers defines the relative vulnerability of a species to develop type 2 diabetes. Taken together, these observations suggest protein misfolding may contribute to beta cell failure leading to type 2 diabetes under conditions of insulin resistance (**Figure 5.14**).

A striking finding from the studies is the high degree of coordinate transcriptome between islets of humans with prediabetes and type 2 diabetes. Since most individuals with prediabetes do not progress to diabetes, these findings imply much of the islet transcriptome in type 2 diabetes may reflect protective pro-survival changes. Furthermore, we found a close overlap in

transcriptome between islets from mice with beta cell overexpression of rIAPP and islets from humans with either prediabetes or type 2 diabetes. Since rIAPP overexpressing mice do not develop diabetes, these findings imply that much of the islet transcriptome in prediabetes and type 2 diabetes reflects adaptive changes to increased beta cell protein synthesis. The importance of successful adaptation of beta cells to increased expression of secretory protein is further illustrated by the large number of genes linked to vulnerability to type 2 diabetes, by GWAS or Mendelian association, that were differentially expressed in rIAPP compared with WT islets.

To better understand the potential role of hIAPP toxicity in beta cell failure and loss in type 2 diabetes, we evaluated the transcriptome of islets in mice overexpressing hIAPP with that in islets of humans with type 2 diabetes. In both hIAPP islets and islets from individuals with type 2 diabetes, there was a strong inflammatory signal, consistent with an ongoing injury response. Network analysis reveals that shared key pro-survival gene networks (**Figure 5.11**) are activated in hIAPP islets and islets from individuals with type 2 diabetes, consistent with the slow progression of beta cell loss in type 2 diabetes. Notably, a decrease in expression of genes that confer beta cells their identity (dedifferentiation) is apparent in both rIAPP mouse and prediabetes human islets, implying that beta cell dedifferentiation maybe a pro-survival adaptation.

By comparing the successful adaptation of beta cells to rIAPP vs hIAPP overexpression, the most prominent difference was the increased expression of genes ascribed to inflammation that were markedly increased in hIAPP islets and type 2 diabetes but only modestly increased in rIAPP islets and in prediabetes. Type 2 diabetes is a heterogeneous disease and there are likely multiple pathways to beta cell toxicity, including toxic actions of lipids and hyperglycaemia [241]. In a recent partial pancreatectomy study in rats, the resulting marked decrease in beta cell

mass and modest hyperglycaemia induced many of the same signals observed in response to beta cell hIAPP toxicity in the present study, including inflammation and partial beta cell differentiation [242]. In an in vitro acute injury study of human islets exposed to glucolipotoxicity, there was also some overlap with islets from humans with type 2 diabetes by RRHO analysis [243], but the overlap was weaker compared with IAPP overexpression in the current study. The difference could be due to in vivo vs in vitro conditions or intrinsic differences between IAPP and glucolipotoxicity.

A limitation of the current study is the use of whole pancreatic islets for RNA-seq, which masks the cell types contributing to the transcriptional signals. Although we used cellular deconvolution and cell marker enrichment methods to mitigate this issue, single cell RNA-seq will be of added value in future studies to confirm the predicted cell type contributions from our deconvolution analysis. In addition, follow-up on the various genes found to be contributing to hIAPP beta cell toxicity through knockdown or overexpression studies will be of use to further confirm their causal role in disease. Another limitation of the present study is the potential bias caused by use of an inbred mouse FVB strain.

In conclusion, the present studies suggest that much of the islet transcriptome in type 2 diabetes is adaptive to the increased beta cell burden of protein synthesis and folding. Beta cell hIAPP toxicity induces a prominent islet inflammatory response, consistent with that observed in type 2 diabetes, implying protein misfolding stress may serve to initiate or contribute to beta cell injury in type 2 diabetes. There are also shared pro-survival gene networks in hIAPP and type 2 diabetes islets. Taken together these studies suggest caution should be taken in interpreting transcriptome changes in islets in type 2 diabetes as therapeutic targets since many, if not most, of these changes are likely

pro-survival adaptations. Strategies to suppress IAPP expression warrant further investigation due to the mounting evidence to suggest its role in type 2 diabetes pathogenesis.

4.4 Methods

Mouse models

Animal studies were approved by the University of California, Los Angeles (UCLA) Office of Animal Research Oversight. The transgenic mice homozygous for human *IAPP* (hIAPP) [219] were originally from Pfizer (available from Jackson Laboratory, Bar Harbor, ME, USA: IMSR cat. no. JAX:008232, RRID:IMSR_JAX:008232) and wild-type FVB (WT) mice (IMSR cat. No. CRL:207, RRID:IMSR_CRL:207) from Charles Rivers Laboratory (Wilmington, MA, USA). The generation of the transgenic mice expressing rodent *Iapp* (rIAPP), human calpastatin (hCAST), and both human *IAPP* and *CAST* (hIAPP:hCAST) on FVB background was described elsewhere [222, 231]. Mice were bred and maintained at UCLA on 12 h day/night rhythm, Harlan Teklad Rodent Diet 8604 (Placentia, CA, USA) and water ad libitum; diabetes was monitored as described [231]. hIAPP transgenic mice develop diabetes (fasting blood glucose > 6.9mmol/l) after 9 weeks of age, while rIAPP mice remained non-diabetic until 18 weeks of age, the end of observation (**Figure 5.15**). Only non-diabetic 9–10-weeks-old male mice were used (Supplement Tables 5.1–5.4). Expression of IAPP (sum of endogenous and transgenic) is comparable in the rIAPP and hIAPP mice [222]. Mice were either subjected to metabolic studies with fasting blood glucose measurements and GTT, or islets and pancreases were collected for analysis of RNA by bulk islet RNA sequencing (RNA-seq) or qPCR, or analysis of protein levels by western blotting (whole cell lysate in RIPA buffer) or immunostaining (4□m thick sections of frozen in OCT 4% paraformaldehyde fixed tissue) [231],[244].

RNA-seq of mouse islets

RNA samples from three mice per group were used for RNA-seq (ESM Table 1). Total RNA was extracted from islet samples (176±11 islets per mouse) using the RNeasy Mini Kit (Qiagen, Germantown, MD, USA). RNA integrity was confirmed using the Agilent Bioanalyzer 2100 (RNA integrity number (RIN) range: 6.8–8.9). RNA-seq libraries were prepared using the TruSeq with Ribo-Zero treatment (Illumina, San Diego, CA, USA) to deplete ribosomal RNA. cDNA libraries were generated using the NuGEN Ovation kit (NuGEN, Redwood City, CA, USA). Illumina's NextSeq 500 platform was used to generate 75 bp, paired-end reads (64 ± 1.4 million reads per sample). Short reads were aligned to the mouse reference genome build GRCm38 (mm10) using the Spliced Transcripts Alignment to a Reference software (STAR aligner) [245]. Between 65% and 75% (mean 70%) of the reads mapped uniquely to the mouse genome. The HT-Seq package [246] was used to count the number of fragments aligned to known exonic regions. Gene expression was measured as total fragment counts per gene. Sample clustering of islet RNA-seq using multidimensional scaling largely reflected genotype (**Figure 5.2**).

RNA-seq of human islets

RNA-seq data from human pancreatic islets were downloaded from the Gene Expression Omnibus (GEO) (GSE50244) [247]**.** Data from 77 samples with available HbA_{1C} values were analysed. Read counts were normalised via the trimmed mean method prior to differential expression analysis using the edgeR package. One type 2 diabetes sample was excluded as an outlier (GSM1216834); therefore 76 samples were included in this manuscript: 51 from normoglycaemic donors with HbA1c levels below 42 mmol/mol (<6%) (HbA_{1c} 5.4±0.1%; BMI 26±0.3; Age 56±2; 18 female/33 male), 15 prediabetic donors with HbA_{1c} levels between 42 and 47 mmol/mol (HbA_{1c} 6.1±0.3%; BMI 26±1; Age 61±2; 6 female/9 male), and ten donors with type 2 diabetes with HbA_{1c}

levels 48 mmol/mol or higher (HbA_{1c} 7.5±0.3%; BMI 30±1; Age 61±3; 6 female/4 male). No additional information about the donors or islet morphology was available.

RNA-seq data analysis

The RNA-seq data from mouse and human islets were subjected to differential expression analysis [248] to find differentially expressed genes (DEGs) between mouse groups and between human groups. Rank-rank hypergeometric overlap (RRHO) [223] analysis was then used to compare human and mouse gene expression signatures to evaluate between-species similarity and differences. In order to understand the biological processes informed by the DEGs, we performed functional enrichment analysis to identify over-representation of Gene Ontology (GO) terms and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways [249]. DEGs from mouse islets were further assessed for enrichment of genes associated with Mendelian or common form of diabetes and metabolic syndrome.

To identify genes with coordinated expression in the form of co-expression modules that are related to diabetes development in mouse models, we utilised weighted gene co-expression network analysis (WGCNA) [80, 250].

To identify cell types contributing to the gene expression changes, we subjected DEGs and co-expression modules (restricting analysis to genes with module membership > 0.5 and false discovery rate (FDR) < 0.05) to two types of analysis. First, we carried out cell type marker enrichment analysis using PanglaoDB cell marker compendium [251] and GeneOverlap R tool [252]. Second, we applied CibersortX [253] to deconvolute mouse bulk islet RNA-seq data into cell type proportions using single cell RNA-seq mouse islet data from GEO (GSM2230762) as reference.

To identify shared regulatory factors between islets from hIAPP transgenic mice and islets from

humans with type 2 diabetes, we performed transcription factor (TF) network analysis using the Enrichr tool [254] with the TF-Gene Co-occurrence extension. To identity additional non-TF regulators, we utilised the key driver analysis function from the Mergeomics R package [45] to identify regulatory genes for the DEG sets and for the co-expression modules using an islet gene regulatory Bayesian network based on a χ^2 like statistic.

Statistical analysis

Statistical analysis was performed as described in the figure legends**.**

4.5 Tables

Table 5.1 Functional characterisation of co-expression modules. Select hub genes with high intramodular connectivity and major biological processes associated with each module by gene set enrichment analysis are reported

Module	Hub genes	Associated biological processes
M1	*Cx3cr1*, *Trf*, *Pla2g7*, *Apoe*, *Trem2*, *Axl*	Immune response, phagocytosis, lysosome
M2	*Tmsb15b2*, *Tk2*, *Micu1*, *Pet112*	Oxidoreductase activity, mitochondria
M3	*Cmklr1*, *Sulf1*, *Mylk*, *Fap*, *Ndrg2*	ERK1/2 cascade, response to growth factor, regulation of angiogenesis
M4	*Arhgap6*, *Tes*, *Myo1b*, *Por*, *Rcan*	Oxidative stress response, protein phosphorylation
M5	*Utp6*, *Slu7*, *Meis2*, *Arhgef9*, *Usp11*	RNA processing
M6	*Nell1*, *Atp8a1*, *Capn9*, *Agt*, *Sorl1*	Oxidative phosphorylation, peroxisome, M phase
M7	*Ttbk2*, *Wdr11*, *Ap1s2*, *Maob*, *Dusp10*, *Cdkn2b*	Cell cycle, microtubule cytoskeleton, phosphatidylinositol signalling
M8	*Ptprz1*, *Jam2*, *Mapk4*, *Calb1*, *Bmp3*, *Bcl2*	Endothelial cell development, cell migration and morphogenesis, MAPK signalling
M9	*Carhsp1*, *Lmna*, *Sqstm1*, *Pink1*, *Psma7*, *Lrp10*	GTPase regulator activity, TGF-β signalling, oxidative stress response
M10	*Setd1b*, *Ncor2*, *Nav2*, *Soga1*, *Vamp2*	Transcription regulation, chromatin organisation, insulin secretion
M11	*Svop*, *Pla2g2f*, *Aldoa*, *Usp7*, *Vcp*, *Sec13*	Protein processing in ER, proteasome, cell cycle regulation
M12	*Ints2*, *Zzef1*, *Gpd2*, *Crhr1*, *Ntrk2*	Protein ubiquitination, chaperonin-mediated protein folding
M13	*Col12a1*, *Mmp2*, *Pld3*, *Cpt1a*, *Calu*, *Nphs1*	Regulation of cell migration, exocytosis, glycerolipid metabolism
M14	*Zfp758*, *Spopl*, *Nemf*, *Clock*, *Slc1a1*	RNA processing, gene expression, G1/S phase
M15	*Gpatch1*, *Nop58*, *Bub3*, *Pdap1*, *Glis1*, *Bag5*	RNA metabolism and splicing, regulation of cell differentiation

4.6 Figures

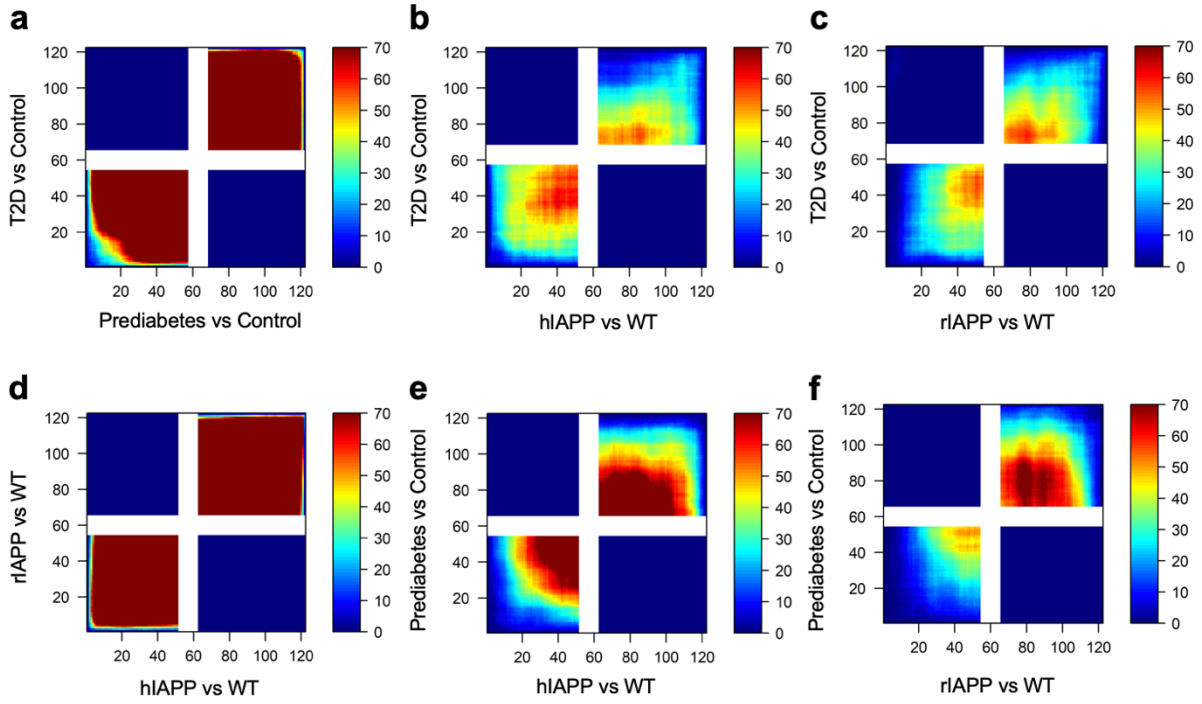

Figure 4.1. Concordant islet transcriptome changes induced by IAPP in mice and in humans assessed by RRHO analysis. Pixels represent the $-\log_{10}(p$ value) of a hypergeometric test (step size=110) and are colour-coded to visualise strength and pattern of overlap. The maximally overlapping sets of upregulated genes (signal in upper right quadrant) and downregulated genes (signal in lower left quadrant) are shown. The expression profile of pancreatic islets from prediabetic and T2D donors (relative to normoglycaemic control) are strikingly similar (**a**), as are the expression profiles of pancreatic islets from the IAPP transgenic mouse models (relative to WT, **d**). The expression profile of pancreatic islets from IAPP transgenic mice (relative to WT) is highly concordant with the islet from humans with T2D (**b**, **c**) and prediabetes (**e**, **f**). The numbers of genes subjected to RRHO analysis and concordantly changed are listed in Figure 5.7. T2D, type 2 diabetes.

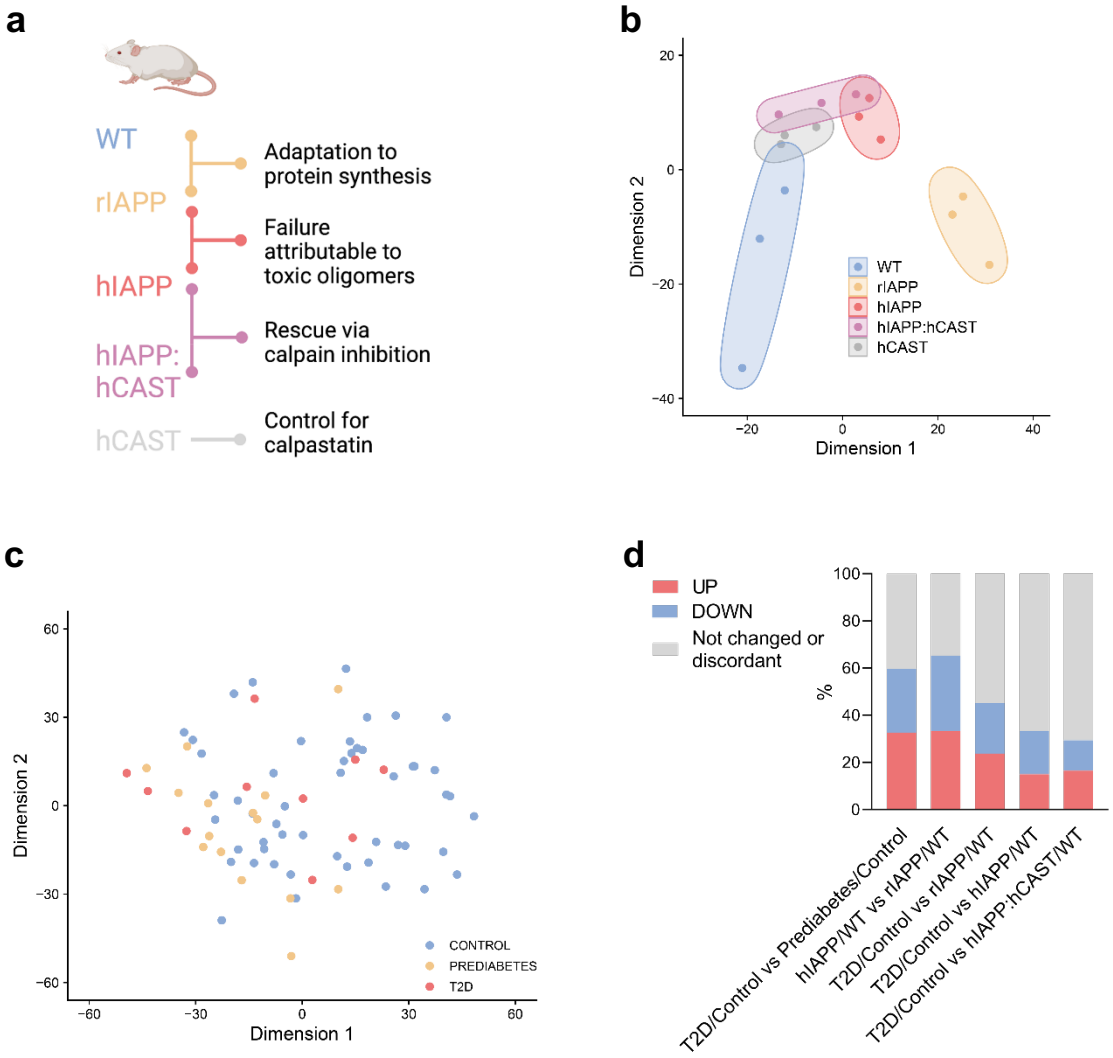

Figure 4.2. (**a**) Schematic depicting experimental and control groups, with rationale and expected output for each comparison. Created with Biorender.com. (**b**) Multidimensional scaling map of islet profiles shows clustering of samples is largely influenced by genotype of mice. (**c**) Multidimensional scaling map of human islet samples. (**d**) The proportion of genes concordantly up- and down-regulated obtained from RRHO analysis of listed pairs of comparisons.

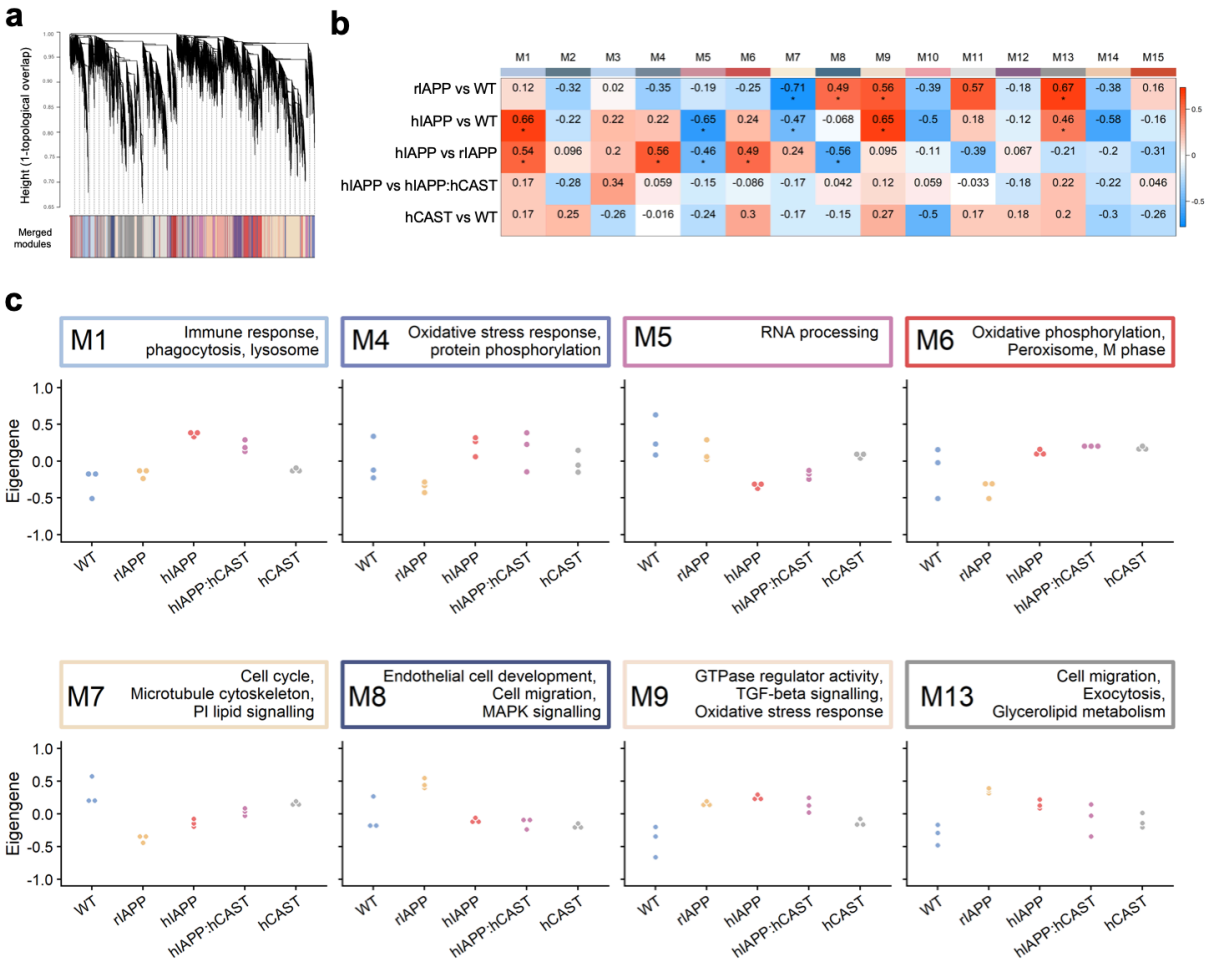

Figure 4.3. Co-expression network construction and analysis. (**a**) Hierarchical cluster dendrogram generated using all samples grouped genes into 15 distinct co-expression modules (M1–M15, labelled with colours). (**b**) Module-trait relationships were assessed by fitting a generalised linear model based on IAPP and CAST status, then comparing module eigengene (ME)—equivalent to the first principal component of a module—between genotype pairs. Module-level differential expression was tested by one-way, nonparametric ANOVA followed by post hoc Tukey test. Differences in ME expression are presented as a heat map, with significant perturbations denoted (*, q<0.05). (**c**) Trajectory plots of perturbed modules display normalised expression across all samples.

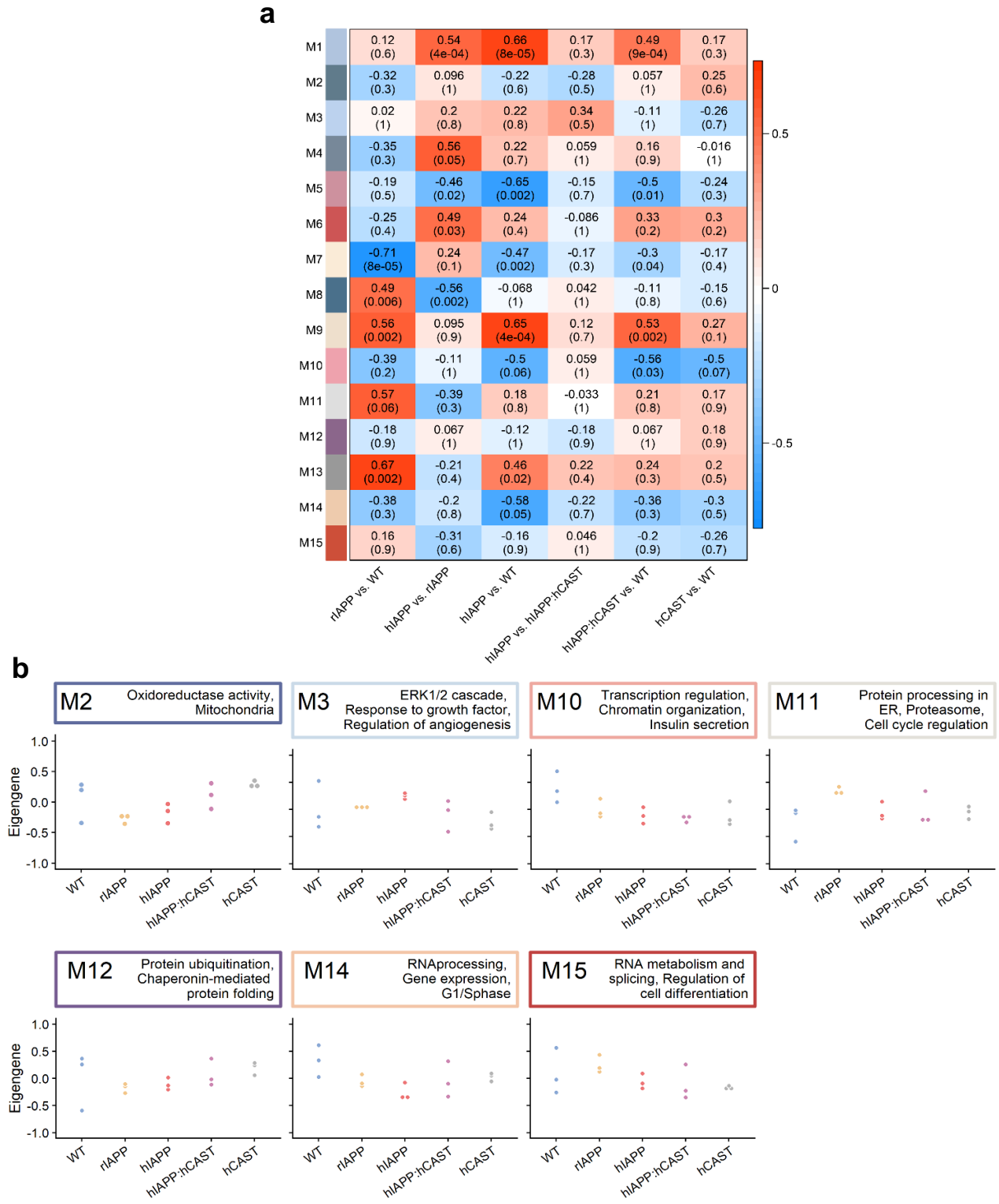

Figure 4.4 Co-expression network construction and analysis. (**a**) Heatmap displays correlation coefficient between module expression and pairwise islet comparisons. Significant module level perturbations are denoted (*, $q < 0.05$). (**b**) Plots of ME trajectory by sample group, for seven modules not shown in the main text.

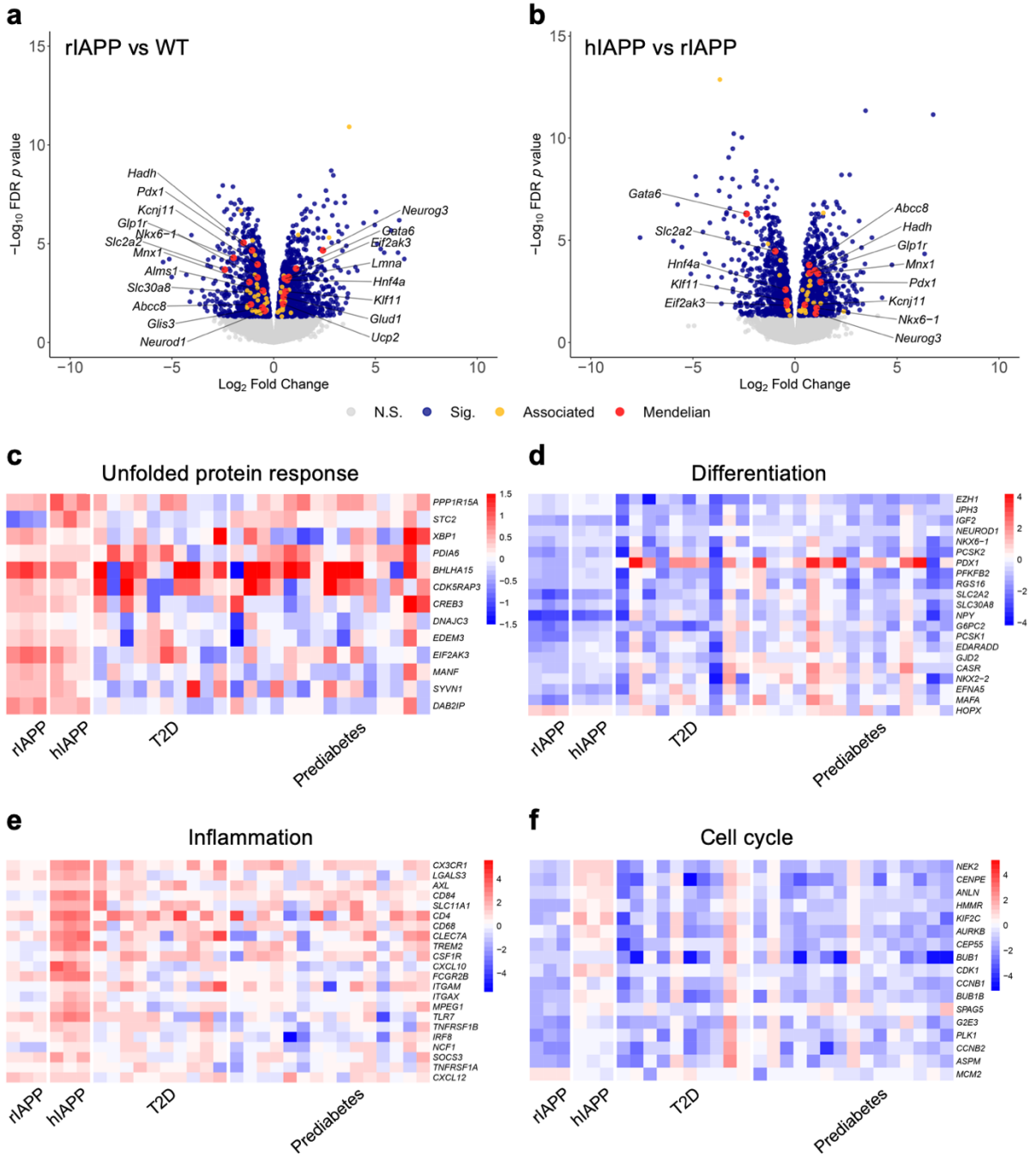

Figure 4.5 Transcriptomic profiles of adaptation to increased secretory workload, and failure in context of protein misfolding toxicity. Volcano plots show relative expression (Log_2 fold change) of 15,731 transcripts plotted against the adjusted *p* value from differential expression. (**a**) Comparing rIAPP islets with those of WT defines the expression profile of islets successfully compensating for increased soluble IAPP. Several genes implicated in Mendelian disease are dysregulated (red, labelled), as are genes linked to type 2 diabetes by GWAS (yellow). (**b**) hIAPP islets compared with rIAPP highlights expression dysregulation corresponding to IAPP-derived oligomer toxicity, now controlling for increased beta cell workload. (**c**–**f**) Successful and failed

adaptation to increased beta cell secretory pathway burden involves activation of the adaptive UPR, inflammation, altered expression of cell cycle-associated genes, and beta cell dedifferentiation. (**c**) Islets of rIAPP mice show enhanced upregulation of key UPR genes compared with hIAPP. Some UPR-related genes appear to be upregulated in both T2D and prediabetes. (**d**) Increased beta cell workload leads to downregulation of key beta cell function and maturity markers, with rIAPP islets demonstrating more profound 'dedifferentiation' than hIAPP islets. (**e**) Increased hIAPP results in transcriptional upregulation of inflammation-associated genes, including several macrophage markers, which is partially attenuated by concurrent overexpression of calpastatin. (**f**) Increased secretory burden drives downregulation of cell cycle-associated genes islets from individuals with T2D (HbA_{1c} level above 48 mmol/mol (>6.5%)) and donors with prediabetes (HbA_{1c} levels between 42 and 48 mmol/mol (6%<HbA_{1c}<6.5%)), as well as rIAPP islets, but not in hIAPP. Data are expressed as a ratio of individual to the mean of WT islets for mouse models, and Control (HbA_{1c} level below 42 mmol/mol (<6%)) for human islets. T2D, type 2 diabetes

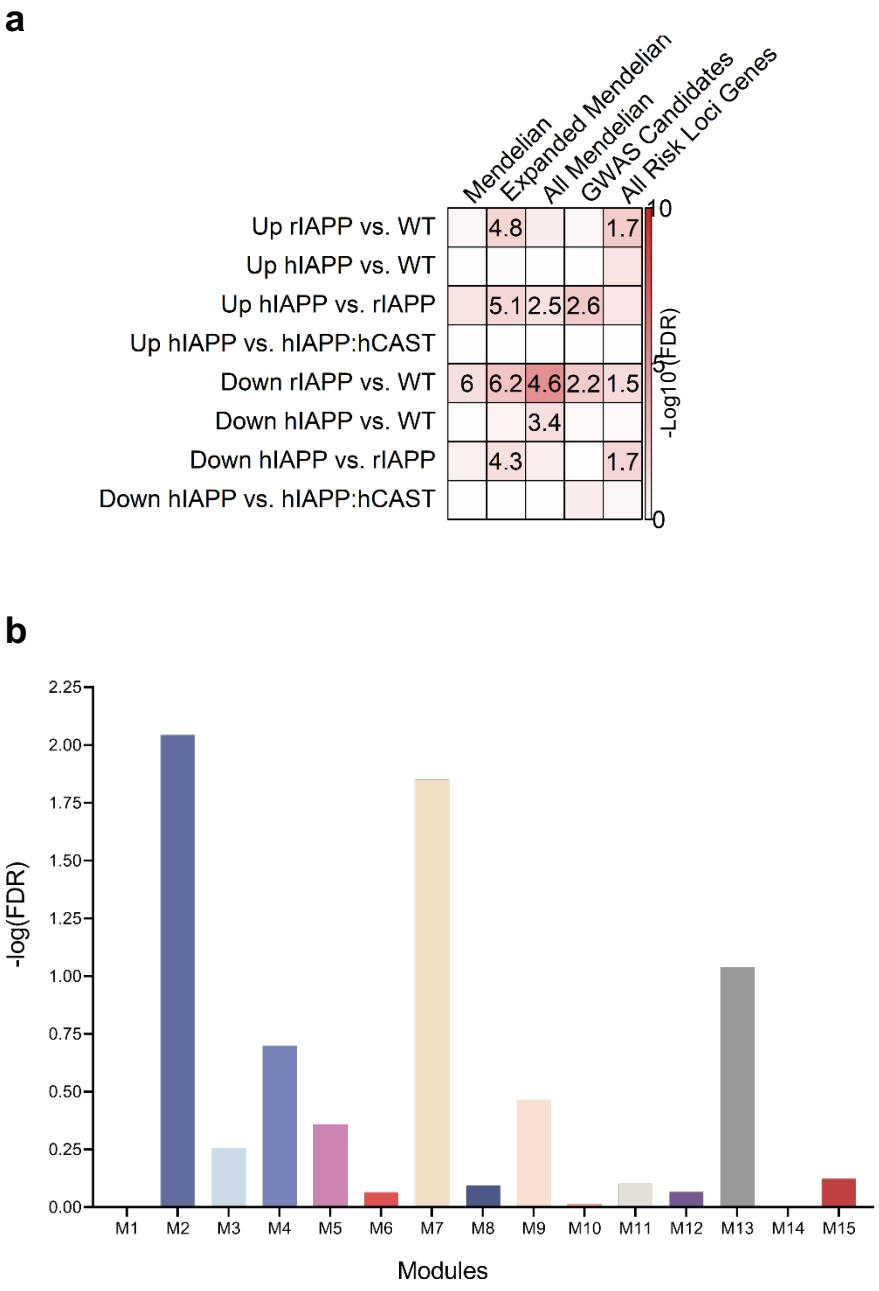

Figure 4.6. (**a**) Common and rare variant enrichment of DEGs. (**b**) Type 2 diabetes GWAS enrichment for 15 WGCNA co-expression modules.

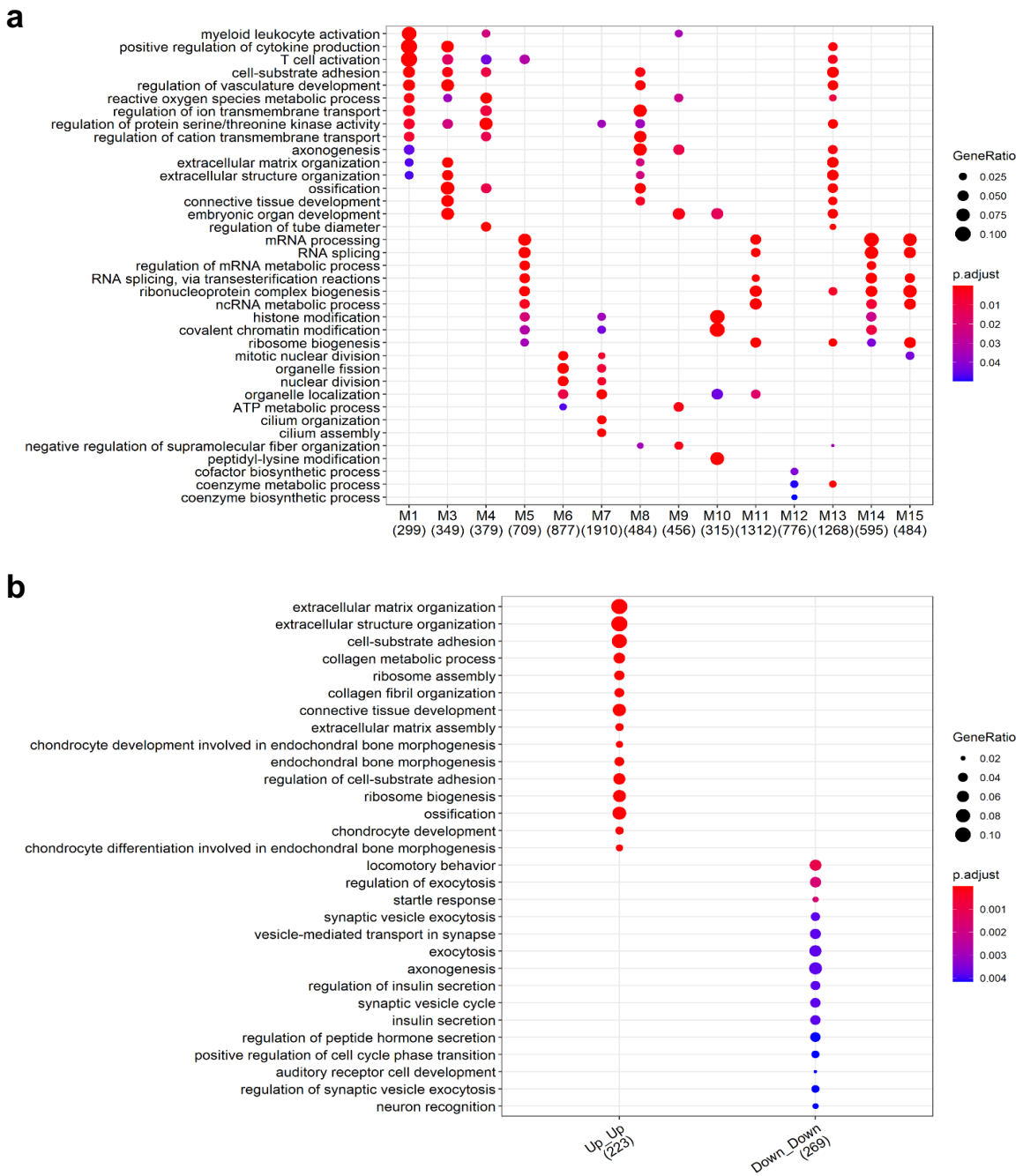

Figure 4.7 (**a**) GO biological process term enrichment of co-expression modules. Enrichment analysis and visualization were performed using ClusterProfiler. No overrepresented terms were identified for M2 (excluded from visualization). (b) Functional annotation of differentially expressed genes (FDR < 0.05) that are upregulated (Up_Up) or downregulated (Down_Down) in both hIAPP/WT and rIAPP/WT. GO biological process term enrichment of DEG sets.

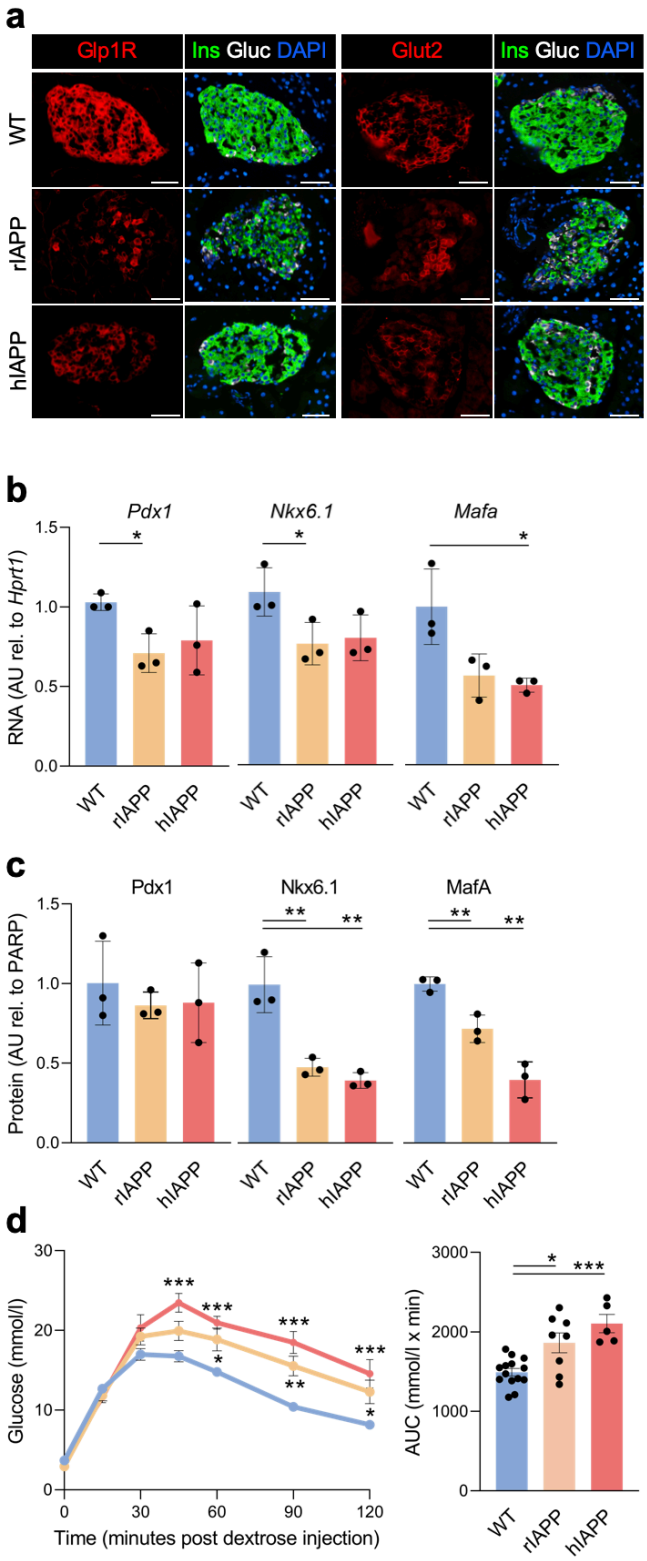
a
Glp1R
Ins Gluc DAPI
Glut2
Ins Gluc DAPI
WT
rIAPP
hIAPP
b
Pdx1
Nkx6.1
Mafa
RNA (AU rel. to Hprt1)
1.5
1.0
0.5
0.0
*
WT
rIAPP
hIAPP
c
Pdx1
Nkx6.1
MafA
Protein (AU rel. to PARP)
**
d
Glucose (mmol/l)
30
20
10
0
***
*
**
0
30
60
90
120
Time (minutes post dextrose injection)
AUC (mmol/l x min)
3000
2000
1000
0

Figure 4.8 (**a**) Immunohistochemistry staining of islets for Glp1r and Glut2 highlights reduced levels of proteins involved in beta cell secretion. Co-staining for insulin (Ins) and glucagon (Gluc) show comparable cell type composition in rIAPP and hIAPP islets. Scale bar, 50 μm. (**b, c**) RNA-seq identified downregulation of the key beta cell TFs (*Nkx6-1*, *Pdx1, Mafa*) in rIAPP and hIAPP islets, and their RNA and protein level expression were tested by qPCR (**b**) and western blotting (**c**), respectively. Data are the mean±SEM, *n*=3 in each group, two-tailed Student's *t* test: *p<0.05, **p<0.01. (**d**) IPGTT, 2 mg dextrose/g of body weight after overnight fast; both hIAPP and rIAPP mice display impaired glucose tolerance compared with body weight matched WT, with the greatest effect observed in hIAPP mice. Data are the mean±SEM, *n*=5–15 per group; one-way ANOVA followed by post hoc analysis: *p<0.05, **p<0.01, ***p<0.001. Separate islet samples from non-diabetic 9-week-old mice were used to generate the data presented in each panel, and they were different from RNA-seq samples (Supplement Tables 5.1–5.3)

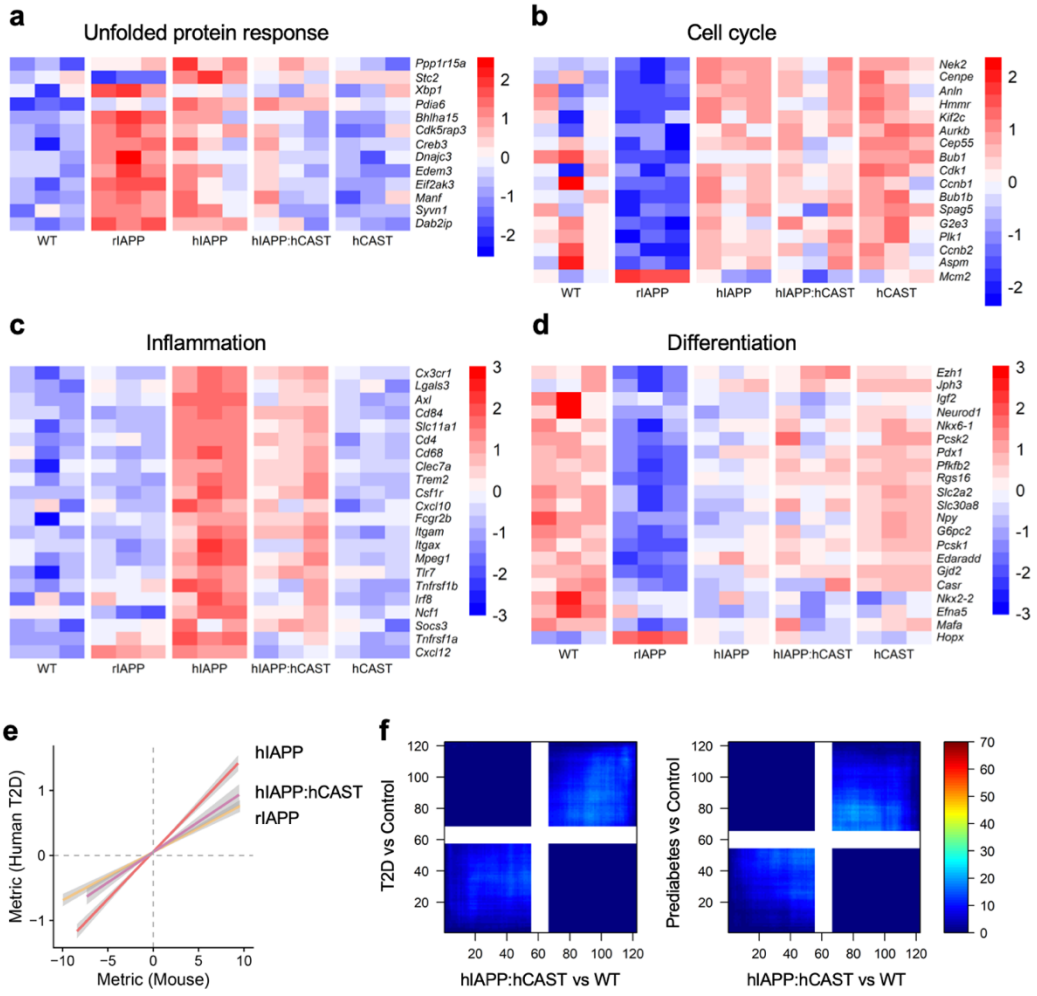

Figure 4.9. Effect of calpain hyperactivation on gene expression. (**a**–**d**) Increased expression of calpastatin in beta cells from hIAPP mice partially rescues phenotype related to UPR, inflammation, cell cycle and beta cell dedifferentiation. (**e, f)** Comparison of islet gene expression profiles affected by IAPP toxicity and calpain hyperactivation with those in human type 2 diabetes. (**e**) Genes measured in two independent experiments are ranked according to degree (nominal *p* value) of differential expression relative to the appropriate control group, multiplied by the sign of the fold change. The type 2 diabetes islet profile is best reflected by the hIAPP islet profile, outperforming the rIAPP and hIAPP:hCAST profiles. (**f**) As an alternative to correlation analysis, we applied the RRHO2 algorithm to test preservation of IAPP toxicity and calpain hyperactivation signatures in islets from humans with type 2 diabetes and prediabetes. Serial hypergeometric tests were performed at gene rank threshold for two ranked lists. The RRHO map was generated by $-\log_{10}$ transformation of the hypergeometric test *p* value (step size=110), and pixels are colour-coded to visualise strength and pattern of overlap. After accounting for the transcriptomic impact of calpain hyperactivity (hIAPP:hCAST), overlap signal between the islet profiles in type 2 diabetes and prediabetes with the hIAPP mouse model of type 2 diabetes (Figure 5.1) significantly

decreases, implying a role for calpain in propagating the inflammatory response in pancreatic beta and other cell types in prediabetes and T2D. T2D, type 2 diabetes

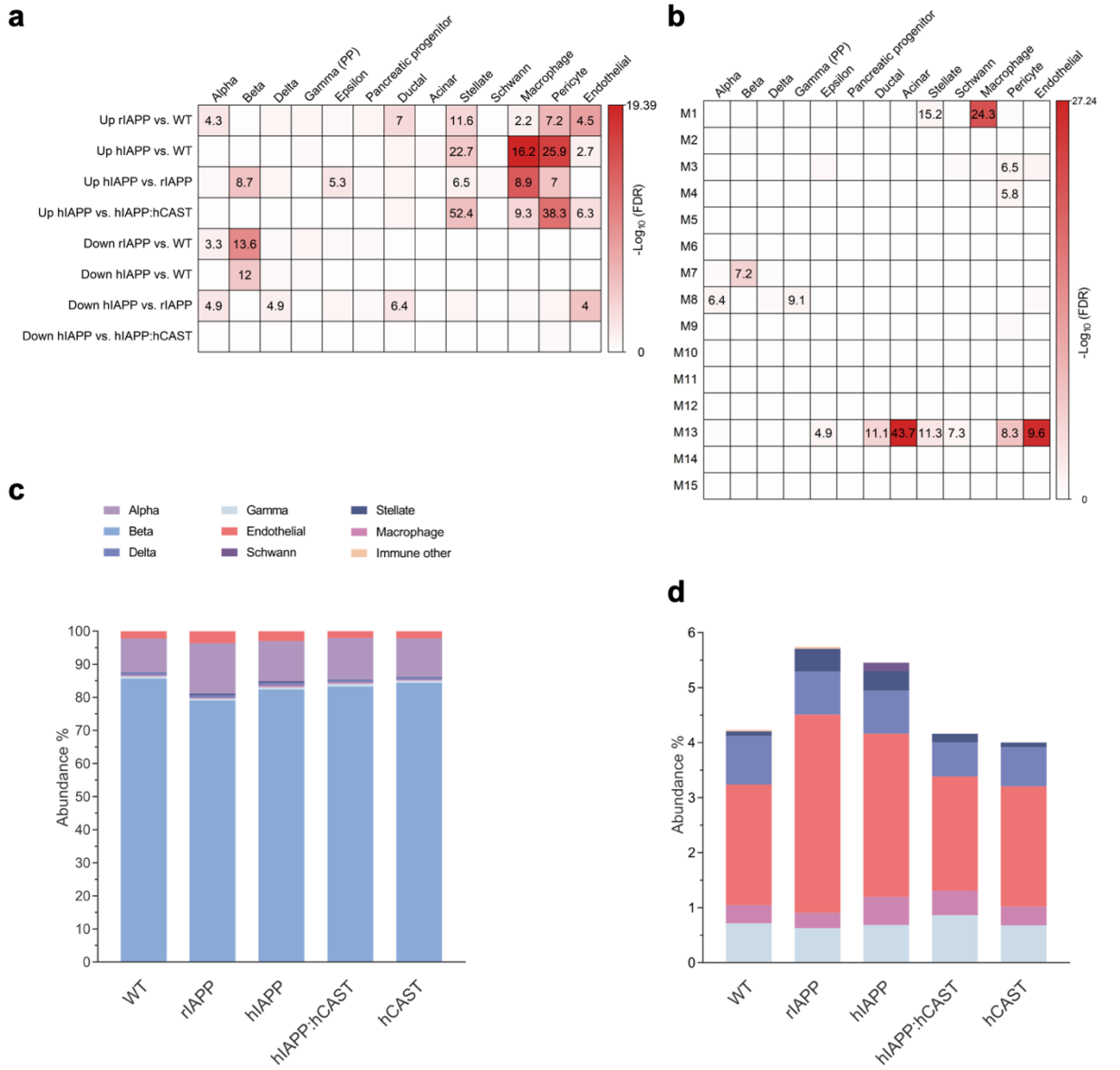

Figure 4.10. (**a**) Cell type marker enrichment of DEGs highlights substantial contribution of non-endocrine cells toward the composite islet profile in the bulk RNA-seq data**.** (**b**) Cell type marker enrichment of modules showcases module links with non-endocrine and endocrine cell types. Module enrichment for high-specificity cell type markers was evaluated by Fisher's exact test. Colour represents the $-\log_{10}$ FDR-corrected *p* value, with the OR provided for FDR<0.05. (**c**) Deconvolution of bulk islet RNA-seq revealed the relative abundance of each cell type captured, highlighting beta cell dominance across the genotypes. (**d**) Magnified view of the deconvolution of bulk islet RNA-seq results on less abundant cell types (alpha and beta cells were excluded) highlighting an increase in endothelial cells, stellate cells and macrophages in hIAPP compared with WT

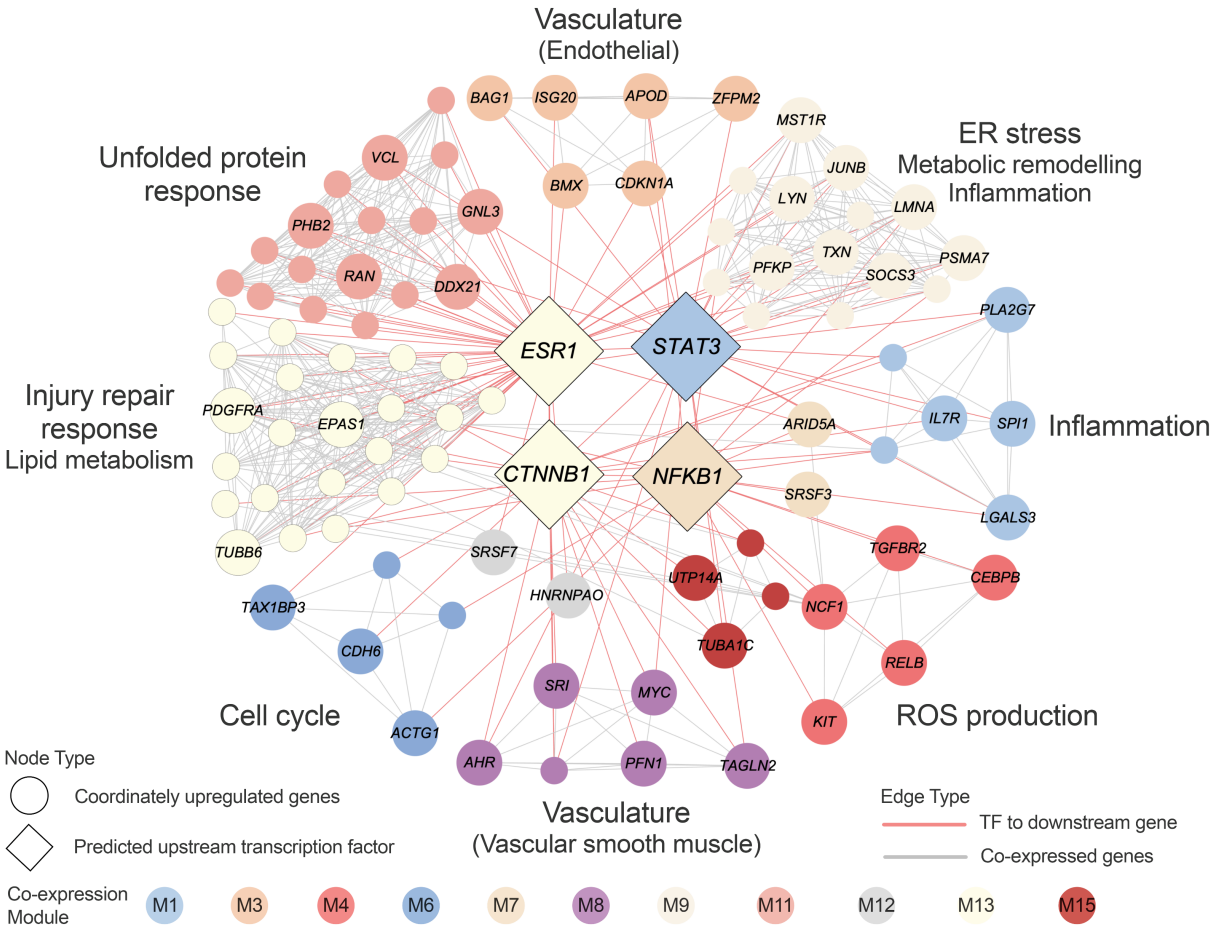

Figure 4.11. Putative regulatory network of genes co-ordinately upregulated in hIAPP and in human type 2 diabetes islets relative to their respective controls (WT and non-diabetic human islets), identified by RRHO analysis. TF binding sites enrichment analysis identified over-represented upstream TFs including NF-κB1, assigned to beta cell-enriched module (M7), and STAT3, a key regulator of inflammation assigned to macrophage- and stellate cell marker-enriched M1. ESR1 and CTNNB1 are both implicated in beta cell stress/survival signalling. Node colour reflects co-expression module assignment. Edges represent experimentally validated transcription factor-target relationships (red) and intramodular co-expression (light grey)

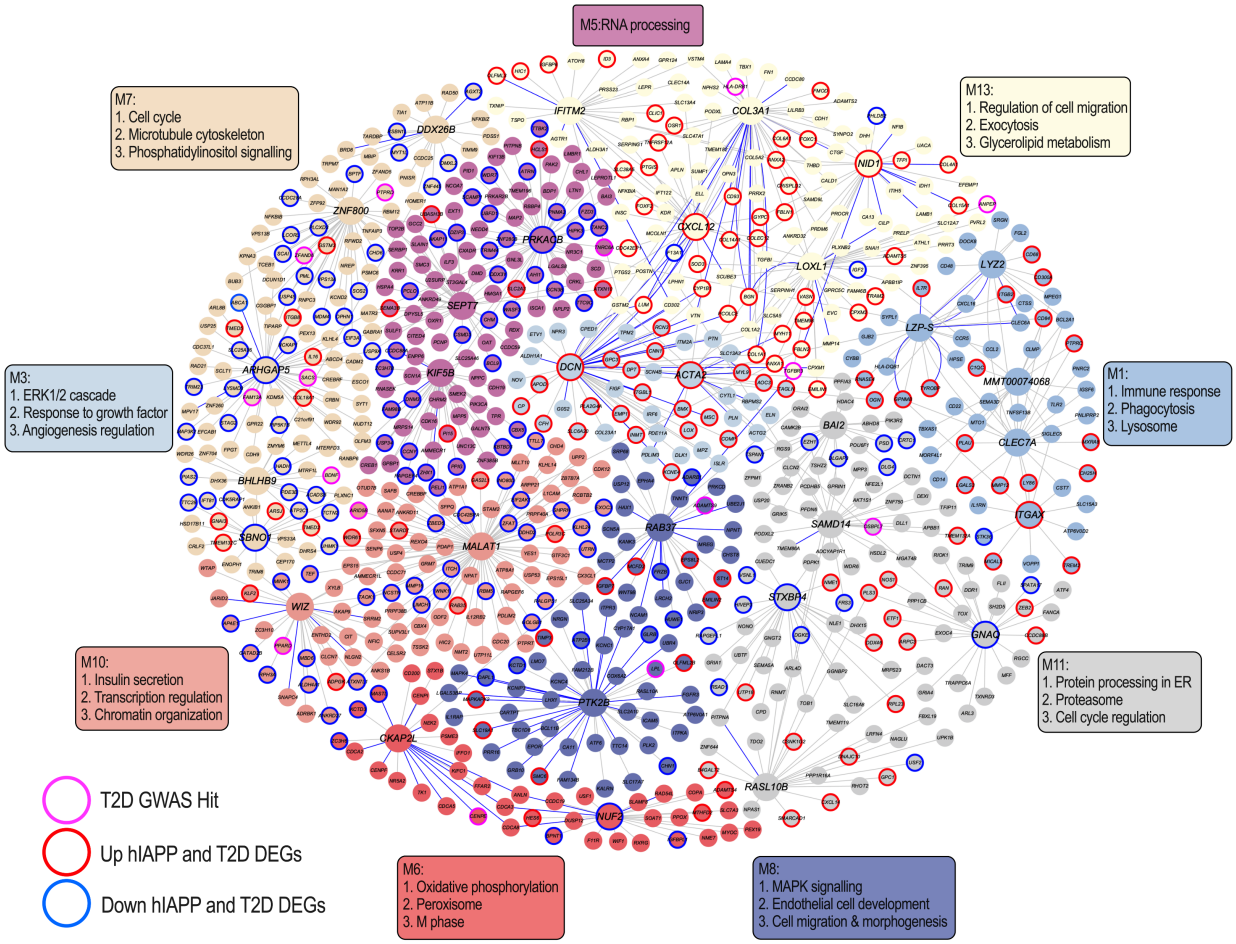

Figure 4.12 An islet Bayesian gene regulatory network illustrating the co-expression module interconnectivity and highlighting key genes potentially important in driving those processes (indicated by the larger node size). T2D GWAS hits (association $p<5\times10^{-8}$) are highlighted on the network with pink rings around the nodes, upregulated hIAPP and T2D DEGs are highlighted with red rings, and downregulated hIAPP and T2D DEGs are highlighted by blue rings. T2D, type 2 diabetes

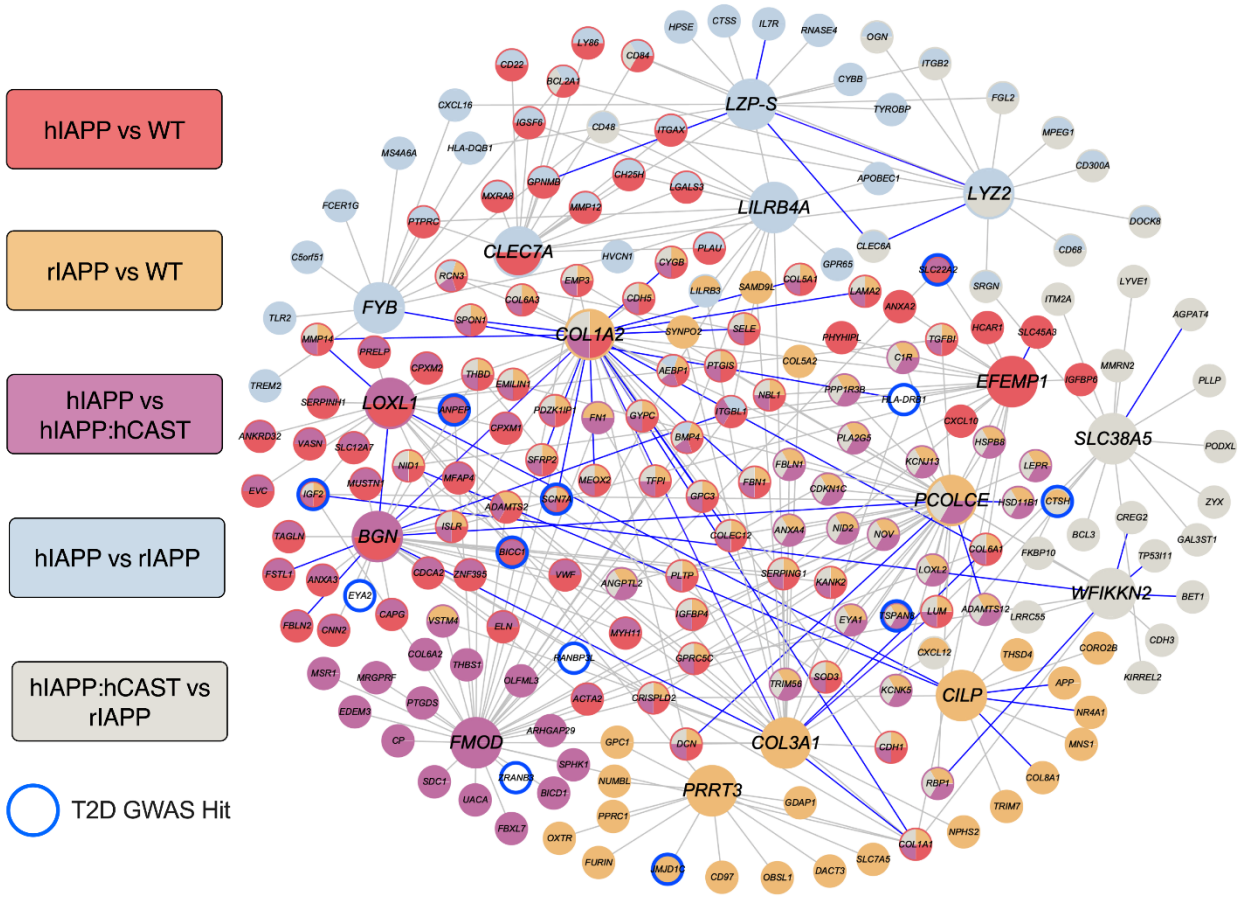

Figure 4.13 Gene-gene regulatory subnetwork (Bayesian Network) and top key drivers of differentially expressed genes from IAPP sequencing between the various genotypes (color indicates comparison(s) in which a gene is differentially expressed; FDR < 0.05). Blue edges indicate connections directly derived from an islet constructed Bayesian network, whereas the grey edges are derived from a brain Bayesian network. Large nodes indicate key driver genes.

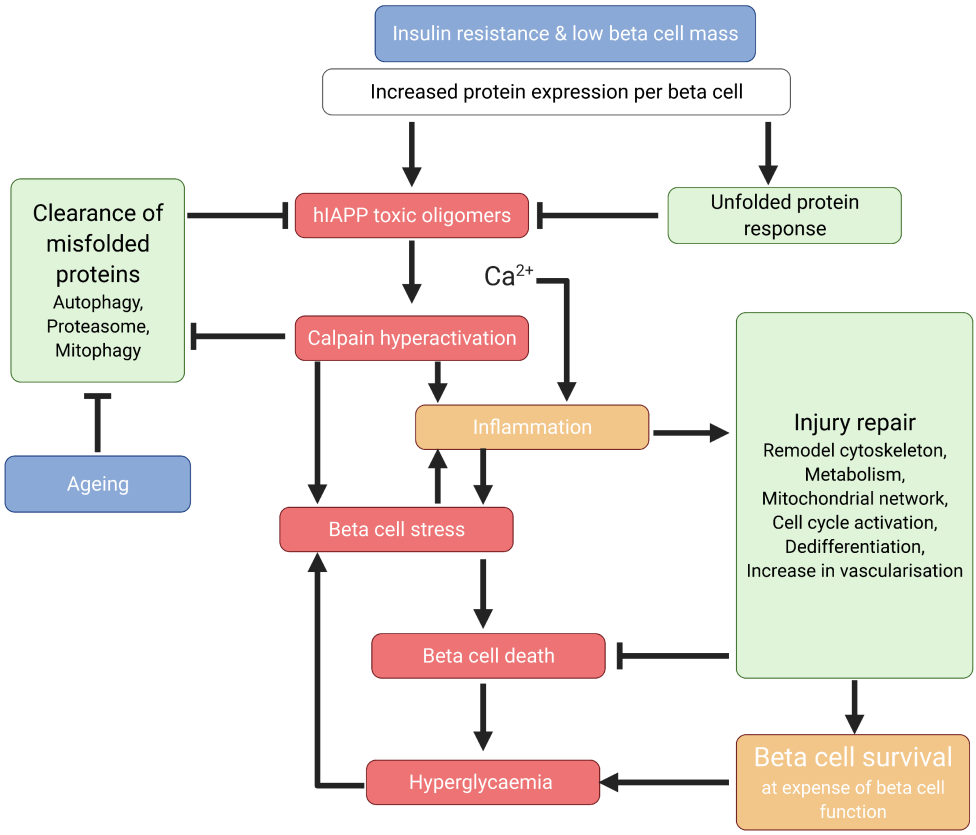

Figure 4.14 Proposed model of IAPP toxicity in type 2 diabetes in relation to the major risk factors insulin resistance and a low innate beta cell mass, which result in very high expression levels of aggregate toxic oligomer-prone IAPP per beta cell in humans. Clearance of misfolded IAPP by autophagy and proteasome declines with ageing. Increased IAPP and insulin expression induces the protective UPR. Membrane permeant toxic oligomers of IAPP lead to aberrant Ca^{2+} signalling that induces injury inflammatory responses directly and via calpain hyperactivation. These initially activate conserved pro-survival injury repair signalling responses that prolong beta cell survival at the expense of function. However, the adverse actions of calpain hyperactivation on defence against proteotoxicity exacerbates IAPP toxicity, gradually overcoming pro-survival responses

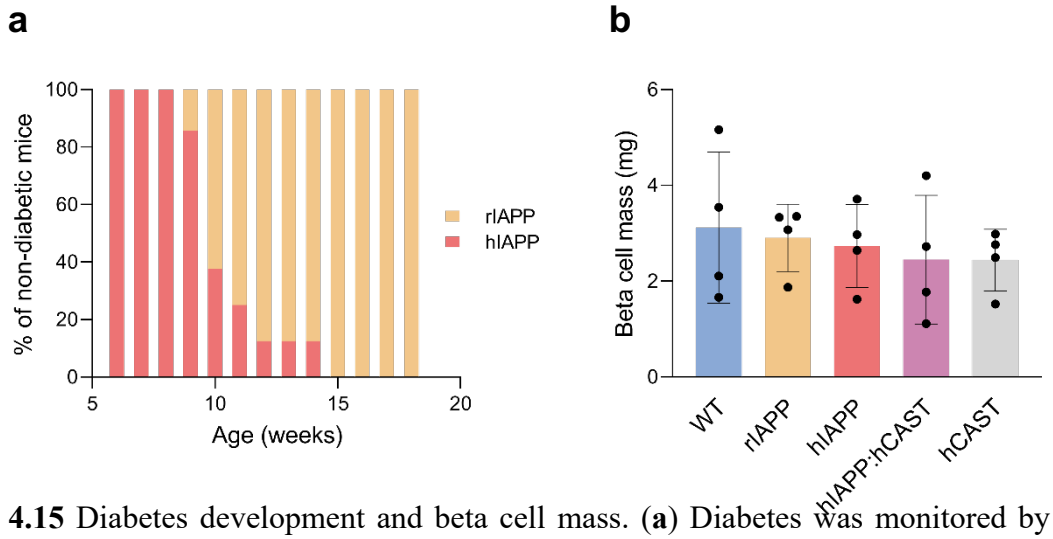

Figure 4.15 Diabetes development and beta cell mass. (**a**) Diabetes was monitored by weekly measurement of tail vein blood glucose after overnight fast in 6-18 weeks-old mice. A mouse was considered diabetic if fasting blood glucose was >6.9mmol/l. Data is % of non-diabetic mice; n=5-16 for rIAPP and 5-14 for hIAPP group per column. (**b**) Beta cell mass in 9-weeks-old mice with fasting glucose and body weight matching to mice used for RNA-seq analysis was comparable in all groups. Mice characteristics are presented in Supplement Table 5.4.

Chapter 5. Sex differences in NASH pathways informed by multi-omics

5.1 Introduction

Non-alcoholic fatty liver disease (NAFLD) encompasses a range of pathologies, from the relatively benign steatosis to the more severe non-alcoholic steatohepatitis (NASH) with or without fibrosis, which can further develop to cirrhosis and eventually hepatocellular carcinoma (HCC). NAFLD has become a significant health burden and continues to increase on a year-by-year basis with currently around 25% of the population falling within the NAFLD definition. While the disease mechanism has been somewhat elucidated, many holes still remain which is largely responsible for the lack of FDA-approved drug options for NAFLD. One area that has had limited exploration is the underlying sex differences that contribute to NAFLD development [255].

Current understanding of the sex differences within NAFLD development presents that males generally are at a higher risk than females in developing NAFLD as well as presenting a more severe phenotype. Following menopause, however, females are found to develop NAFLD at a higher rate compared to males. This difference has been partially attributed to the protective effects of estrogen as well as metabolic differences between the male and female livers though there is still much to uncover [255]. Our previous research [165], in line with others, has uncovered mechanistic differences in the development of steatosis between sexes, and therefore, approach to treatment and biomarkers in a sex specific manner may be beneficial [174]. NAFLD is projected to be the number one cause for liver transplantation over the next few years, where the major cause of liver transplantation is liver fibrosis. Liver fibrosis is the typical response to chronic liver disease and is characterized by an increase in extracellular matrix (ECM) constituents that collectively form the hepatic scar. To develop target therapies to reverse the fibrotic response and improve the outcomes of patients with chronic liver disease, it is important to uncover the mechanisms that

underlie liver fibrogenesis. Briefly, the cellular and molecular mechanisms of hepatic fibrosis include the activation of hepatic stellate cells (HSCs), which secrete autocrine and paracrine growth factors, chemokines, and ECM. The prominent transcriptional targets during HSC activation include type I collagen, a-SMA, TGFβ1, TGFβ receptors, MMP2, TIMP1, and TIMP2. Among the transcription factors that activate these downstream targets are Ets1, Mef2, CREB, Egr1, Vitamin D receptor, Foxf1, JunD, and C/EBPβ. However, the full picture behind fibrosis is far from elucidated with a desperate need for better biomarkers and drug treatment options.

To comprehensively understand the mechanisms in hepatic fibrosis, we utilize a systems genetics approach, allowing us to examine its associated genetic factors and pathways, systematically taking into account potential sex differences too. Hui et al. reported a systems genetics analysis of NASH/fibrosis using a hyperlipidemic human CETP and APOE*3-Leiden transgenic mice that develop many features and molecular signatures characteristic of human NASH pathophysiology [256]. Overall, 619 male mice from 102 strains of the hybrid mouse diversity panel (HMDP) were surveyed. This cohort of mice was termed the fibrosis HMDP. A wide spectrum of fibrosis, predominantly pericellular fibrosis, was observed among the strains, showing that hepatic fibrosis in mice is strongly dependent on genetic background, and hepatic steatosis and NASH/fibrosis are mediated by distinct genetic factors, consistent with a multistep model of NAFLD development. Through genome-wide association mapping, significant genetic loci uniquely associated with fibrosis were identified, including rs50309490, rs31853140, and rs29935539. This study produced a rich multi-omics data resource for liver fibrosis, including dense genotyping of common genetic variants, liver transcriptome data, and the corresponding expression quantitative trait loci (eQTLs) that reflect genetic regulation of gene expression. To add on from this study, 232 female mice from 102 strains of the same transgenic background as

the males were also conducted with the same analyses and data generated to allow for direct comparison of sex differences.

To go beyond the top significant genetic loci and better understand the underlying pathways and key drivers for liver fibrosis using the HMDP, we applied an integrative genomics approach to fully incorporate the whole spectrum of liver fibrosis genetic association with functional genomics information from liver fibrosis eQTLs and from gene networks constructed using liver transcriptome data from 102 strains as well as from a multitude of existing genomic studies. This multi-omics integration revealed coordinated gene-gene interactions in liver tissues that are perturbed by polygenic risks of liver fibrosis and uncovered hidden biology missed by traditional genomic analysis as well as key sex differences.

This data-driven integrative approach not only highlighted shared mechanisms for liver fibrosis such as ECM related pathways, but also revealed mechanistic differences between sexes, with males possessing more immune related processes as well as potential protein metabolism perturbations and females possessing more carbohydrate metabolism perturbations as well as showing enrichment for diseases with known enzyme deficiencies for long chain sugars. These results are in line with the phenotypic differences seen between males and females, with males showing a much more severe fibrosis, guiding us to realize that the mechanisms resulting in fibrosis may differ between sexes. This is further highlighted by the differences found when exploring the key driver genes revealed utilizing Bayesian network modeling. Additionally, using the key driver genes uncovered, we perform drug repositioning in an attempt to identify potential therapeutic treatments for NAFLD.

5.2 Methods

Study Overview

We modeled fibrosis gene networks using the multi-omics HMDP data along with additional public gene expression datasets to identify pathways and predict potential 'key driver' genes underlying hepatic fibrosis in both males and females (**Figure 6.1**). In brief, we first constructed gene co-expression networks based on liver fibrosis expression data across the HMDP strains. Then, we integrated these networks along with curated canonical pathways (KEGG, Reactome, and Biocarta) and GWAS analyses of hepatic fibrosis as well as fibrosis eQTL information using the Mergeomics platform [257, 258]. This integration led to the identification of co-expression modules (groups of co-expressed genes) and canonical biological pathways that are enriched for hepatic fibrosis GWAS signals. In addition, we utilized the correlation of liver transcriptomics with liver fibrosis to understand the current biological pathways enriched within the diseased mice to complement the genetically linked approach. Subsequently, we mapped the fibrosis-associated network modules and pathways to gene regulatory Bayesian networks of liver tissues that are based on numerous genetics and gene expression datasets to predict potential key regulators, termed key drivers (KDs), of the hepatic fibrosis processes. We then prioritized the resulting predicted KD genes for experimental validation and mechanistic studies in mice. Taking the KDs identified, we also performed drug repositioning using our PharmOmics platform in order to uncover potential therapies for NAFLD [64].

GWAS and eQTL Analyses

For 102 mouse strains, genotypes were obtained from The Jackson Laboratory using the Mouse Diversity Array (Yang et al., 2009). Single Nucleotide Polymorphisms (SNPs) were removed if

they had a minor allele frequency (MAF) < 5% and a missing genotype rate of > 10% as well as being flagged for poor quality, resulting in around 200,000 SNPs as previously described [256]. GWAS mapping of fibrosis liver in the HMDP as well as tissue-specific eQTLs were previously generated in Hui et al., 2018 [256]. Using the Factored Spectrally Transformed Linear Mixed Models (FaST-LMM) approach, both GWAS and eQTLs were generated as previously described. eQTLs utilized were of a cis background defined as those within a ±1Mb region of the transcription start and end sites of the genes. P values were adjusted to estimate the false discovery rate (FDR) to correct for multiple testing. We included 258,743 male-specific cis-eQTL associations (84,164 unique cis-eSNPs and 2,463 cis-genes) in liver at P<1E-6 (FDR<0.01) and 216,706 female-specific eQTLs (78,300 unique cis-eSNPs and 2,054 cis-genes) in the current study.

Construction of co-expression modules from liver transcriptome data

Sex-specific liver transcriptome co-expression modules were constructed from female-specific gene expression data and male fibrosis HMDP data. To construct these co-expression modules, we utilized two different methods based on hierarchical clustering to identify co-regulated gene sets for liver fibrosis: Multiscale Embedded Gene Co-expression Network Analysis (MEGENA) and Weighted Gene Co-expression Network Analysis (WGCNA). Collectively, we generated a total of 150 female co-expression modules and 62 male co-expression modules. WGCNA is the more commonly used of the two network methods and has shown importance with inferring biologically relevant processes. At the same time, WGCNA generally produces very large modules with many genes where a gene can only be allocated to one module, not multiple, which is not reflective of true biology. To account for this limitation, we utilize MEGENA, which produces smaller, more coherent modules to capture more discrete biological processes as well as allow genes to be

assigned to multiple modules. Therefore, utilizing both approaches is complementary and provides the full potential to capture any unknown biology as well as confirm known processes.

Both network methods utilize hierarchical clustering to identify co-regulated gene sets from the correlations of gene pairs where they will ultimately assign the co-expressed genes into modules. The difference is that WGCNA is based on agglomerative clustering in which genes are clustered by merging, whereas MEGENA utilizes divisive clustering in which genes are clustered by splitting. In both, gene clustering is determined by a distance measure: one minus topological overlap matrix (TOM) in WGCNA, conferring dissTOM = 1 - TOM, and shortest path distance (SPD) in MEGENA. For WGCNA, the distance between two clusters is calculated by taking the average of the dissTOM scores of all gene pairs (one gene from each cluster), where TOM is determined by considering the correlation scores, or edge weights, between two genes, or nodes, as well as that of their common neighbors. For MEGENA, a nested k-medoids clustering is used, which seeks to minimize SPD in each k-best cluster, running until no further child clusters can be identified. MEGENA also differs from WGCNA in that it executes multi-scale clustering, allowing us to obtain alternate modules at different scales while employing the same input. This allows us to assign a single gene into multiple modules, whereas WGCNA is limited to one gene-one module.

Functional Annotation of the NAFLD Correlated Co-expression Modules

To annotate the liver fibrosis-based co-expression modules with corresponding biological pathways, we utilized MatrisomeDB, Biocarta, Reactome, KEGG, and PID databases from the MSigDB via the hypergeometric test. Adjusted P-values were obtained using Bonferroni

correction. Pathways passing an adjusted p < 5% and shared a gene number > 5 were considered as significant pathways.

Knowledge-Based Biological Pathways Curated

We used a total of 1827 canonical pathways from Reactome (Version 45), Biocarta (Nishimura, 2001) and the Kyoto Encyclopedia of Genes and Genomes (KEGG) databases.

Liver Transcriptomic Data Correlated with Liver Fibrosis

To provide potential insights into the most relevant genes associated with liver fibrosis, we correlated liver gene expression data with the fibrosed area of the liver up to a p-value cutoff < 1e-5 using the bicor function from the WGCNA R package. Liver fibrosis was measured as a quantitative trait as previously described in Hui et al., 2018 [256].

Marker Set Enrichment Analysis (MSEA)

To identify co-expression modules and pathways that show evidence for genetic association and gene expression correlation with liver fibrosis, we performed a marker set enrichment analysis (MSEA) via the Mergeomics R package, using GWAS in conjunction with the eQTL data and correlated genes with fibrosis separately. Utilizing this approach would provide two lists of co-expressed modules and pathways: 1) genetic and causal pathways to fibrosis and 2) pathways affected and linked with fibrosis through transcriptomics. MSEA employs a chi-square-like statistic with multiple quantile thresholds to assess whether a co-expression module or pathway shows enrichment of functional disease SNPs (i.e., SNPs that likely regulate gene expression as captured in eQTLs) compared to random chance. 10,000 permuted gene sets were generated for

each co-expression module and pathway. As detailed in Shu et al., the enrichment statistics from the permutations were used to approximate a Gaussian distribution from which enrichment P-values were determined. Benjamini-Hochberg (BH) false discovery rate (FDR) was estimated across all co-expression modules and pathways tested for each of the two approaches. Gene sets were considered to be statistically significant if FDR < 0.05 in at least one SNP-gene mapping set or gene correlation set. To evaluate gene sets across the two approaches and between sexes, we followed up with a meta-analysis at the module/pathway level using the meta-MSEA function in Mergeomics to retrieve robust gene sets across both male and female cohorts as well as the three approaches. Stouffer's Z-score method was used to calculate meta-P-values based on the P-values from the multiple MSEA runs. Meta-FDR was calculated using the Benjamini-Hochberg method as described above.

Merging overlapping pathways into Supersets

The curated pathways and gene co-expression modules may carry redundant information. For example, the KEGG pathway "insulin signaling" can have largely overlapping genes with the Reactome pathway "insulin receptor signaling". To reduce redundancy, we compared the significant modules and pathways associated with liver fibrosis at FDR < 0.05 and merged the overlapping ones using a merging algorithm in Mergeomics to produce independent, non-overlapping "supersets". The algorithm employs an overlap ratio *r* between two gene sets A and B as $r = (r_{AB} \times r_{BA})^{0.5}$, where r_{AB} is the proportion of genes in A that are also present in B and r_{BA} is the proportion of genes in B which are also in A. The overlap ratio cut-off was set to r >= 0.33 and Fisher's exact test was used for assessing the statistical significance of gene overlap between modules/pathways. BH FDR < 0.05 was considered significant. Resultant supersets containing

more than 500 genes were trimmed down to contain core genes shared among the overlapping gene sets.

Liver Gene Regulatory Networks and Key Driver Analysis (KDA)

We performed a key driver analysis using the KDA algorithm in Mergeomics to identify potential KDs whose network neighbors are enriched for genes within the fibrosis-associated supersets uncovered by MSEA. The algorithm employed a chi-square-like statistic similar to that described for MSEA, and FDR < 0.05 was used to focus on the top robust KDs. Liver-specific Bayesian gene regulatory networks (BNs) were utilized. BNs used in our study were derived from mapping our liver fibrosis-associate supersets onto liver BNs, which were constructed based on human and mouse datasets from previous studies using the established method RIMBANet. A BN from a dataset represents a consensus network in which only edges that passed a probability of > 0.30 across 1000 BNs generated starting from different random seed genes were kept. BNs from individual studies [12, 117-119, 259-261] were combined without considering the edge weights (as the edges included in each BN were considered robust) to form a union network. This strategy has been successfully used previously to derive meaningful biological insights. Since the directions of the interactions might be conflicting in some of these previous studies, we omit the directionality in these BNs when applying KDA. Because these BNs were collected from both mouse and human studies, gene symbols in network figures are given in human orthologs.

Drug Repositioning

Using the sex-specific KDs found in the KDA, we utilized the PharmOmics tool to uncover potential therapeutic drugs [64]. We performed both network-based and overlap-based

repositioning for female-specific, male-specific, and shared KDs. For network-based repositioning, we queried meta drug signatures in mice/rats using a liver network. For overlap-based repositioning, we inputted the KDs found as upregulated genes since we omit directionality of BNs during the KDA.

5.3 Results

Identification of Co-Expression Modules and Pathways Genetically Associated with Liver Fibrosis

To identify the potential causal pathways for fibrosis in both males and females, we utilized genetic information in the form of GWAS, liver fibrosis eQTLs, fibrosis-correlated co-expression modules, and canonical pathways through integration (for males and females separately) to infer functionally connected gene groups that show a strong genetic association with liver fibrosis. Out of the 1827 curated canonical pathways, we identified 41 pathways enriched from male mice, and from the 62 fibrosis correlated co-expression modules, we found 47 enriched modules at an FDR < 0.05. For females, we identify 22 canonical pathways and 42 co-expression modules enriched at an FDR < 0.05 (**Figure 6.2**).

In terms of unique pathways, for males we find the top enriched canonical pathway to be "tight junction interactions" followed by a host of immune related signals in line with previous literature that males show a greater immune response in fibrosis than females. These pathways include "apoptosis", "cytokine-cytokine receptor interaction", and "pathogenic E. coli infection". The co-expression modules follow similar results with largely immune related signals, but also showcase strong enrichment for extracellular matrix related pathways as well, which was not captured within the canonical pathway analysis.

In females, the pathways enriched to be causal through our canonical pathway analysis are heavily carbohydrate metabolism related, with “Propanoate metabolism”, “Butanoate metabolism”, “gluconeogenesis”, and “glycerolipid metabolism” among others showing significance at an FDR < 0.05 (**Figure 6.2**). Some immune/inflammation causal pathways are present too such as “systemic lupus erythematosus”, “T cell receptor signaling” and “Fc Epsilon Ri signaling pathway”. The co-expression module analysis similar to the males was able to highlight many of the ECM related pathways such as “ECM receptor interaction”, “Core Matrisome”, as well as more lipid metabolism related pathways such as “cholesterol biosynthesis” and “metabolism of lipids”.

When comparing pathways shared between males and females, we found the causal pathways to be mostly unique, with only two overlapping canonical pathways “apoptotic cleavage of cellular proteins” and “neurotrophin signaling”, which has previously found by our research to be conserved between mice and humans (**Figure 6.2**) [256]. In our co-expression analysis, we were able to uncover more overlap between sexes, particularly the ECM related processes as well as other more novel including “CSK pathway”, important in cell growth and immune response. In addition, we uncovered more metabolic processes related to vitamins, lipids and lipoproteins to be shared between sexes.

Liver Fibrosis Transcriptome Correlations to Predict the Current Biological Pathways and Co-Expression Modules

While utilizing genetic associations advises on the potential causal pathways for liver fibrosis, utilizing liver transcriptome data correlated with liver fibrosis can help better indicate the current

biology within the liver. Using the same approach as above, we find a total of 102 canonical pathways and coexpression modules enriched for males and 48 enriched for females at an FDR < 0.05 (**Figure 6.2D**). Here, the transcriptome data showcases an increase in the number of overlapping pathways compared to those genetically associated, with a total of 32 overlapping between sexes compared to 21 via GWAS. However, the 32 overlapping pathways are dominated with ECM related processes, including "ECM organization", "Collagen formation", and "integrin cell surface interactions", which would reflect the current pathology of the liver.

For males, there are 70 unique pathways which include "metabolism of proteins", "bile acid metabolism", and translation related processes such as "viral mRNA translation" and "peptide chain elongation". Interestingly, a majority of these 70 unique pathways were found within our canonical analysis (51/70). The co-expression analysis captures many of the immune related processes including "IL7 Signaling", "TGF-Beta Pathway" and "IL4 Receptor in B Lymphocytes", which again overall mimics the sentiments of the causal pathways being uniquely more immune related for males.

Females, show a similar enrichment to the causal pathways, including "carbohydrate metabolism", "glycosaminoglycan metabolism", and "MPS diseases" amongst their 16 unique signals. MPS, or Mucopolysaccharidoses, are inherited lysosomal storage disorders that arise due to a lack of functional enzymes resulting in abnormal accumulation of glycosaminoglycans, which helps explain why we also find enrichment of glycosaminoglycan metabolism. The common link found utilizing the transcriptome as well as genetic data within females is a potential deficit in lysosomal enzyme function in females with resultant liver injury.

Between sexes, the data alludes to the same end result of liver fibrosis through major ECM reconstruction, however through sex specific mechanisms. Specifically, in males, both transcript

and genetic data showcase more protein related defects, a stronger immune response and perhaps greater cell death. Whereas, females showcase a strong metabolism defect particularly with glycosaminoglycan, butanoate and propanoate, with a less complex immune response noted. In fact, upon further investigation of the genes enriched for the overlapping pathways such as "Collagen Formation", we largely see non-overlapping genes between males and females despite the pathway being strongly enriched in each.

Merging of Genomics and Transcriptomics for combined pathways

To try to capture a holistic picture utilizing all genomics and transcriptomics datasets available, we can conduct a meta-MSEA, which calculates the cumulative significance of each pathway enrichment across all the above datasets to give an overall value, the Meta P-value. Unsurprisingly, the top enriched pathways for both males (**Table 6.1**) and females (**Table 6.2**) include ECM organization and ECM receptor interactions. Perhaps as a reflection of the current biology of the liver, the collective results are dominated by the transcriptome enrichments with minor causal (GWAS) contribution amongst the top pathways (**Table 6.1, 6.2**). The top 5 pathways for females are all ECM related, whereas that for males also include protein formation in addition.

Key Driver Analysis

Utilizing our liver Bayesian gene regulatory network constructed from tens of human and mouse datasets, we utilized our key driver analysis (KDA) to pinpoint key regulators/genes of liver fibrosis by overlaying genes enriched from all the significant canonical pathway and co-expression modules specific to GWAS and transcriptome results separately. We compared for overlap within

sexes for GWAS and transcript informed results as well as between sexes to understand how the different omics layers contribute to disease development and progression.

Male Networks

Firstly, for our male GWAS informed network, we found a large number (810) of KDs identified from our co-expression modules particularly those informing on immune signals and ECM related functions where the top KDs informed from these signals include *SYK, TLR2* and *EVL* (FC-gamma-R-mediated phagocytosis) and *COL6A3*, *COL1A1* and *ADAMTS2* (ECM glycoproteins) (**Figure 6.3A**). Importantly, we found many additional signals beyond immune and ECM, which involved lipid metabolism processes, including *ACSS2* which is important in acetyl-coA generation and has been functionally validated for its role in steatosis, as well as *FASN* another gene validated for its role in steatosis from our previous study [153] and perhaps more novel *HCC*. These are interesting candidates as it poses the potential for these genes to be important throughout NAFLD development.

Due to the dominant enrichment of key drivers informed from our co-expression analyses, we re-ran our KDA analysis exclusively for canonical pathways to ensure that no other biology was missed or overshadowed. Here, we generally found complementary results with our combined network (co-expression and canonical informed) as expected, highlighting more immune processes such as hemostasis, cytokine-cytokine receptor interaction and chemokine signaling pathways with genes highlighted including *CCL4, CCL7, INPP5D, PTPRC* and *APBB1P* among others.

Besides the causal networks informed by GWAS, it is appropriate to understand which genes are driving current disease progression processes. We therefore overlaid the transcript

informed mechanisms on our liver network. As would have been expected, the network was heavily enriched for ECM related processes of which *COL6A3, ADAMTS2* and *COL1A1* were again top candidates reflecting the current fibrosis liver state. Many of the overall network genes reflected the GWAS analysis and to understand the overlap and differences better, we carried out an overlap network and a unique network. The GWAS only network highlighted the uniquely causal hubs including again *ACSS2* and many inflammatory signals such as *IFIT1, ISG15, CXCL10* and interestingly circadian rhythm related genes such as *ARNTL1.*

Female Networks

Following suit from the male analysis to inform on the causal networks in female liver fibrosis, we examined the results suggested through GWAS. Generally, the overall theme is similar to male GWAS with many of the top KDs recapitulated albeit not all from the same co-expression modules or canonical pathways (although similar overall functions) (**Figure 6.3B)**. An example of this is *SYK,* which again is one of the top ranked KDs but is now listed as part of leishmania infection (derived from MEGENA) rather than FC-gamma-R-mediated phagocytosis (derived from WGCNA). On top of this the ranked order of key driver significance (FDR<0.05) is somewhat different perhaps suggesting different levels of importance of genes fundamental to fibrosis development such as *ACSS2* being top 10 in female but top 30 in males. Moreover, the female related processes have slightly less emphasis on the immune pathways, with many of the top networks covering metabolic processes from lipid and lipoprotein metabolism as well as cholesterol synthesis including genes *FDFT1* and *GPAM.* Again, separating out the co-expression module analysis from our canonical pathways, the canonical analysis only highlighted six processes including arginine and proline metabolism (*GOT1, ASS1 and CPS1*), Butanoate

metabolism (ACSM3), Propanoate metabolism (*ACADM*), systemic lupus erythematosus (*HIST1H4F, HIST1H2BK and HIST1H2BO*), Fc epsilon RI signaling pathway (*NCKAP1L*), and T cell receptor signaling (*CD3G*). The canonical network highlights this underlying metabolic issue, but again coupled with immune irregularities.

Examining the current biology of female liver fibrosis, we again looked at the transcript network, which highlighted the ECM related processes and those major genes *COL6A3, ADAMTS2,* and *THBS2*. The Transcript analysis highlighted that the genes within male and female liver working in fibrosis processes are similar as a whole. Importantly, we screened for differences with our GWAS network, which highlighted that GWAS was specific to a number of immune signals and metabolism signals with more emphasis on the key drivers derived from canonical pathways with all bar *NCKAP1L* being specific to GWAS.

Sex Comparison between networks

Comparing first the overlap between the GWAS networks, we find that for females 88% of the key drivers predicted overlap with males whereas for males there is a 64% overlap with females, showcasing the additional complexity in male liver fibrosis (**Figure 6.4A**, **6.4B**). The top overlapping KDs include the ECM related collagen genes (*COL6A3, PCOLCE, ADAMTS2, COL1A1*), immune related (*SYK*) and metabolism processes (*ACSS2, FASN, GPAM, FDFT1*). We also find *PNPLA3* as a KD in both sexes, which is interesting due to high prevalence in NAFLD as a major GWAS hit. The top ranked male specific KDs include *CCNA2, C15orf23, RACGAP1* and *PBK*, whereas for females *UGT2A3, HIST1H2BO* and *MAST2* are noted. Beyond these top ranked (FDR<0.05), we find *CHCHD6* to be specific to females, which is interesting due to its validation in our previous study in steatosis and *FADS3* a known GWAS hit in NAFLD to be

female specific. *PRODH*, which has recently been implicated in NAFLD development was noted to be male specific.

For the transcriptome comparison, we find that 64% of the female key drivers overlap with males whereas males have an 85% overlap with females in this case (**Figure 6.4C, 6.4D)**. This is almost a reverse of the key driver analysis for GWAS, this is possibly explained by females having a more complex pathogenesis in fibrosis development, where more cross talk is apparent from different mechanisms and genes within those mechanisms. We find again that the top shared overlapping key drivers include *COL6A3*, *ADAMTS2*, *COL1A1* and *THBS2.* The top KDs unique to males includes *CPS1, ASL, HGD* and *ASS1*, and the top unique to female includes *FERMT3, IGSF6*, *CTSS* and *NCF2*.

Drug Repositioning through Pharmomics

Next, we utilized the drug repositioning tool Pharmomics [64] to predict potential drugs that can be used to alleviate symptoms of NAFLD. Performing drug repositioning using male-specific KDs, the top drugs found include corticosteroids like Fludrocortisone and Dexamethasone; anti-hypertensives like Ramipril, Pentoxifylline, Candesartan, and Losartan; NSAIDs like Naproxen, Sulindac, and Fenoprofen; and diuretics like Eplerenone and Furosemide. Using female-specific KDs, the top drugs we identified include corticosteroids like Fludrocortisone, Betamethasone, Prednisolone, and Dexamethasone; antivirals like Penciclovir and Ritonavir; and immunomodulators like Glatiramer among other drugs. Searching through the genes overlapped between sexes, the top drugs we found include corticosteroids like Fluocinolone and Dexamethasone, and Enzalutamide which is a nonsteroidal antiandrogen drug. Through drug

repositioning and subsequent experimental validation, we hope to be able to find potential therapeutic treatments for NAFLD.

5.4 Discussion

In this study, we utilized a multi-omics approach utilizing our Mergeomics pipeline in an attempt to distinguish key differences in fibrosis pathogenesis between sexes. To do this, we leveraged hepatic fibrosis GWAS and fibrosis eQTL data and integrated these datasets with knowledge derived canonical pathways (KEGG, Reactome, and Biocarta) and data driven sex specific liver fibrosis co-expression modules created using WGCNA and MEGENA. We also utilized liver fibrosis transcriptome data to pinpoint the biology relevant to liver fibrosis during disease as opposed to inherited and predicted causal pathways/genes from GWAS. A combination of multi-omics allows us to gain a more holistic picture of disease progression starting from baseline inherited genetics to the mechanisms and genes contributing to fibrosis in its current state. Taking all of these resources together we finally mapped all of the significant co-expression modules and canonical pathway genes to liver BN's in our KD analysis, where we were able to pinpoint the central hubs/key driver genes which are likely to play a significant role in disease generation, both shared and sex specific.

Through our pathway analysis, we found that males have a greater number of immune response pathways, which is consistent with prior knowledge suggesting that liver fibrosis in males generates a more intense immune response. Additionally, we found expected lipid metabolism pathways as well as previously unobserved protein/translation-related pathways among our male-specific pathways, which suggests that differential expression of protein biosynthesis and degradation pathways could potentially play a role in fibrosis pathogenesis. In females, we

identified more carbohydrate metabolism-related pathways including a large array of mucopolysaccharide diseases, which are inherited lysosomal storage disorders where lysosomal enzymes responsible for breaking down glycosaminoglycans are impaired or absent. This alludes to a potential pathway for fibrosis progression, as buildup of glycosaminoglycans can cause cell injury, leading to fibrosis. Previous studies [262-264] have also alluded to this notion, finding NAFLD patients to have a higher heparan sulfate and chondroitin sulfate concentration. Through our GWAS analysis we observed two shared pathways, "Apoptotic Cleavage of Cellular Proteins" and "Neurotrophin Signaling Pathway", which has been noted in our previous study [256]. While "Neurotrophin Signaling Pathway" may seem like an unorthodox pathway for liver fibrosis, studies have shown neurotrophin regulator *p75NTR* to be a liver regeneration regulator increasingly expressed in the hepatic stellate cells (HSCs) of cirrhotic liver. Comparing our transcriptome data with our liver fibrosis GWAS, we identified many shared pathways that were ECM-related, including "ECM Organization", "Collagen Formation", and "Focal Adhesion", in reference to the pathophysiology of liver fibrosis.

Through the Key Driver Analysis function of the Mergeomics pipeline, we identified potential key regulators of NAFLD progression between sexes. In males, we identified many cell cycle and mitosis related genes, suggesting that these genes, specifically in regard to fibroblast proliferation, more heavily impact male liver fibrosis pathogenesis. We also found male-specific protein/translation-related genes consistent with pathways from our male-specific pathway analysis. In females, we found a variety of genes including metabolism related genes as well as complement system genes, highlighting the complement system's role in promoting inflammation leading to fibrosis. Shared between males and females, we find an abundance of ECM-related

genes, consistent with liver fibrosis pathophysiology as aforementioned. These shared genes, however, are annotated with different pathways between sexes, suggesting that these genes both contribute to liver fibrosis, though through different pathways in NAFLD progression between sexes. Additionally, we also identified *NCKAP1L* and *FASN*, which have both been previously found by our group where we have experimentally validated *FASN* with knockdown high fat, high sucrose-induced NAFLD mice, finding them to have improved steatosis and insulin resistance.

Looking into the known biology of each key driver, we chose to explore several less-studied genes in regard to NAFLD found in our multi-omics analysis using high fat, high sucrose-induced NAFLD mice. We are currently pursuing validation of *VNN1* and *MAOB*, both shared causal genes between sexes. *VNN1* is a GPI-anchored molecule that plays a role in hematopoietic cell trafficking and the oxidative stress response. In a prior study, lipid-induced toxicity in mice was shown to induce hepatocytes to secrete microparticles on which the protein coded by *VNN1* is the most abundant protein; these microparticles promote angiogenesis, which plays a major role in NAFLD progression [265]. *MAOB* encodes a monoamine oxidase where elevated levels serves as a biomarker for liver fibrosis [266] and has been shown to be tied with reactive oxygen species production leading to mitochondrial dysfunction [267].

Using the PharmOmics pipeline [64], we performed drug repositioning to uncover potential therapies for NAFLD using the sex-specific and overlap KDs found in our KDA. One important thing to note is that our pipeline does not show directionality, meaning that the drugs found could have beneficial effects or adverse effects, or both. In all three drug repositioning forms used, we find a lot of corticosteroids like Fluocinolone and Dexamethasone, which are expected due to inflammation being a key player in NAFLD progression. We also find many NSAIDs in both sex-

specific repositioning for the same reason as well. Interestingly, we find Enzalutamide as an overlap drug. Enzalutamide is a nonsteroidal antiandrogen, which may allude to the adverse effects of testosterone in both sexes as well as the protective effects of estrogen seen in females. As we will with the key driver genes, we will be pursuing the validation of less-studied drugs for a potential treatment of liver fibrosis.

In this study, we utilized an integrative and systems biology approach through genetic and transcriptomic data in an attempt to holistically differentiate liver fibrosis pathogenesis between males and females. Our study overall highlights a greater immune response in males along with more protein and lipid metabolism abnormalities; for females, we found more carbohydrate metabolism related abnormalities contributing to liver fibrosis. Through our KD analysis, novel key drivers were found. More research into these genes can help identify plausible targets and create sex-specific therapeutic treatments for NASH.

5.5 Tables

Table 5.1 Top consistent pathways derived from GWAS and TWAS within males.

MODULE	P GWAS	FDR GWAS	P Transcript	FDR Transcript	METAP	FDR	DESCR
WGCNA_3	1.37E-05	0.00310957	1.34E-47	4.55E-45	5.71E-40	5.03E-37	ECM glycoproteins, Core matrisome, ECM receptor interaction
WGCNA_4	0.0002158 7	0.00756445	1.33E-27	2.25E-25	1.79E-24	7.87E-22	Core matrisome, ECM glycoproteins, Matrisome associated
rctm0388	0.0053118 9	0.0496367	3.90E-24	3.78E-22	2.24E-19	3.29E-17	Extracellular matrix organization
rctm0385	0.0230956 6	0.1010456	1.97E-21	1.34E-19	3.22E-16	4.05E-14	Eukaryotic Translation Termination
rctm0790	0.0487548 5	0.13866756	4.19E-19	1.89E-17	5.30E-14	5.19E-12	Nonsense-Mediated Decay
MEGENA_7	0.1421078 9	0.24013052	2.89E-21	1.78E-19	6.82E-14	5.72E-12	NCAM1 interactions, Core matrisome, Integrin1 pathway
rctm0576	0.2092814 1	0.29983929	2.96E-22	2.23E-20	7.80E-14	5.72E-12	Influenza Viral RNA Transcription and Replication
rctm0788	0.0375928	0.12678989	2.25E-18	9.00E-17	7.60E-14	5.72E-12	Nonsense Mediated Decay Enhanced by the Exon Junction Complex
M189	0.1321793 4	0.23255821	1.31E-20	6.83E-19	1.25E-13	8.50E-12	Ribosome
rctm0789	0.0443685 7	0.13131997	3.09E-18	1.17E-16	1.38E-13	8.71E-12	Nonsense Mediated Decay Independent of the Exon Junction Complex

Table 5.2 Top consistent pathways derived from GWAS and TWAS within females

MODULE	P GWAS	FDR GWAS	P Transcript	FDR Transcript	METAP	FDR	DESCR
WGCNA_3	4.46E-14	1.16E-11	1.27E-24	3.48E-22	5.59E-36	4.37E-33	ECM glycoproteins, Core matrisome, ECM receptor interaction
MEGENA_26	0.00407012	0.04877883	2.60E-21	2.39E-19	7.88E-18	3.08E-15	Core matrisome, ECM glycoproteins, ECM organization
rctm0388	0.006496	0.07063284	3.54E-09	8.85E-08	2.46E-09	1.75E-07	Extracellular matrix organization
MEGENA_550	0.00063689	0.01271074	2.24E-05	0.0003238	1.21E-07	6.79E-06	Axon guidance, Basement membranes, Developmental biology
MEGENA_586	0.00017071	0.00458565	0.00012143	0.00151793	1.47E-07	7.66E-06	Integrin3 pathway, Core matrisome, Integrin1 pathway
MEGENA_870	2.34E-05	0.00126442	0.00438287	0.02869734	1.11E-06	5.12E-05	CSK pathway; Calcium dependent events; Alanine, aspartate, and glutamate metabolism
MEGENA_186	0.00447908	0.0522157	6.34E-05	0.00087167	2.58E-06	9.61E-05	Melanoma, Glioma, Axon guidance
M4086	2.76E-06	0.00031125	--	--	2.76E-06	9.81E-05	Propanoate metabolism
M3397	9.07E-06	0.00088271	--	--	9.07E-06	0.00026255	Butanoate metabolism
MEGENA_618	2.33E-05	0.0012643	--	--	2.33E-05	0.0006069	Biosynthesis of unsaturated fatty acids, Alpha-linolenic acid (ALA) metabolism

Table 5.3 Top consistent drugs derived from GWAS and TWAS within males

Drug (GWAS)	Class of Drug	Network-based P-Value	Overlap-based P-Value	Associated KDs
Fludrocortisone	Corticosteroid	1.32E-05	2.33E-43	ACAT2, ACSS2, ANXA2, ARNTL, VIM, CASP1, CCL2, FVER1G, FASN, EVL
Prednisolone	Corticosteroid	2.34E-04	1.03E-42	ADAM8, AIF1, BTG2, COL1A1, ELOVL5, MVD, NCAM1, PTPRC, VIM
Urokinase	Urokinase-type plasminogen activator	2.11E-03	3.03E-41	ACACB, AIF1, ANXA3, AXL, DDR1, CYBA, ELOVL5, FASN, NCF2
Ritonavir	HIV Protease Inhibitor	1.22E-03	2.10E-52	BEX2, ANXA3, COL4A1, CTPS, FOLR2, PDGFRB, PLEK, MVD, SMOC2, THRSP
Oxaliplatin	Alkylating Antineoplastic Agent	4.19E-03	9.49E-35	ACCS2, ADAMTS2, BTG2, CYB5R, COL1A2, FCGR1, GSN, MGP, TLR1
Sulindac	NSAID	8.17E-03	2.45E-40	ACAT2, ACSS2, ATF3, COL3A1, FCER1G, ELOVL6PKLR, VTN, VIM
Doxorubicin	Topoisomerase Inhibitor Antineoplastic Agent	3.71E-02	1.52E-56	ADAMTS2, ANXA2, ARNTL, BGN, COL1A1, EVL , FASN, FERMT3
Dexamethasone	Corticosteroid	2.01E-02	1.42E-70	ANXA1, CASP1, CCDC3, CCL3, COL6A3, LPL, GPAM, NCKAP1L
Prednisone	Corticosteroid	3.14E-02	4.26E-22	CCNA2, FCGR3, HK3, LOXL1, NCAM1, PNPLA5, RAC2, SMOC2, TYROBP
Decitabine	Antimetabolite Antineoplastic Agent	3.60E-02	1.65E-38	ACACB, ACSS2, ADAMTS2, AXL, COL6A3, DDR1, GPAM, IDI1, MVD

Table 5.4 Top consistent drugs derived from GWAS and TWAS within females

Drug (GWAS)	Class of Drug	Network-based P-Value	Overlap-based P-Value	Associated KDs
Fludrocortisone	Corticosteroid	3.36E-02	2.44E-31	AACS, ACSS2, CCL2, COL3A1, DDR1, CTSS, HCK, MVD, MOGAT1, PGD
Prednisolone	Corticosteroid	4.45E-03	2.83E-28	AIF1, ANXA2, BGN, CCDC3, DCN, DECR1, FERMT3, PGD, VIM
Ritonavir	HIV Protease Inhibitor	6.08E-03	4.57E-45	AIF1, AXL, FERMT3, GPAM, MOGAT1, PCOLCE, PGD, SAMSN1
Chloroquine	Quinolone	8.06E-03	6.26E-21	AACS, ACSS2, ANXA2, COL3A1, FCER1G, FOLR2, LSS, PCOLCE
Neomycin	Aminoglycoside Antibiotic	9.56E-03	1.35E-34	ACADM, AACS, CCL2, CYBA, EVL, NCAM1, NCF4, NCKAP1L, PKLR
Doxorubicin	Topoisomerase Inhibitor Antineoplastic Agent	9.87E-03	5.21E-47	ACACA, ADAMTS2, BGN, EVL, FASN, ELOVL5, LSS, JUN, PAM, PTPRC
Sorafenib	Tyrosine Kinase Inhibitor Antineoplastic Agent	9.98E-03	2.77E-49	ACSS2, ART4, AIF1, CLIP2, COL1A1, CIDEC, HCK, GPAM, VIM, THBS2
Pyrazinamide	Antitubercular Agent	1.19E-02	7.81E-25	ACAT2, ANXA2, CCL2, CIDEC, EVL, FASN, JUN, PLEK, SCARA3, TYROBP
Betamethasone	Corticosteroid	1.79E-02	1.22E-27	ACLY, ACSS2, CASP1, EVL, FCER1G, FOLR2, MOGAT1, PTPRC, SPARC
Glatiramer	Immunomodulator	3.07E-02	9.73E-37	ACAT2, CCL2, CLIp2, CFP, COL1A2, GPAM, ITGAL, PGD, PLK3, SCARA3

5.6 Figures

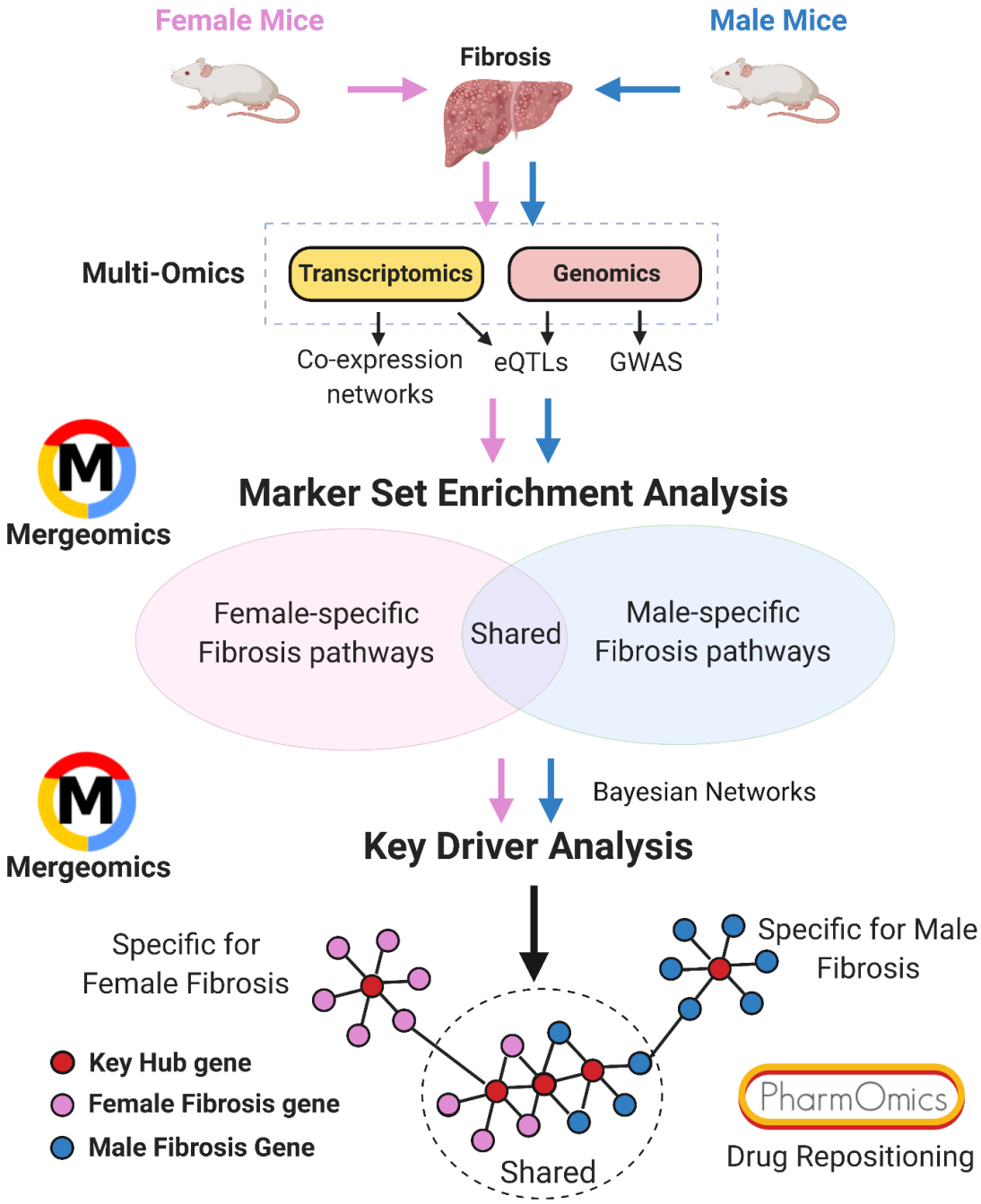

Figure 6.1. Study Overview.

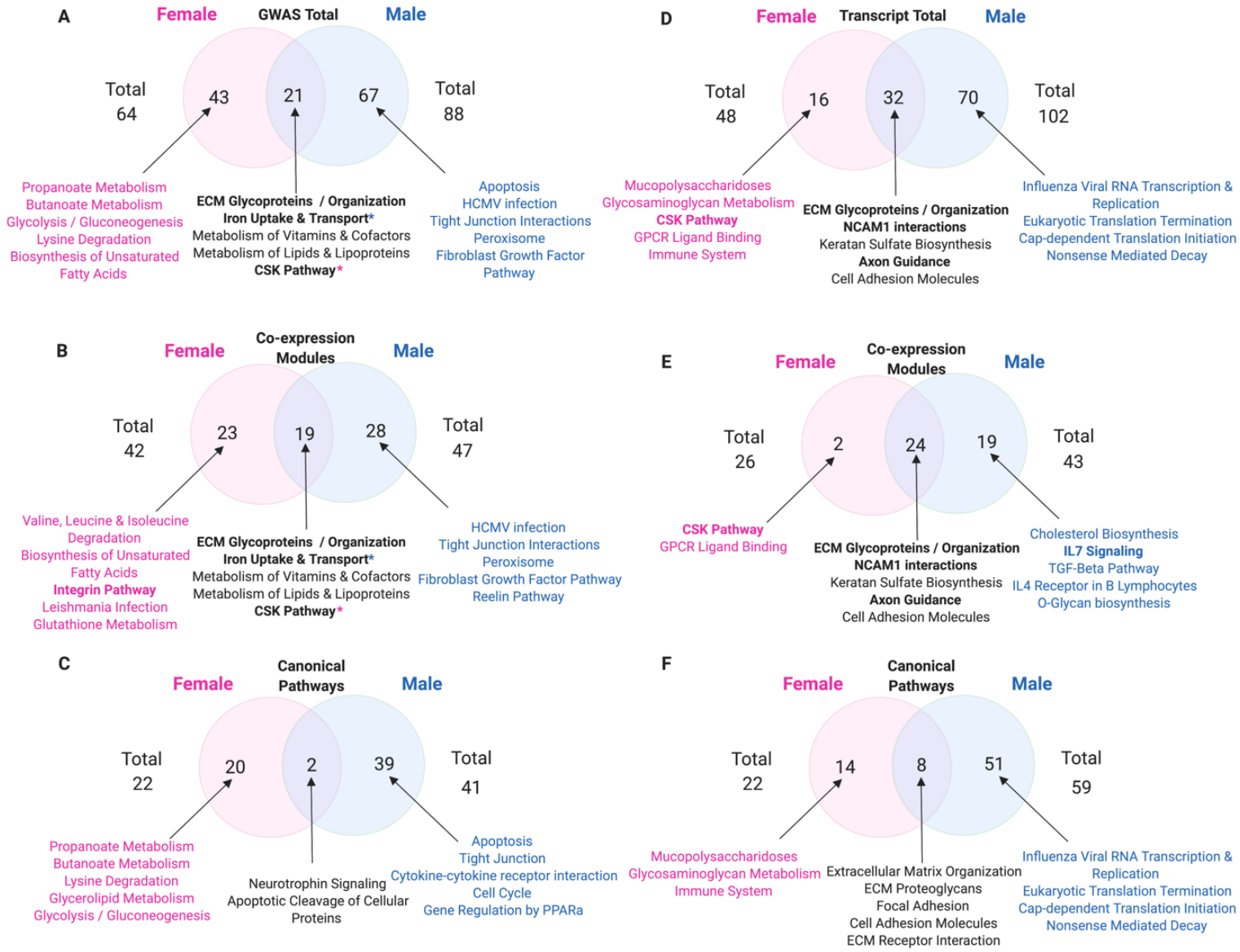

Figure 5.2 Overlap results of MSEA results between male and female mice for GWAS and TWAS

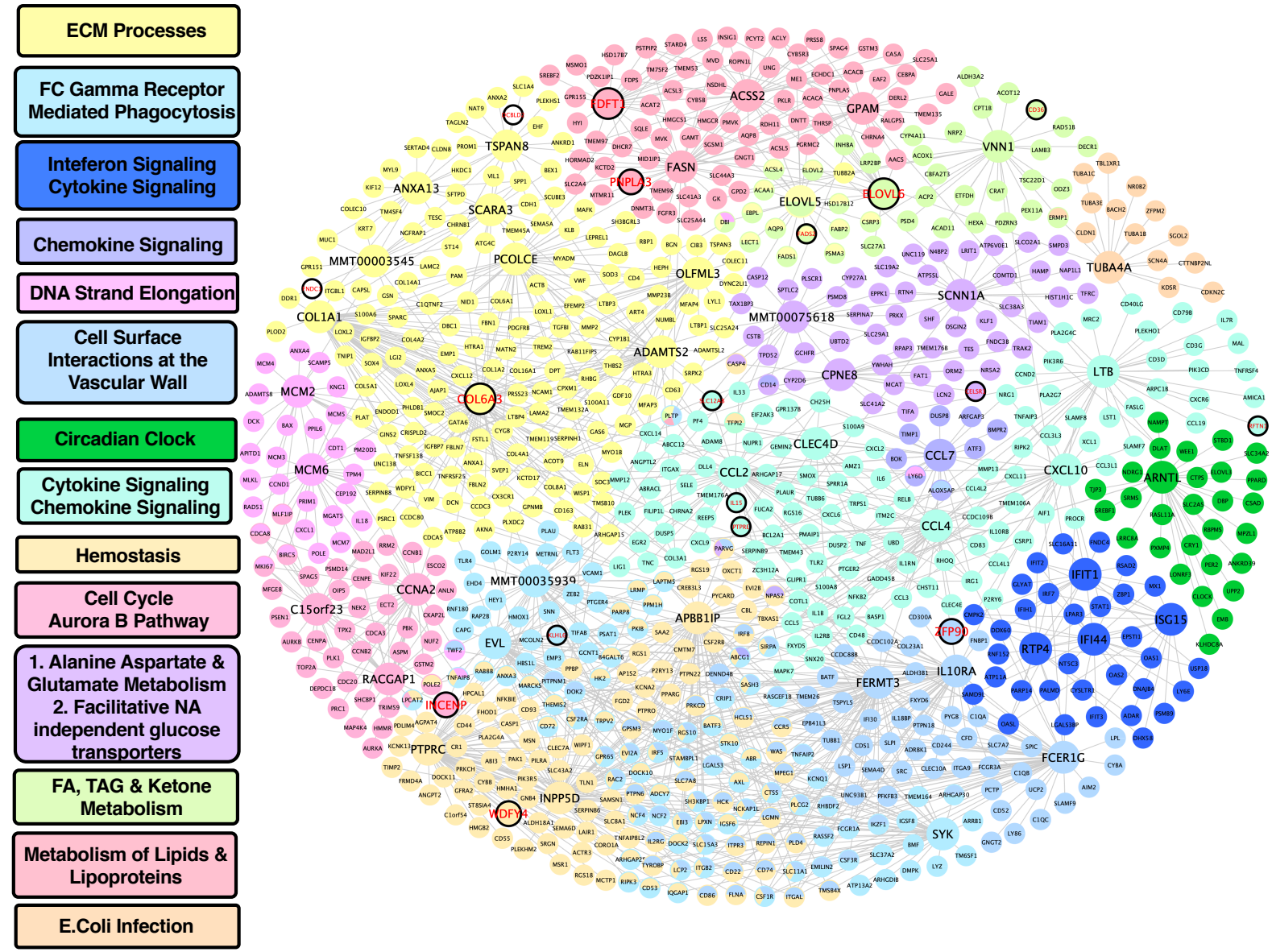

Figure 5.3 Male GWAS liver Bayesian network

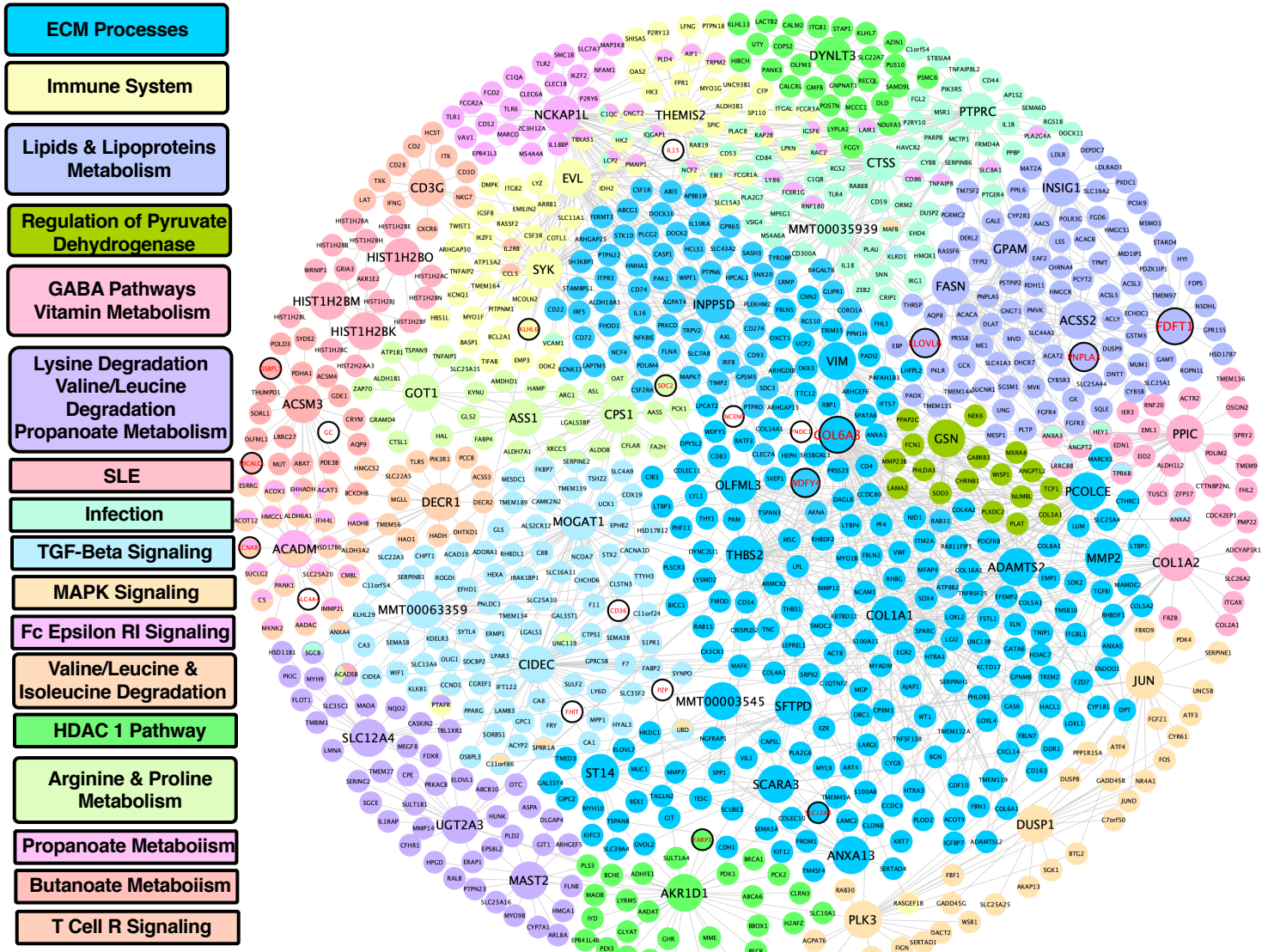

Figure 5.4 Female GWAS liver Bayesian network

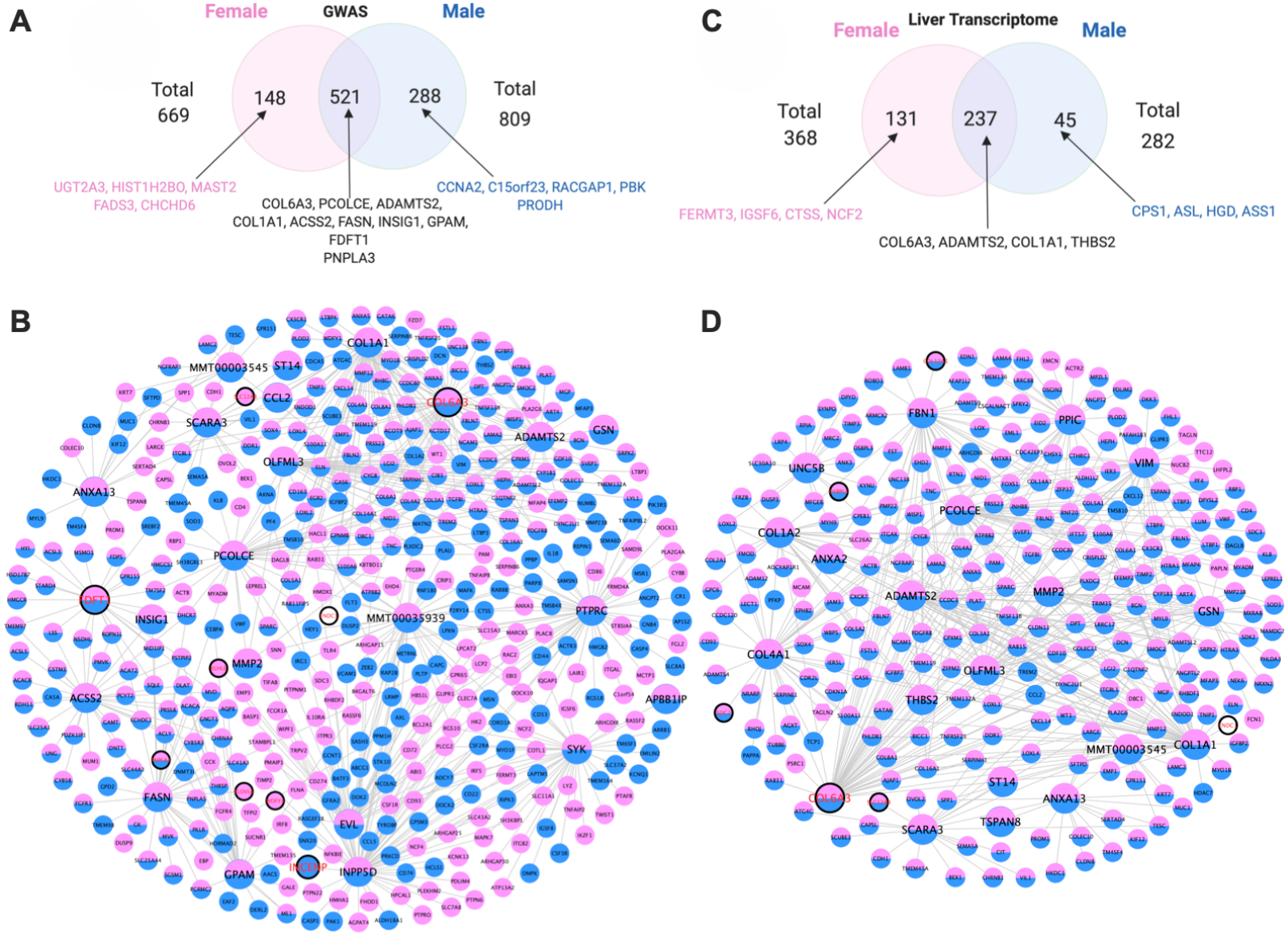

Figure 5.5 Overlap analysis of KD genes between males and females and visualization of network structure

Chapter 6. Relative contributions of sex hormones, sex chromosomes, and gonads to sex differences in tissue gene regulation

6.1 Introduction

Females and males differ in the risk, incidence, and progression of complex diseases such as obesity, non-alcoholic fatty liver disease, and diabetes [268-270]. Thus, one sex may have endogenous protective or risk factors that could become targets for therapeutic interventions. Current sexual differentiation theory suggests that three major classes of factors cause sex differences [271-275]. First, some sex differences are caused by different circulating levels of ovarian and testicular hormones, known as "activational effects". These differences are reversible because they are eliminated by gonadectomy of adults. Second, certain sex differences persist after gonadectomy in adulthood and represent the effects of permanent or differentiating effects of gonadal hormones, known as "organizational effects," that form during development. A third class of sex differences are caused by the inequality of action of genes on the X and Y Chromosomes in male (XY) and female (XX) cells, and are called "sex chromosome effects".

To date, few studies have systematically evaluated the relative importance of these three classes of factors acting on phenotypic or gene regulation systems [276]. The activational effects of hormones have been established as a significant contributor to sexual dimorphism in metabolic diseases, with additional evidence pointing to sex chromosome effects on obesity and lipid metabolism [277-279]. Previous studies have also emphasized the importance of organizational or activational hormone effects on liver gene expression [280-284]. However, the tissue-specific contributions and the interactions of activational, organizational, and sex chromosome effects on gene regulation are poorly investigated.

Here we conduct a systematic investigation to understand the relative contribution of the three sex-biasing factors in gene regulation (**Figure 6.1)**. We used the Four Core Genotypes (FCG) mouse model, in which the type of gonad (ovary or testis) is independent of sex chromosome complement (XX or XY) [285, 286]. The model separates the effects of sex chromosome complement by fixing the gonadal status (XX vs. XY with ovaries; XX vs. XY with testes) from the effects of gonads by fixing the sex chromosome type (ovaries vs. testes with XX genotype; ovaries vs testes with XY genotype). By varying adult gonadal hormone levels via gonadectomy and subsequent hormonal treatments we also asked how androgens and estrogens influence gene expression as a function of sex chromosome complement and gonadal sex. The design allows comparison of the magnitude of effect of each sex-biasing factor and the interactions among different factors.

Using the FCG model, our aim is to assess the role of the three sex-biasing factors and their interactions on gene expression, molecular pathways, and gene network organization in the liver and adipose tissue, which are central tissues for metabolic and endocrine homeostasis, with adipose tissue additionally contributing to immune functions. We further aim to understand the relationship of each sex-biasing factor with various human diseases.

6.2 Results

Overall study design

In FCG mice, the Y Chromosome (from strain 129) has sustained a spontaneous deletion of *Sry,* and an *Sry* transgene is inserted onto Chromosome 3 [286] (**Figure 6.1A;1B**). Here, "male" (M) refers to a mouse with testes, and "female" (F) refers to a mouse with ovaries. FCG mice include XX males (XXM) and females (XXF), and XY males (XYM) and females (XYF; **Figure 6.1C**). A total of 60 FCG mice were gonadectomized (GDX) at 75 days of age and implanted

immediately with medical grade Silastic capsules containing Silastic adhesive only (blank control; B) or testosterone (T) or estradiol (E) (**Figure 6.1D**). This study design produced 12 groups, with 4 groups of FCG mice (XXM, XXF, XYM, XYF) and each group subdivided into B, T or E based on hormonal treatment: XXM_B, XXM_T, XXM_E, XYM_B, XYM_T, XYM_E, XXF_B, XXF_T, XXF_E, XYF_B, XYF_T, XYF_E (n=5/genotype/treatment). Liver and inguinal adipose tissues were collected 3 weeks later for transcriptome analysis (**Figure 6.1E**). All liver samples passed quality control (n=5/group) whereas 5 adipose samples across 4 of the 12 groups failed quality control (n=3-5/group; **see Methods**) (**Figure 6.1F**). The design allowed detection of differences caused by three factors contributing to sex differences in traits (**Figure 6.1G**). (1) "Sex chromosome effects" were evaluated by comparing XX and XY groups (n=~30/sex chromosome type/tissue). (2) "Gonadal sex effects" were determined by comparing mice born with ovaries vs. testes (n=~30/gonad type/tissue). Since mice were analyzed as adults after removal of gonads, the gonadal sex effects represent organizational (long-lasting) effects of gonadal hormones, such as those occurring prenatally, postnatally, or during puberty. This group also includes effects of the *Sry* gene, which is present in all mice with testes and absent in those with ovaries. Any direct effects of *Sry* on non-gonadal target tissues would be grouped with effects of gonadal sex. (3) "Hormone treatment effects" refers to the effects of circulating gonadal hormones (activational effects) and were evaluated by comparing E vs. B groups for estradiol effects, and T vs. B groups for testosterone effects, with n=~20/hormone type/tissue.

Global effects of sex chromosome complement, gonadal sex, and hormonal treatments on liver and adipose tissue gene expression

To visualize the overall gene expression trends due to effects of the three primary sex-biasing components, we conducted principal component analysis (PCA; **Figure 6.2**). For adipose

tissue, hormonal treatment (**Figure 6.2A**), sex chromosomes (**Figure 6.2B**) and gonadal sex (**Figure 6.2C**) did not clearly separate the groups. However, in the liver there was a separation of groups based on gonadal hormones, particularly in response to testosterone treatment (**Figure 6.2D**), but not based on chromosomal or gonadal factors (**Figure 6.2E, 6.2F**).

We then asked which individual genes in liver and adipose tissues were affected by adult hormone level, gonadal sex, and sex chromosome complement, as well as interactions between these factors, using three sets of ANOVA tests to address biological questions at different resolution. We defined a differentially expressed gene (DEG) as a gene that passed a false discovery rate (FDR)<0.05 for individual sex-biasing factors and the interaction terms from the ANOVAs. First, we used a 3-Way ANOVA (3WA) to test the main effects of sex hormones, gonad type, and sex chromosome as well as the interaction terms. Tens to thousands of DEGs were identified in liver (**Table 6.1**) and adipose tissue **(Table 6.2)**. In both tissues, hormonal treatments affected the largest numbers of genes, followed by fewer genes that were responsive to gonadal/organizational effects or sex chromosome complement (**Figure 6.3**). Testosterone treatment in the liver induced the largest number of DEGs (**Figure 6.3A**), whereas in adipose tissue estradiol treatment affected the greatest number of DEGs (**Figure 6.3D**). These trends remained when different statistical cutoffs (unadjusted $p<0.05$, $p<0.01$, FDR<0.1, FDR<0.05) were used (**Figure 6.4**). These results support tissue-specific sensitivity to different hormones.

Next, we asked if the sex chromosome and gonadal effects are more evident in specific hormonal treatment groups using a 2-way ANOVA (2WA). In the liver, the organizational effects of gonad type were strongest in gonadectomized mice without hormone replacement (blank group) (**Figure 6.3B**). By contrast, in adipose tissue the gonadal sex effect was most prominent in the estradiol treated groups (**Figure 6.3E**), suggesting that estradiol levels augment the enduring

differential effects of gonads on the adipose transcriptome. Sex chromosome effects were limited regardless of hormonal treatment status.

Lastly, we examined whether the effects of testosterone and estradiol are dependent on genotypes using a 1-way ANOVA (1WA) followed by post-hoc analysis. More liver genes were affected by testosterone than by estradiol regardless of genotype, although XYM liver appeared to be less responsive to testosterone than liver from other genotypes (**Figure 6.3C**). By contrast, in adipose tissue, estradiol affected more DEGs in XX genotypes (XXM and XXF) than in XY genotypes (XYM and XYF), whereas testosterone had minimal impact on adipose tissue gene expression in all four genotypes (**Figure 6.3F**). These results further support tissue-specific effects of estradiol in adipose tissue and testosterone in liver, and indicate that activational effects of hormones also depend on sex chromosome complement and hormonal history (gonadal sex) of the animal.

Genes and pathways affected by hormonal treatment

In the liver, the 3WA analysis showed that testosterone treatment induced the greatest number of DEGs with 1378 compared to 333 DEGs from estradiol treatment (**Table 6.1; Figure 6.3A)**. The testosterone DEGs were enriched for metabolic pathways (lipid metabolism, organic acid metabolism, bile acid biosynthesis), development, and immune response (**Table 6.1**). The estradiol liver DEGs showed enrichment for metabolic (organic acid metabolism, carboxylic acid metabolism) and immune pathways (complement and coagulation).

In contrast to liver, we found that the effect of estradiol treatment was more profound (2029 DEGs) than that of testosterone (275 DEGs) in 3WA of the inguinal adipose tissue (**Table 6.2; Figure 6.3D**). The estradiol DEGs were enriched for protein metabolism, focal adhesion, and

transport pathways. Testosterone DEGs were enriched for cell-cell adhesion, development, regulation of transcription, and protein signaling pathways.

Overall, both estradiol and testosterone affected genes involved in metabolism, development, and immune function. However, estradiol primarily affected these processes in the adipose tissue, whereas testosterone exhibited influence in the liver.

Genes and pathways affected by gonadal sex

In the liver, 3WA analyses revealed 93 DEGs influenced by gonadal sex when testosterone and blank treatment groups were considered, and 209 DEGs in the analysis of estradiol and blank groups (**Table 6.1**). These genes were enriched for immune/defense response and lipid metabolism pathways. By 2WA, we found that gonadal sex has the strongest influence on inflammatory and metabolism genes in the absence of hormones (blank group; 115 DEGs), but the effect was reduced by estradiol treatment (53 DEGs) and minimized by testosterone treatment (9 DEGs; **Table 6.1**; **Figure 6.3B**).

For the inguinal adipose tissue, gonadal sex had more than twice as many DEGs as in liver tissue in the 3WA analysis (**Figure 6.3A** vs. **6.3D**). Further dissection of the gonadal sex effect in individual hormonal treatment groups in a 2WA analysis showed that the effects of gonadal sex were strongest in the estradiol group (400 DEGs), followed by testosterone group (161 DEGs), and lastly by the blank group (70 DEGs; **Figure 6.3E**). Genes affected by gonadal sex are mainly relevant to developmental processes, with arginine and proline metabolism genes also affected in the estradiol group and cancer-related genes in the testosterone group.

These results support the importance of gonadal sex in regulating development, metabolic, and immune processes in both tissues. However, in the liver, hormonal treatments minimized the

effects of gonadal regulation of gene expression, whereas in the adipose tissue, hormones amplified the gonadal influence on gene expression. In both tissues, the gonadal sex effect was more prominent in the estradiol-treated group than in the testosterone-treated group (**Figure 6.3B** vs. **6.3E**).

Genes and pathways affected by sex chromosome complement

In both the 3WA and 2WA analyses, ten or fewer genes were found to be significantly affected by sex chromosome complement at FDR < 0.05 in the liver (**Table 6.1; Figure 6.3C**) and 10-22 DEGs were influenced by sex chromosomes in the adipose tissue (**Table 6.2; Figure 6.3F**). These genes were mainly sex chromosome genes known to exhibit sex differences, including *Xist*, *Ddx3y*, *Kdm6a*, *Hccs, Cited1, Tlr7* and *Eif2s3x/y* [278, 287-289]. However, autosomal genes were also influenced by sex chromosome type in both liver (e.g., *Ntrk2* and *H2-DMb1)* and adipose tissue (e.g., *Pals1*, *Esrp1*, and *Dnai1)*. Genes influenced by sex chromosome complement are involved in inflammation/immune response (*Tlr7*, *H2-Dmb1, Cited1*), GPCR signaling (*Esrp1*), metabolism (*Hccs*), and cell junction organization (*Pals1*).

Genes and pathways affected by interactions of sex-biasing factors

The interactions among the sex-biasing factors are supported by numerous DEGs with significant effects from the interaction terms in the ANOVA analyses (FDR<0.05; **Supplemental Table S6.1; S6.2**). For instance, in adipose tissue 31 DEGs were affected by interactions between estradiol and gonad type. These DEGs were enriched in pathways such as VLDL particle assembly and regulation of leukocyte chemotaxis. DEGs *Dnai1* and *Cited1* were expressed in female gonads (XXF or XYF) when no sex hormones were provided; genes such as *Ctns, Slc2a3, S100a14* and

Ier3 showed a significant increase in expression when estradiol treatment was provided to female gonads (**Figure 6.5**). In the liver fewer genes showed significant interaction effects between pairs of sex-biasing factors (FDR<0.05) (**Supplemental Table S6.2**). For instance, expression of *Cyp3a41a, Sult3a1* and *Cyp17a1* was downregulated by testosterone in mice with female gonads; *Obp2a* expression was upregulated by testosterone in mice with male gonads; expression of *Igfbp2* was upregulated by testosterone on female gonads but downregulated by testosterone on male gonads (**Figure 6.6**).

Comparison of mouse DEGs affected by sex-biasing factors with human sex-biased genes

To cross-validate the DEGs identified in our FCG mouse model, we compared them with sex-biased genes identified in human GTEx studies of liver (**Supplemental Table S6.4**) and adipose tissues (**Supplemental Table S6.5**) [290]. We found 80 out of 500 sex-biased genes (16%) in GTEx liver and 116 out of 500 sex-biased genes (23.2%) in GTEx adipose tissue were identified as DEGs affected by one or more sex-biasing factors in our FCG model. It is important to note the key difference between studies: the sex-biased genes in GTEx are the results of the combined effects of all sex-biasing factors whereas our FCG mouse study focuses on the effect of individual sex-biasing factors.

As GTEx studies cannot isolate specific sex biasing factors, our FCG model suggests the particular factors contributing to the sex biased genes found in humans. For instance, in adipose tissue, the GTEx female biased genes *ASAH1, PRDX2* and *LOXL1* might be explained by an effect of estradiol. In contrast, the male-biased adipose gene *HSD11B1* in GTEx can be explained in the FCG by the effect of testosterone (**Figure 6.7**). In the liver, the human male-biased genes *ADH4, GNA12*, *HSD17B12* can be explained in our mouse model by an effect of testosterone, whereas

the female-biased human genes *AS3MT, ZFX* and C*XCL16* were found to be affected by estradiol in FCG mice (**Figure 6.8**). Therefore, the FCG mice not only can recapitulate certain sex-biased genes in human studies but suggest the specific sex-biasing factors that contribute to the sex bias.

Coexpression modules affected by each sex-biasing factor

The above DEG analyses focused on genes that were individually influenced by sex-biasing factors as well as their interactions. Sets of genes that are highly coregulated or co-expressed can offer complementary information on coordinated gene regulation by sex-biasing factors that might be missed by the DEG-based analyses. To this end, we constructed gene coexpression networks for each tissue using MEGENA and identified 326 liver and 131 adipose coexpression modules. The first PCs of the coexpression modules were assessed for influence by sex chromosome, gonadal sex, and hormonal treatment factors using 3-, 2-, and 1-Was (**Figure 6.3G, 6.3H**). We confirmed the large effect of hormonal treatment in regulating modules enriched for diverse biological pathways. In the liver, testosterone affected modules involved in metabolism (RNA, lipid, protein), development, protein assembly, chemical response, immune system (inflammation, adaptive immune response), apoptosis, and transcription/translation. In adipose tissue, estradiol influenced modules related to focal adhesion, development, metabolism (protein, lipid, oxidative phosphorylation), immune system (complement and coagulation), and translation.

Gonadal sex also showed considerable influence on liver modules related to protein metabolism/assembly, development, stress/immune response, apoptosis, and transcription/translation regulation, whereas in adipose tissue gonadal sex mainly affected developmental and focal adhesion processes, and to a lesser degree, lipid metabolism, biological oxidation, and intracellular signaling modules (**Figure 6.3G**).

The coexpression network analysis also confirmed the limited effect of sex chromosomal variation on altering coexpression modules (**Figure 6.3G, 6.3H**). However, in adipose tissue, sex chromosomes showed weak effects on modules related to lipid metabolism and intracellular signaling when estradiol and blank groups were considered, but not when the testosterone group was included **(Figure 6.3H)**.

Overall, the gene coexpression network analysis offered clearer patterns of tissue-specificity and functional specificity of each sex-biasing factor compared to the DEG-based analysis.

Bulk tissue deconvolution to understand cellular composition changes through sex-biasing factors

To explore whether the DEGs and pathways/modules identified in FCG can be explained by cellular composition changes affected by each sex biasing factor, we carried out cell composition deconvolution analysis on the bulk tissue transcriptome data using CIBERSORTx based on single cell reference datasets of the corresponding tissues (**see Methods**). We subsequently assessed the hormonal, gonadal, and sex chromosomal effects on individual cell types.

In both the liver (**Figure 6.9**) and adipose tissue (**Figure 6.10**), hormones affected the largest number of cell types in terms of their abundance, including various immune cell populations such as the hepatocellular stellate cells (HSCs) and neutrophils in the liver, and macrophages, CD4 T-cells, dendritic cells, and antigen presenting cells in adipose tissue. Hormones also affected dividing cell populations and endothelial cells in both tissues. These cell populations affected by hormones support the DEGs and pathways involved in immune functions

and development that are influenced by the same sex-biasing factor. Similar to the findings based on DEG and pathways analysis, the gonadal effect on cell populations is also dependent on the tissue and other sex-biasing factors: female gonads exhibit increases in hepatocytes, endothelial, and HSCs in the liver on an XX background, whereas male gonads showed an increase in macrophage proportion in adipose tissue on an XY and testosterone background. Lastly, the sex chromosome effect can be noted in immune cell populations, but it is generally dependent on the interactions with other sex factors. Overall, the changes in cellular composition support the changes in the pathways highlighted through our DEGs and coexpression modules including immune, developmental and metabolic signals in both tissues.

Effect of hormonal treatment on gene expression direction across genotypes

Due to the dominant effect of hormonal treatment as compared to gonadal sex or sex chromosome differences based on the above analyses, we further investigated the differences between testosterone and estradiol treatments in terms of the gene sets they target and the direction of gene expression change within and between tissues.

Overlapping DEGs between testosterone and estradiol treatment

Comparing groups of DEGs regulated by testosterone or estradiol in the 3WA (**Figure 6.11**), 226 overlapped in the liver and 383 overlapped for adipose tissue. However, estradiol DEGs in individual genotypes had limited overlap with those caused by testosterone in 1WA (**Figure 6.12**). In particular, for the XYF mouse we found no overlapping DEGs in either the liver or adipose DEGs between testosterone and estradiol (**Figure 6.11**). For other genotypes, the overlapping DEGs in the liver (**Figure 6.12A; 6.12B; 6.12C**) and adipose tissues (**Figure 6.12D; 6.12E; 6.12F**) mostly had consistent directions of expression changes between hormones, except

that *Fmo3* (Flavin containing monooxygenase 3, important for the breakdown of nitrogen-containing compound) in XXM liver (**Figure 6.12A**) and all the shared DEGs in XXF liver (*C1qb, C1qc*, and *Vsig4*; complement pathway genes) (**Figure 6.12C**) were affected by testosterone (down) and estradiol (up) oppositely.

Identification of potential regulators of sex-biasing factors

Transcription factor (TF) network analysis

To understand the regulatory cascades that explain the large numbers of sex-biased genes affected by hormone treatments (**Figure 6.13**), we performed TF analysis using as input DEGs that passed an FDR<0.05 from 1WA specific to testosterone effects in the liver and estradiol effects in adipose tissue (**Table 6.1; Table 6.2**). For the testosterone liver DEGs, we identified 67, 66, 60 and 62 TFs for XYM, XXM, XXF and XYF respectively (**Figure 6.14A-D; Supplemental Table S6.6**). As expected, we captured gonadal hormone receptors including Androgen Receptor (AR) as a highly ranked TF in all genotypes and estrogen receptors (ESR1, ESR2, ESRRA) to be TFs with lower rank. We also found NR3C1 (Nuclear receptor subfamily 3; the glucocorticoid receptor important for inflammatory responses and cellular proliferation) to be among the top 5 TFs for all four genotypes and the top-ranked TF for XXF and XYF, which is consistent with a female bias for this TF found in the GTEx study [290]. A number of circadian rhythm TFs were found throughout all genotypes in the liver including CRY1, CRY2, PER1, and PER2, which is consistent with sex differences in body clocks [291]. Additional consistent TFs for testosterone effect in liver across multiple genotypes, where sex bias has been documented previously, include FOXA1/2, XBP1, HNF4A, SPI1, and CTCF.

An analysis of TFs that may mediate estradiol effects in adipose tissue identified 64, 61, 44 and 53 TFs for XYM, XXM, XXF and XYF respectively (**Figure 6.14E-H; Supplemental Table S6.7**). We found ESR1 and ESR2 as consistent TFs throughout the genotypes, except for XYF, where no classical estradiol or androgen receptor TF was captured. We also identified AR as a top TF in XYM and XYF. Notably, we found many TFs across our genotypes to be consistent with the TFs for female-biased genes in the Anderson et al. human adipose study [292]. Out of their top 20 ranked TFs for female-biased genes, we found 17 in our results for estradiol treatment in our genotypes, including ESR1, H2AZ, SUZ12, KDM2B, CEBPB and PPARG. The top TFs were generally consistent across genotypes, except KDM5A, POLR2B, KMT2C and CLOCK were particular to XXF.

When looking into the TFs that mediate estradiol's effects in XYM for potential male-biased regulation in adipose tissue, we found matches with 13 of the top 20 TFs from the Anderson et al. human adipose study. These included AR, CTCF, SMC1A, EZH2, ESR1, RAD21 and TP63, and many were also consistent in additional mouse [292, 293] and human studies including the GTEx [290, 292].

Gene regulatory network analysis

An alternative and complementary approach to the TF analysis above is to utilize a gene regulatory network approach to decipher the key drivers (KDs) that may drive sex-biased gene alterations in each genotype based on the DEGs found in 1WA (**Table 6.1; Table 6.2**). We note that these KDs did not overlap with the TFs identified above due to the incorporation of genetic regulatory information in network construction.

In the liver (**Figure 6.14I**), we saw overlapping KDs for testosterone DEGs across all four genotypes. *Cyp7b1*, which is important in converting cholesterol to bile acids and metabolism of

steroid hormones, was among the top 5 KDs for all genotypes. *Mgst3* (involved in inflammation), *C6* and *C8b* (complement genes), and *Ces3b* (xenobiotics detoxification) were top 5 KDs for 3 of the 4 genotypes (**Figure 6.14I**). We also identified KDs specific to particular genotypes (**Supplemental Table S6.8**) such as *Ces3a* (xenobiotics detoxification) for female gonads, *Slc22a27* (anion transport) for XXF, *Serpina6* (inflammation) for XYF, and *Hsd3b5* (steroid metabolism) for male gonads. Among these KDs, *Slc22a27* was previously found to be expressed predominantly in females and *Hsd3b5* and *Cyp7b1* were male specific [294], thus agreeing with our results.

For estradiol, 31 KDs were found for adipose tissue DEGs from the XXF, XXM and XYM genotypes (**Figure 6.14J; Supplemental Table S6.9**). The KDs included *Mrc1* (response to infection), which is the only overlapping top KD between genotypes XXF and XXM. KDs that were more highly ranked for XXM but still statistically significant in XXF included genes involved in extracellular matrix organization (*Prrx2, Mfap2, Col1a2*, and *Gas7*), and those specific to XXF are relevant to lipid synthesis/metabolism (*Tbxas1, Pla1a*) and immune function (*Adgre1* and *Mcub*). *Irf7* is the only KD for XYM, which has been recently suggested to be a TF in adipocytes with roles in adipose tissue immunity as well as obesity [295].

Disease association of the genes affected by sex-biasing factors

Finally, to test the disease relevance of the genes affected by sex-biasing factors, we used a marker set enrichment analysis (MSEA; **details in Methods**) to detect whether the DEGs highlighted in the 1WA overlap with genes previously identified to have SNPs associated with human diseases/pathogenic traits by GWAS. In brief, we mapped each of the GWAS SNPs to genes using liver and adipose eQTLs to represent disease-associated genes informed by GWAS. The mouse orthologs of these human GWAS disease genes were then compared with sex-biased

DEGs from FCG to connect the genes affected by individual sex factors with human disease genes. Of the 73 disease/traits screened for which full GWAS summary statistics was available, we focused on two broad categories, "cardiometabolic" (**Figure 6.15A;6.15B**) and "autoimmune" (**Figure 6.15C;6.15D**), both of which are known to show sex differences. For hormone DEGs, we focused on those that are directly relevant to the general human population to understand how testosterone or estradiol can affect disease outcomes on XYM (physiological males) or XXF (physiological females).

Disease association for hormone DEGs

When cardiometabolic diseases were considered, testosterone and estradiol DEGs in the adipose tissue from both the XYM and XXF genotypes showed extensive disease associations **(Figure 6.15A; Supplemental Table S6.10**). In contrast, liver DEGs for both hormones showed limited cardiometabolic associations, with specificity of testosterone DEGs for both T2D and LDL but no association for estradiol DEGs (**Figure 6.15B; Supplemental Table S6.11)**. In terms of autoimmune diseases, testosterone DEGs in both the adipose (**Figure 6.15C; Supplemental Table S6.12)** and liver (**Figure 6.15D; Supplemental Table S6.13**) from both XXF and XYM genotypes showed enrichment for disease associations. The estradiol DEGs in both tissues, however, had a genotype-dependent pattern for disease association. In particular, estradiol liver DEGs from XYM had no association with autoimmune diseases but DEGs in XXF were associated with all autoimmune diseases.

Overall, adipose DEG sets altered by both hormones for both cardiometabolic and autoimmune processes. For liver DEGs, the most significant associations were with autoimmune diseases and subtle T2D and LDL associations were identified for liver testosterone DEGs.

Disease association for gonadal sex DEGs

We also used MSEA to detect whether gonadal DEGs highlighted in 2WA (FDR<0.05) overlap with human disease genes informed by GWAS. For both adipose tissue (**Figure 6.15E; Supplemental Table S6.14**) and liver (**Figure 6.15F; Supplemental Table S6.15**), the gonadal DEGs on an estradiol background showed associations with cardiometabolic diseases or traits, whereas gonadal DEGs on a testosterone or blank background had limited or no disease association.

Disease association for sex chromosome DEGs and interaction DEGs

Due to the low number of DEGs captured for the sex chromosome effect or interactions among the sex-biasing factors, no enrichment results are possible through MSEA, therefore we queried whether these DEGs have been previously implicated in human diseases by overlapping the DEGs at FDR<5% with candidate genes from the GWAS catalog for 2203 traits. Both adipose tissue and liver DEGs demonstrating sex chromosome effects, or interactions between gonad and hormone, or interactions between sex chromosome and gonad, overlapped with GWAS candidates for numerous cardiometabolic and autoimmune diseases (**Supplemental Table S6.16-18)**.

3.3 Discussion

The variation in physiology and pathophysiology between sexes is established via the modulatory effects of three main classes of sex-biasing agents. The manifestations of these sex-dependent modulators impact disease incidence and severity, including metabolism-related diseases and autoimmune diseases [296, 297]. In this study, we separated the effects of these sex-biasing components using the FCG model, thus enabling the analysis of each contributing factor as well as their interactions in altering gene expression in inguinal adipose and liver tissues, which are relevant in systems metabolism and immunity.

Our data revealed distinct patterns between tissues in the relative contribution of each sex-biasing factor to gene regulation (**Table 6.1; Table 6.2**). In particular, the liver transcriptome is mainly affected by acute effects of testosterone, followed by acute effects of estradiol, organizational effect of gonadal sex, and sex chromosome complement, whereas inguinal adipose gene expression is primarily regulated by acute effects of estradiol, followed by gonadal sex, acute effects of testosterone, and sex chromosome complement. The genes and pathways regulated by the sex-biasing factors are largely different between factors, although metabolic, developmental, and immune functions can be regulated by both activational effects of sex hormones and gonadal sex (organizational effects). Sex chromosome effects were primarily associated with genes that reside on X and Y Chromosomes, along with a handful of autosomal genes involved in inflammation and metabolic processes that are downstream of the sex-biasing effects of X and Y genes. Cell deconvolution analysis supports that sex-biasing factors influence the proportion of diverse cell populations such as immune cells, hepatocytes, and dividing cells, suggesting that cellular composition changes may partially explain the observed genes and pathways. Lastly, the liver and adipose tissue genes affected by the sex-biasing factors were found to be downstream targets of numerous TFs and network regulators, not just the sex hormone receptors, and show association with human cardiometabolic and autoimmune diseases (**Figure 6.16**). Previously, sex differences in the liver transcriptome have been largely attributed to sex differences in the circadian rhythm and levels of Growth Hormone, which are established because of perinatal organizational masculinization of hypothalamo-pituitary mechanisms controlling Growth Hormone [280-282]. Genes regulated in this manner would be expected to appear in the gonadal effect DEGs. Our results suggest, however, that the acute activational effects of gonadal hormones might be a more important influence, because of the larger number of testosterone or estradiol DEGs compared to

gonad DEGs. Our results are in line with previous evidence that removal of gonadal hormones in adulthood eliminates most sex differences in mouse liver gene expression [284, 298], and that liver-specific knockout of estrogen receptor alpha or androgen receptor altered genes that underlie sex differences in the liver transcriptome [283]. It is possible that the effects of gonadal steroids during adulthood are required for some of the organizational effects of testosterone mediated via Growth Hormone action. In contrast to liver, gonadectomy does not eliminate sex differences in the adipose transcriptome [298], which agrees with our finding that the organizational effects of gonads play a strong role, in addition to estradiol, in adipose gene regulation. The striking tissue-specificity for each of the sex-biasing factors observed here highlights that individual tissues have unique sex-biased regulatory mechanisms.

We found that the gonadal sex factor primarily affects developmental pathways, cell adhesion, and metabolic pathways in adipose (**Table 6.2; Figure 6.3H**), which corroborates past evidence indicating that early gonadal sex status and associated hormonal release play critical roles in the development of sex differences and disease outcomes [299-301].

Compared to the organizational gonadal sex effects and activational hormone effects, the sex chromosome effects were minimal, and no coherent pathways were found for the sex chromosome-driving DEGs (**Table 6.1; Table 6.2**) or co-expression modules (**Figure 6.3G-H)**. The DEGs include those known to escape X inactivation (*Kdm6a*, *Eif2s3x*, *Ddx3x*) [278, 287] and their Y paralogues (*Eif2s3y, Ddx3y*). The X escapees are expressed higher in XX than XY cells, causing sex differences in several mouse models of metabolic, immune, and neurological diseases [277, 288, 302, 303].

As our comparative analysis of the three classes of sex-biasing factors clearly determined that the activational effects of gonadal hormones are the dominant factors, we further investigated

potential upstream regulatory factors that may control the sex-biased genes, using a gene regulatory network analysis and a TF analysis, revealing both expected and novel findings. In concordance with the importance of hormonal effects and consistent with recent human studies including GTEx searching for tissue-specific sex bias [290, 292], TFs for hormone receptors (AR and ESR1/2) were captured in the majority of genotypes (**Figure 6.14A-H**). Beyond the major hormonal receptors, within the liver numerous circadian related TFs were captured (PER1, PER2, CRY1 and CRY2). Although it is known that males and females have differing biological clocks [291], the contribution of hormones particularly in this rhythm is far from fully elucidated and our findings support that hormones need to be taken into account in liver circadian rhythm studies. In adipose tissue for estradiol treatment, however, we found that the XXF genotype has no significant signal for ERs, which may imply that estradiol's major contribution in adipose gene regulation is more importantly through TFs such as H2AZ, which have been shown to be essential for estrogen signaling and downstream gene expression [304]. In addition to TFs, we utilized a GRN analysis, revealing non-TF regulators. For the liver GRN (**Figure 6.14I**), key driver genes for testosterone DEGs are involved in immune processes (*Mgst3, C6, C8b*), steroid metabolism (*Cyp7b1* and *Hsd3b5*), and xenobiotic detoxification (*Ces3b* and *Ces3a*). In adipose tissue (**Figure 6.14J**), there were far fewer shared key drivers for estradiol DEGs across genotypes relative to the results in the liver with testosterone treatment, indicating that estradiol has more finely tuned interactions with the gonadal sex and sex chromosome genotypes than the broad effect of testosterone.

Lastly, to provide context to the health relevance of the liver and adipose sex-biasing DEG sets, we looked for GWAS association of these genes with human diseases/traits. We found that hormone-affected genes in adipose tissue were enriched for genetic variants associated with numerous cardiometabolic diseases/traits, but the enrichment was weaker for the liver DEGs

(**Figure 6.15A; 6.15B**). Another important area of sex difference is found within autoimmunity, which occurs more in females [305]. While both adipose and liver DEGs from multiple hormone-genotype combinations were enriched for autoimmune diseases, the liver DEGs, particularly those from the XXF genotype, had more prominent autoimmune association. Beyond the hormonal DEG enrichment in human disease/trait, we also found that DEGs caused by gonad type from both adipose and liver are involved in cardiometabolic disease (**Figure 6.15E; 6.15F**). Finally, despite minimal DEGs captured for the sex chromosome effect as well as the interactions between each sex biasing factors, we found overlap of these DEGs with various disease traits. The DEGs underlying disease associations may explain the differential susceptibility of males and females to these major diseases, and warrant further investigation to distinguish risk versus protection through the genes identified in this study.

The analyses presented in this study show an extensive dissection of the relative contribution of three classes of sex biasing factors on liver and adipose gene expression, their associated biological processes and regulators, and their potential contribution to disease. Importantly, many of the genes identified in our study were replicated in independent human studies such as GTEx, and our mouse study offers unique insights into the particular sex-biasing factors (hormones, sex chromosomes, or gonads) that likely contribute to the sex-biased gene expression in humans. Despite retrieving numerous new insights, we acknowledge the following limitations. First, gonadectomy and subsequent treatment of hormones may have caused activational effects that do not match the effects of endogenous physiological changes in the same hormones, leading to more predominant activational effects being observed. Second, the relative effects of testosterone and estradiol are affected by the doses of each hormone used. Testing additional doses is required for detailed comparison of effects of the two hormones. Third, we used

DEG counts as a measure of overall effect size to compare the various sex-biasing factors, which may be influenced by sample size and statistical power. Therefore, caution is needed when interpreting the results. However, we note that the sample sizes are comparable across sex-biasing factors and are adequate for mouse transcriptome studies with sufficient statistical power [306, 307]. Fourth, the comparison of mice with testes vs. ovaries does not map perfectly onto mice that had organizational effects of testicular vs. ovarian secretions because of the potential effects of the *Sry* transgene, which was present in tissues only of mice with testes. Lastly, only liver and inguinal adipose tissues were investigated, and other tissues warrant examination in future studies.

Overall, our data revealed tissue-specific differential gene expression resulting from the three sex-biasing factors, thereby distinguishing their relative contributions to the differential expression of key genes in a variety of clinically significant pathways including metabolism, immune activity, and development. Importantly, in addition to establishing the critical influence of hormones and their effect on the transcriptome in a tissue specific manner, we also uncovered and highlighted the underappreciated role of the sex chromosomal effect and organizational gonadal effect as well as interactions among sex-biasing factors in global gene regulation. Our findings offer a comprehensive understanding of the origins of sex differences, and each of their potential associations with health and disease.

3.4 Methods

Animals

Mouse studies were performed under approval of the UCLA Institutional Animal Care and Use Committee. We used FCG mice on a C57BL/6J B6 background (B6.Cg-TgSry2Ei Srydl1Rlb/ArnoJ, Jackson Laboratories stock 10905; backcross generation greater than 20), bred

at UCLA [285, 286]. Gonadal females and males were housed in separate cages and maintained at 23°C with a 12:12 light: dark cycle.

A total of 60 FCG mice, representing 4 genotypes (XXM, XXF, XYM, XYF), were gonadectomized (GDX) at 75 days of age and implanted immediately with medical grade Silastic capsules containing Silastic adhesive only (blank control, (B) or testosterone (T) or estradiol (E). Mice were euthanized 3 weeks later; liver and inguinal adipose tissues were dissected, snap frozen in liquid nitrogen and stored at -80°C for RNA extraction and Illumina microarray analysis.

RNA isolation, microarray hybridization, and quality control

RNA from liver and inguinal adipose tissue was isolated using TRIzol (Invitrogen, Carlsbad, CA). Individual samples were hybridized to Illumina MouseRef-8 Expression BeadChips (Illumina, San Diego, CA) by Southern California Genotyping Consortium (SCGC) at UCLA. Two adipose samples were removed from the total of 60 after RNA quality test (degradation detected). Principal Component Analysis (PCA) was used to identify three outliers among the adipose sample, which were removed from subsequent analyses. PCA was conducted using the prcomp R package [308] with the correlation matrix.

Identification of differentially expressed genes (DEGs) affected by individual sex-biasing factors

To identify DEGs, we conducted 3-way ANOVA (3WA), 2-way ANOVA (2WA), and 1-way ANOVA (1WA) using the aov R function. The 3WA tested the general effects of 3 factors of sex chromosomes, gonad, and hormonal treatments, as well as their interactions. The 2WA tested the effects of sex chromosomes and gonads as well as their interactions within each hormonal treatment group (T, E, or B) separately. For 1WA, we tested the effects of T (comparing T vs. B) and E (comparing E vs. B) within each genotype. Multiple testing was corrected using the

Benjamini-Hochberg (BH) method, and significance level was set to FDR <0.05 to define significant DEGs.

Co-expression network construction and identification of co-expression modules affected by individual sex-biasing factors

We used the Multiscale Embedded Gene Co-expression Network Analysis (MEGENA) [309], a method similar to WGCNA [310], to recognize modules of co-expressed genes affected by the three different sex-biasing factors. The influence of each sex-biasing factor on the resulting modules was assessed using the first principal component of each module to represent the expression of that module, followed by 3WA, 2WA, 1WA tests and FDR calculation as described under the DEG analysis section to identify differential modules (DMs) at FDR <0.05 that are influenced by each sex-biasing factor.

Annotation of the pathways over-represented in the DEGs and DMs

For each of the DEG sets and DMs that were significantly affected by any of the sex-biasing factors, we conducted pathway enrichment analysis against Gene Ontology (GO) Biological Processes and KEGG pathways derived from MSigDB using Fisher's exact test, followed by BH FDR estimation.

Gene regulatory network analysis

To predict potential regulators of the sex-biased DEGs, we used the Key Driver Analysis (KDA) function of the Mergeomics pipeline [45] and liver and adipose Bayesian networks. In brief, the Bayesian networks were built from multiple large human and mouse transcriptome and genome datasets [187, 311-314]. To identify the key driver (KD) genes within these networks, the KDA uses a Chi-square like statistic to identify genes that are connected to a significantly larger number

of DEGs than what would be expected by random chance. KDs were considered significant at FDR<0.05 and top KD subnetworks were visualized using Cytoscape [315].

Transcription factor (TF) analysis

To predict TFs that may regulate the sex-biased DEGs sets, we used the Binding Analysis for Regulation of Transcription (BART) computational method [316]. We followed the tool's recommendation of a minimum of 100 DEGs as input and an Irwin-Hall p-value cut off ($p < 0.01$) for identify TFs.

Marker Set Enrichment Analysis (MSEA) to connect sex biasing DEGs with human diseases or traits To assess the potential role of the DEGs affected by each of the sex-biasing factors in human diseases, we collected the summary statistics of human GWAS for 73 diseases or traits that are publicly available via GWAS catalog [317]. SNPs that have linkage disequilibrium of $r^2>0.5$ were filtered to remove redundancies. To map GWAS SNPs to genes, we used GTEx Version 7 eQTL data for liver and adipose tissues [318] to derive tissue-specific genes potentially regulated by the SNPs. We then used the MSEA function embedded in Mergeomics [45] to compare the disease association p-values of the SNPs representing the DEGs with those of the SNPs mapped to random genes to assess whether the DEGs contain SNPs that show stronger disease associations than random genes using a Chi-square like statistic.

Deconvolution of bulk liver and inguinal adipose tissue

We downloaded single cell RNA-seq data for mouse liver from GEO (GSE166178) and mouse inguinal adipose from GEO (GSE133486) as our reference datasets, and utilized the deconvolution tool CIBERSORTx [253] to impute cell fractions in each sample. Cell proportion estimates were compared across groups to identify cell types influenced by sex hormones using 1WA with posthoc analysis and by gonads or sex chromosomes using t-test.

3.5 Tables

Table 6.1 Liver DEGs affected by sex-biasing factors and the associated GO/KEGG pathways.

Analysis	Treatment/ Genotype group	Sex Factor	DEGs at FDR < 0.05	Top DEGs Based on FDR	Top DEGs Based on LogFC	Top GO/KEGG Pathways FDR<0.05
3-way ANOVA	T + B	Horm	**1378**	*Trim80, Cux2, Elovl3, Cyp3a41a, Sult3a1*	**Up**: *Elovl3, Serpina4-ps1, Cyp4a12a, Cyp2u1, Slco1a1* **Down**: *Cyp2b13, Fmo3, Cux2, Trim80, Eci3*	Lipid metabolic process Organic acid and metabolic process Cellular lipid metabolic process Primary bile acid biosynthesis
		Gonad	**93**	*Cyp3a41a, Sult3a1, Cxcl9, St3gal6, Ly6a*	**Up**: *Nat8, Cyp4a12a, Asns, Obp2a, Cpne8* **Down**: *Gbp2b, Themis2, Spic, Sult3a1, Cyp3a41a*	Defense response Immune system process Response to stress
		Chr	**8**	*Ddx3y, Eif2s3y, Xist, Kdm6a, Tmsb4x*	**Up:** *Eif2s3y, Ddx3y, Tmsb4x* **Down:** *Xist, Ntrk2, Kdm6a, Eif2s3x, Wfdc2*	Rig I like receptor signalling pathway Transmembrane receptor protein tyrosine kinase signalling pathway MAPK signaling pathway
	E + B	Horm	**333**	*Cyp17a1, Slc11a1, Trp53inp2, C1qb, Clmn*	**Up:** *Prtn3, Obp2a, Isyna1, Ear4, Acot11* **Down:** *Adam11, Gsta1, Cfap126, Pde6c, Tppp*	Carboxylic acid metabolic process Organic acid metabolic process Monocarboxylic acid metabolic process Complement and coagulation cascades
		Gonad	**209**	*Cxcl9, Cyp3a41a, Cd74, Slc11a1, Ccr5*	**Up:** *Nat8, Cyp4a12a, Tiam2, Susd4, Agpat6* **Down**: *Cyp3a41a, Thy1, Themis2, Tbc1d10c, Cyp4f18*	Defense response Lipid metabolic process Immune response (SLE)
		Chr	**8**	*Eif2s3y, Xist, Ddx3y, Kdm6a, Eif2s3x*	**Up**: *Eif2s3y, Hccs, Ddx3y, Tmsb4x, Mgrn1* **Down**: *Xist, Kdm6a, Eif2s3x*	Electron transport Ubiquitin mediated proteolysis Organ morphogenesis
2-way ANOVA	T	Gonad	**9**	*Obp2a, Nat8, St3gal6, Cd74, Cxcl9*	**Up**: *Asns, Obp2a, Nat8, Cyp4a12a, Gm4956* **Down**: *Li, Cxcl9, St3gal6, Qpct*	Cellular response to stress/nutrient levels/extracellular stimulus Amino Acid Synthesis
		Chr	**6**	*Ddx3y, Eif2s3y, Xist, Kdm6a, Tlr7*	**Up**: *Eif2s3y, Tlr7, Ddx3y, Tmsb4x* **Down:** *Kdm6a, Xist*	Traf6 mediated irf7 activation in Tlr7 signaling Interleukin 8 biosynthetic process
	E	Gonad	**53**	*H2-Ab1, H2-Eb1, H2-DMb1, Cd44, Cd74*	Up: *Cyp7b1, Plekhb1, Actr1b, Sh3glb2* **Down**: *Thy1, Tbc1d10c, Cd79b, AI467606, S100a8*	Asthma Leishmania Infection Allograft rejection Systemic Lupus Erythematosus
		Chr	**6**	*Eif2s3y, Xist, Ddx3y, Als2cl, Kdm6a*	**Up:** *Eif2s3y, Als2cl, Ddx3y* **Down**: *H2-DMb1, Kdm6a, Xist*	Asthma Allograft rejection
	B	Gonad	**115**	*Cyp3a41a, Sult3a1, Slc11a1, Susd4, Nox4*	**Up:** *Nat8, Cyp4a12a, Cyp2d9, Igfbp2, Susd4* **Down:** *Sult3a1, Cyp3a41a, A1bg, Themis2, Spic*	Systemic Lupus Erythematosus Functionalization of Compounds CYP450 arranged by substrate type Biological oxidations
		Chr	**5**	*Eif2s3y, Xist, Ddx3y, Kdm6a, Ntrk2*	**Up:** *Eif2s3f, Ddx3y* **Down:** *Xist, Ntrk2, Kdm6a*	Rig I like receptor signaling pathway Aging, Circadian Rhythm Fatty Acid Metabolism
1-way ANOVA Post-hoc	XXF	T	**139**	*Aqp4, A1bg, Gas6, Parp1*	**Up**: *Cyp4a12b, Serpina4-ps1, Elovl3, Cyp2u1, Aqp4* **Down**: *A1bg, Cyp2b13, Fmo3, Slc22a26, Cux2*	Biological Oxidations Metabolism (xenobiotic/drug/glutathione) Immune (complement/interferon)
	XYF	T	**123**	*Heg1, Aqp4, Cyp2u1, Il1a, Fam111a*	**Up**: *Cyp4a12a, Elovl3, Gm4956, Cyp2u1, Fst* **Down:** *Cyp2b13, Fmo3, A1bg, Cux2, Eci3*	Biological Oxidations Complement cascade Bile acid metabolism Metabolism (steroid/drug)
	XXM	T	**124**	*Spry4, Cux2, Sybu, Akr1d1, Slco1a1*	**Up**: *Serpina4-ps1, Elovl3, Obp2a, Slco1a1, Cyp4a12a* **Down**: *Cux2, Eci3, Trim80, Irx3, Akr1d1*	Biological Oxidations Immune system (complement/SLE) Cell development/Wnt signaling Drug/Protein/Bile acid/salt metabolism
	XYM	T	**44**	*Atp2a3, Unc79, Trim80, Slco1a1, Obp2a*	**Up**: *Serpina4-ps1, Elovl3, Slco1a1, Obp2a, Gadd45g* **Down:** *Fmo3, Cux2, Slc22a26, Trim80, Unc79*	Cell development / Localization Metabolism (Drug/Riboflavin) Immune (Complement/Innate)
	XXF	E	**20**	*Zfp367, Lamb2, Ahcyl2, Prtn3, Pdlim4*	**Up:** *Prtn3, C1qb, Li, C1qc, Vsig4* **Down**: *Ahcyl2, Lamb2, Zfp367*	Signal transduction Metabolic Pathways Complement Pathways Class B2 Secretin Family Receptors
	XYF	E	**2**	*Klk1b27, Reck*	**Down Only:** *Klk1b27, Reck*	Developmental Processes
	XXM	E	**29**	*Slco1a, Tasor2, Ywhae, Sult3a1, Obp2a*	**Up**: *Sult3a1, Obp2a, Slco1a1, Slc11a1, Serpina1e* **Down**: *Ywhae, Tasor2, Ankrd12, Spc24, Serpina6*	Nitrogen/Vitamin/ Steroid Hormones Metabolism Complement/Inflammatory Pathways Signal Transduction/Transcription
	XYM	E	**8**	*Adam11, Sult3a1, Spc24, Ifit2, Gsta4*	**Up Only**: *Ifit2, Sult3a1, Cyp17a1, Gsta4, Serpina6*	Immune response (interferon signaling) Protein complex assembly Biological Oxidations

Table 6.2: Inguinal adipose DEGs affected by sex-biasing factors and the associated GO/KEGG pathways.

Analysis	Treatment/ Genotype Group	Sex Factor	DEGs at FDR < 0.05	Top DEGs Based on FDR	Top DEGs Based on LogFC	GO/KEGG Pathways at FDR<0.05
3-way ANOVA	T + B	Horm	**275**	*Pals1, Hp, Lrg1, Serpina3n, Prtn3*	**Up**: *Serpina3m, Mt2, Agt, Odf3l2, Krt36* **Down:** *Fabp5, Crtac1, Clca5, Rad54b, P2rx5*	Cell Cell Adhesion Protein Signaling Pathway
		Gonad	**268**	*Pals1, Dnai1, Aida, Cited1, Slc12a2*	**Up:** *Rab9b, Sntg2, Dusp15, Pinx1, Wfdc21* **Down:** *Dnai1, Aida, Nrxn3, Rab25, Upk2*	Multicellular Organismal Development Cell Cell Adhesion Anatomical Structure Development Pathways in Cancer
		Chr	**15**	*Pals1, Xist, Ddx3y, Eif2s3y, Hccs*	**Up:** *Eif2s3y, Kdm5d, Ddx3y, Pals1, Tlr7* **Down:** *Xist, Aatk, Kdm6a, Eif2s3x, Il11ra1*	Interleukin 8 biosynthetic process
	E + B	Horm	**2029**	*Dnai1, Gas6, Greb1, Fbxw17, Lrg1*	**Up**: *Adamts19, Hpca, Rpp25, Greb1, Thbs1* **Down**: *F2rl3, Fabp5, Clca5, Crtac1, Ucp1*	Protein Metabolic Process Respiratory Electron Transport Focal Adhesion
		Gonad	**449**	*Dnai1, Prlr, Prr15l, Sptbn2, Ercc2*	**Up**: *Dusp15, Mtap7d3, Rab9b, Gm525, Kcnh2* **Down:** *Prr15l, Ercc2, Cldn3, Prom2, Rab25*	Anatomical Structure Development Nitrogen Compound Metabolic Process Aldosterone Sodium Reabsorption
		Chr	**10**	*Xist, Dnai1, Ddx3y, Eif2s3y, Eif2s3x, Hcc*	**Up:** *Eif2s3y, Ddx3y, Tlr7, Hccs, Gprasp1* **Down**: *Xist, Gm525*	Rig I like receptor signaling pathway Generation of precursor metabolites and energy
2-way ANOVA	T	Gonad	**161**	*Esrp1, Sephs1, Fermt1, Wnt7b, Lrrc26*	**Up:** *Rab9b, Mlec, Por, Cdsn, Rprml* **Down:** *Rab31, Wnt7b, AI646023, Esrp1, Fermt1*	Epithelial Cell Differentiation Excretion Cell Cell Communication Cell Carcinoma/ Junction Organization
		Chr	**11**	*Eif2s3y, Rbm35a, Sephs1, Fermt1, Ddx3y*	**Up:** *Eif2s3y, Ddx3y, Wnt7b, Fermt1, Lrrc26* **Down:** *Xist, Kdm6a*	Seleno amino acid metabolism Electron transport Basal Cell Carcinoma
	E	Gonad	**400**	*Mtap7d3, Elf3, Ercc2, Pkp1, Slc5a7*	Up: *Kcne11, Dusp15, Kcnh2, Mtap7d3, Gm525* Down: *Pkp1, Fermt1, Atad4, Elf5, Mfsd2a*	Arginine & Proline Metabolism Cell Junction Organization Cell Cell Communication Epidermis Development
		Chr	**22**	*Mtap7d3, Xist, Eif2s3y, Ddx3y, Elf3*	Up: *Eif2s3y, Ddx3y, Tlr7, Hccs, Mxra7* Down: *Xist, Mtap7d3, Kdm6a, Fermt1, Eif2s3x*	Neurotransmitter release Cell Differentiation
	B	Gonad	**70**	*Rnf208, Cited1, Dnai1, Prr15l, Pals1*	Up: *Cd300ld3, Pals1, Bik, Adck5, Capn5* Down: *Prr15l, Dnai1, Cited1, Rnf208, Tcfap2b*	Cell Cell Communication YAP1 and WWTR1 TAZ stimulated gene expression Epidermis Development
		Chr	**10**	*Xist, Pals1, Rnf208, Cited1, Dnai1*	**Up:** *Pals1, Ddx3y, Rnf208, Prr15l, Cited1* **Down:** *Xist, Eif2sx, Kdm6a*	Tight Junction interactions Electron Transport Rig I like receptor signaling
1-way ANOVA post-hoc	XXF	T	**26**	*Rad9b, Rfwd3, BC049762, Dusp15, Crtac1*	**Up**: *Slc47a1, Pitpnc1, Pals1, scl0003251.1_21, Rfwd3* **Down**: *H2-Q10, Crtac1, Slc4a1, Dusp15, Rad9b*	Cell Carcinoma Cell Development/Differentiation/Wnt Signaling Immune (Antigen Presentation)
	XYF	T	**1**	*Fabp5*	**Down Only:** *Fabp5*	Development Processes
	XXM	T	**13**	*P2rx5, Krt36, Grik5, F2rl3, AA792892*	**Up**: *Mt2, Krt36, Hp, Serpina3n, Lrg1* **Down**: *Aspg, Clca5, F2rl3, Grik5, AA792892*	Organ Development/Cell Division/cycle Immune (haemostasis, antigen presentation)
	XYM	T	**13**	*Sephs1, Col4a5, Abp1, Tnnt1*	**Up:** *Mt2, Apom, Tnnt1, Fam25c, Col4a5* **Down:** *Cxcl15, Aoc1, Sephs1, Rnf144a*	Cell proliferation/Migration
	XXF	E	**395**	*Echdc1, Ptpn6, Il10, Rad9b, S100a14*	**Up**: *Ppp1r27, Adamts19, Saa3, Thbs1, Krt15* **Down**: *Dnai1, Cited1, Aida, Lratd1, Clca2*	Immune System (Interferon, Antigen presentation)/ ERK pathways Metabolism (Porphyrins, tryptophan) Cancer/Protein Stabilization
	XYF	E	**100**	*Fam83f, Csf3r, Galnt9, S100a14, Efna4*	**Up:** *Tph1, Areg, Aoc1, Ly6g6c, Lgals7* **Down**: *Actc1, Aqp4, Gdap1, A930018M24Rik, Dnai1*	Complement Cascade Transport by Aquaporins Development (Cell migration, cell junction organization, ECM R interaction, Focal Adhesion)
	XXM	E	**228**	*Lrrc75b, Ahsp, P2rx5, Ildr1, 6330403K07Rik*	**Up:** *Ifi205, Hpca, Rpp25, Il1rl1, Nxnl2* **Down**: *Anxa8, Calca2, F2rl3, Grik5, Grb14*	Complement Cascade/Interferon Signaling ECM Organization Metabolism (glucose, glutathione, lipid)
	XYM	E	**110**	*1700088E04Rik, 6330403K07Rik, Dlx3, Mtcl1, Cacna1g*	**Up**: *Hpca, Spon1, Rpp25, Thbs1, Il31ra* **Down**: *Fabp5, Ccno, Trfr2, F2rl3, Cidea*	Immune (Interferon, Tlr4/ERK signaling) Metabolism (Nicotinamide, Glutathione, lipid)/ EGF Pathway/Cell Proliferation

6.6 Figures

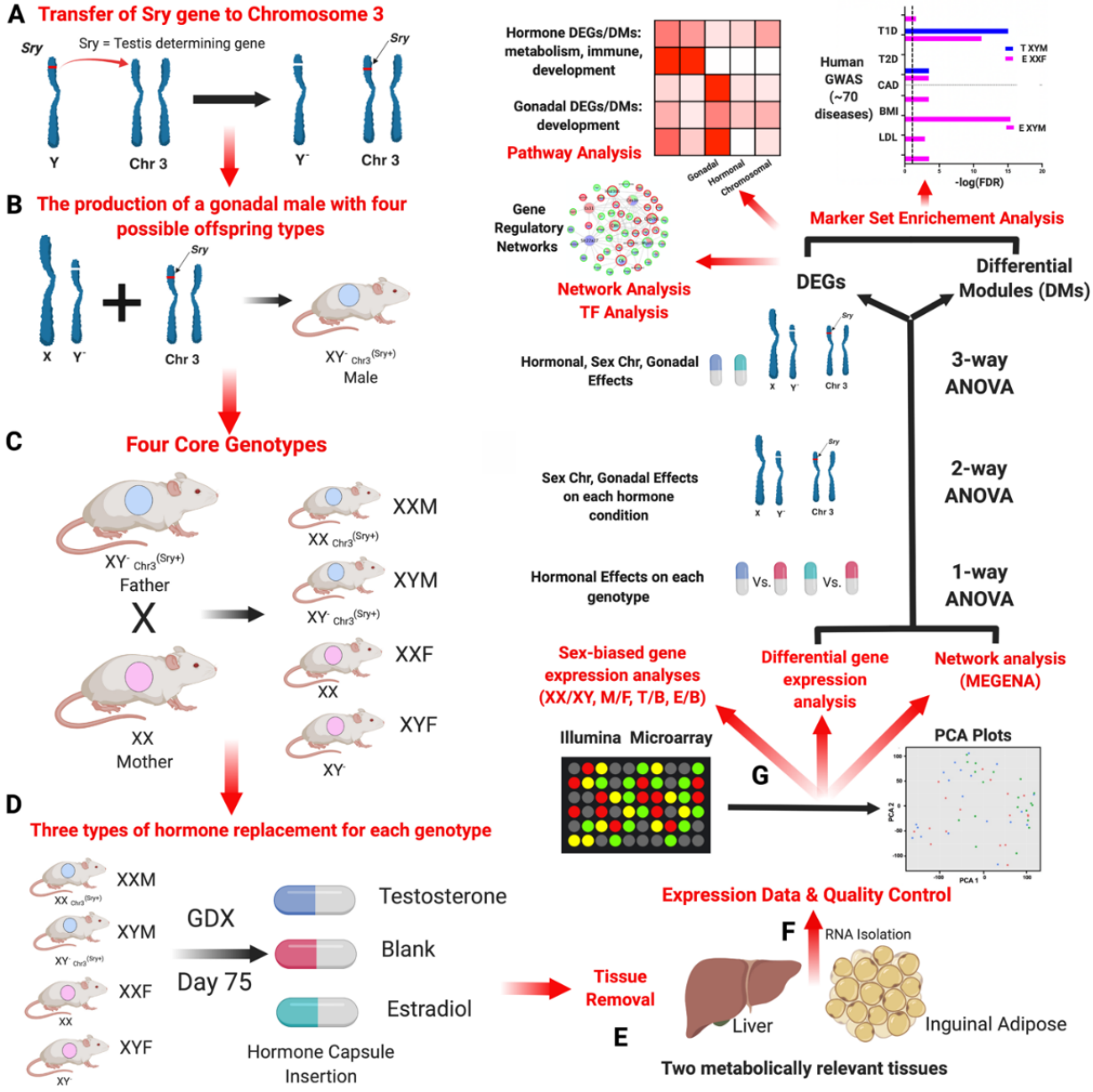

Figure 6.1. Overall Study Design. A. Transfer of the *Sry* gene to Chromosome 3. *Sry* which is usually located on the Y Chromosome was deleted (a spontaneous deletion) and inserted as a transgene onto Chromosome 3, making *Sry* independent of the Y Chromosome. **B.** The production of a gonadal male XY^{-} $_{Chr3}^{Sry+}$, which has the ability to produce 4 types of gametes resulting in the four core genotypes (FCG). **C.** The generation of the FCG mice. Mating of XY^{-} $_{Chr3}^{Sry+}$ male and XX female produces four types of mouse offspring (two gonadal males and two gonadal females): XY^{-}$_{Chr3}^{Sry+}$ (XYM), XX $_{Chr3}^{Sry+}$ (XXM), XX (XXF), XY^{-} (XYF). **D.** Modulation of sex hormones in mouse offspring of each genotype after gonadectomy (GDX). Each of the four core genotypes underwent GDX at day 75 and was implanted with a capsule that contained either estradiol, testosterone or blank (n = 5/genotype/treatment). **E.** Dissection of the liver and inguinal adipose

tissue for RNA isolation. **F.** Gene expression profiling and quality control. Using an Illumina microarray, we measured the transcriptome and then carried out a principal component analysis (PCA) to identify outliers and global patterns. **G.** Bioinformatics analyses. Differentially expressed genes (DEGs) influenced by individual sex-biasing factors were identified using 3-way ANOVA (chromosomal, gonadal and hormonal effects), 2-way ANOVA (gonadal and chromosomal effects under each hormone condition), and a 1-way ANOVA (estradiol and testosterone treatment effects in individual genotypes). Gene coexpression networks were constructed using MEGENA and differential coexpression modules (DMs) affected by individual sex-basing factors were identified using 3-way, 2-way, and 1-way ANOVAs. DEGs and DMs were analyzed for enrichment of functional categories or biological pathways. The relevance of the DEGs to human disease was assessed via integration with human genome-wide association studies (GWAS) for >70 diseases using the Marker Set Enrichment Analysis (MSEA). Transcription factor analysis and gene regulatory network analysis were additionally conducted on the DEGs derived from the one-way ANOVA.

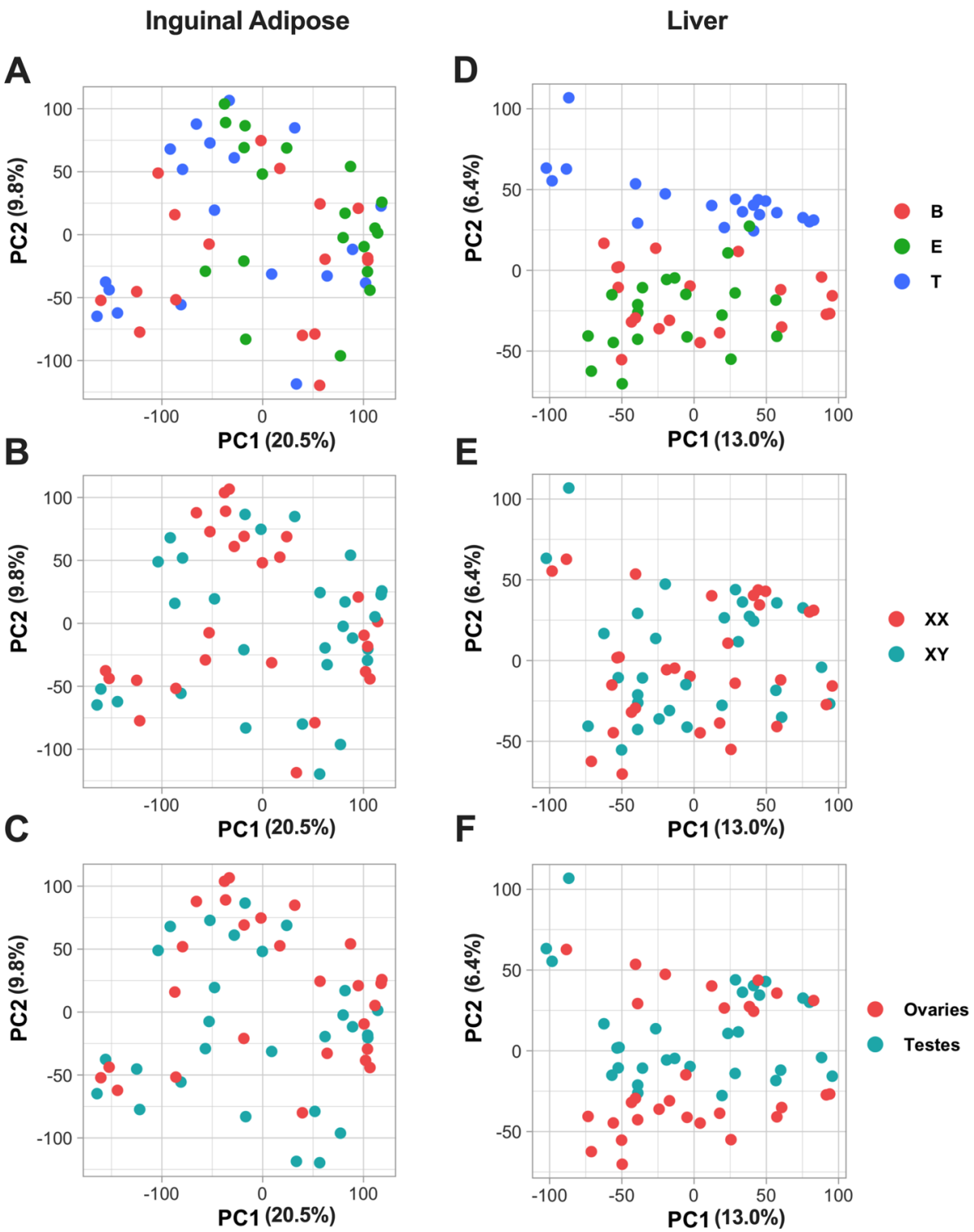

Figure 6.2. PCA plots of inguinal adipose and liver samples colored by different factors A. Inguinal adipose tissue PCA plot, with samples labelled by hormone types - blank (red), estradiol (green), and testosterone (blue). B. Inguinal adipose tissue PCA plot, with samples were labelled by sex chromosome categories - XX (red) and XY (blue). C. Inguinal adipose tissue PCA plot, with samples labelled by gonadal sex categories - ovaries (red) and testes (blue). D. Liver tissue

PCA plot, with samples were labelled by hormone types - blank (red), estradiol (green), and testosterone (blue). **E**. Liver tissue PCA plot, with samples labelled by sex chromosome categories - XX (red) and XY (blue). **F**. Liver tissue PCA plot, with samples labelled by gonadal sex categories - ovaries (red) and testes (blue).

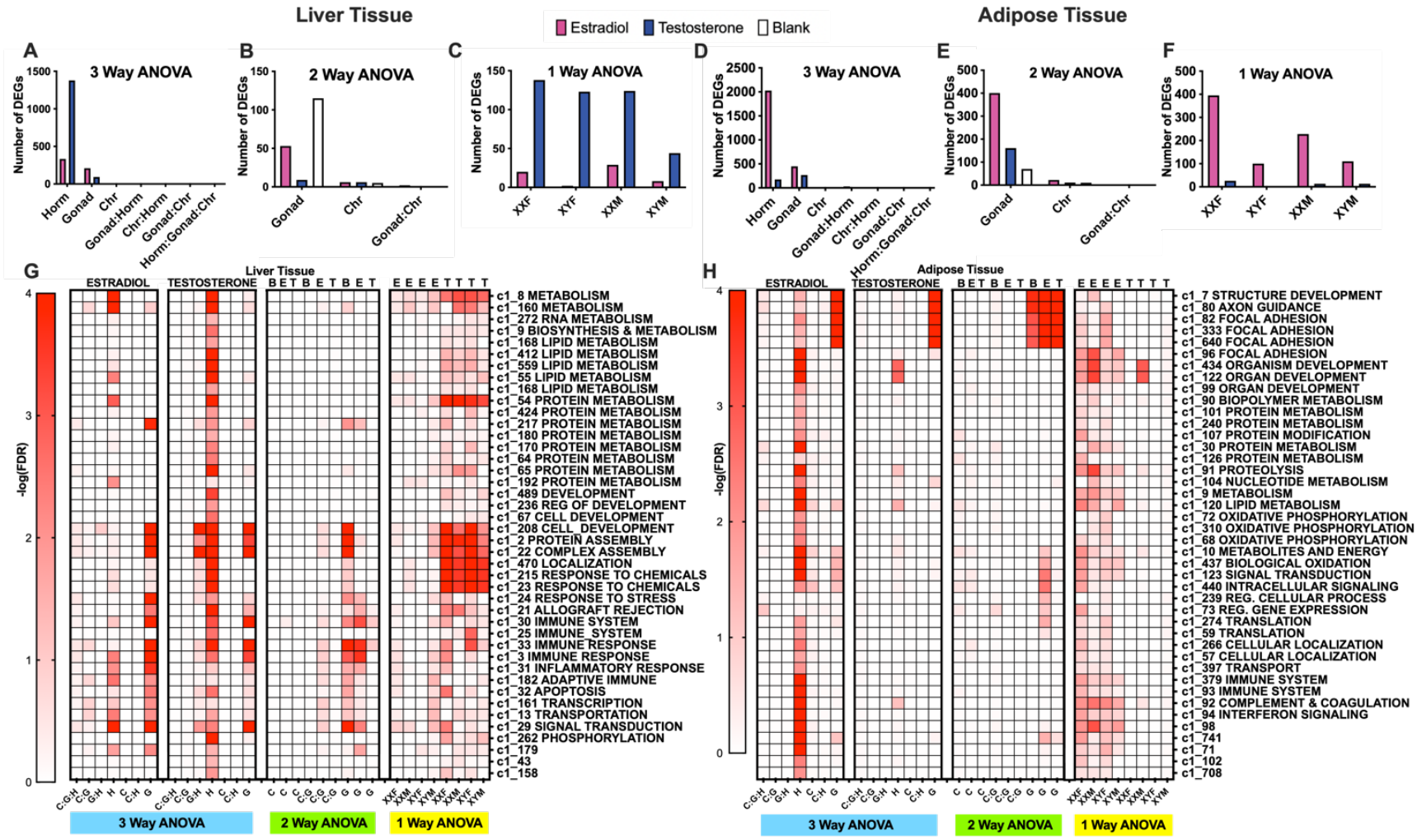

Figure 6.3. Bar graphs (A-F) and heatmaps (G-H) representing the number of DEGs for each sex-biasing factor and differential co-expression modules from a 3-way, 2-way, and 1-way ANOVA, respectively. Each bar graph represents the number of DEGs based on each specific statistical analysis at FDR<0.05. **A**, **D** represent results from 3-way ANOVAs run separately in testosterone vs. blank

groups, and estradiol vs. blank groups to examine hormone, gonad, and sex chromosome effects as well as the interaction terms. **B** and **E** represent results from 2-way ANOVAs with factors of gonadal sex and sex chromosomes as well as the interaction term, run separately on data from testosterone (T), estradiol (E), and blank (B) treatment groups. **C** and **F** represent results from 1-way ANOVAs testing effects of hormonal treatments (vs. Blank) in each of the four genotypes for liver and inguinal adipose tissue. In A and D, pink bars indicates estradiol vs. blank; blue bars indicates testosterone vs. blank. In panels **B** and **E**, colors represent the hormonal treatment condition (testosterone groups blue, estradiol groups pink, and blank groups white). In panels **C** and **F**, colors show effects of testosterone vs blank (blue) or estradiol vs blank (pink) in each of the four genotypes. Horm = Hormone, Chr = Sex Chromosome, M = Testes/*Sry* present, F = Ovaries present, no *Sry*. **G** represents the heatmap for liver. **H** represents the heatmap for adipose tissue. Each heatmap shows results from 1-way, 2-way, and 3-way ANOVAs for hormone (H), chromosome (C), and gonad (G) when treated with testosterone (T), estradiol (E) and blank (B). Interaction terms among H, C, and G were also tested. For instance, C:G:H indicates the interaction term among the 3 factors in 3-way ANOVA. The influence of each sex-biasing factor on the coexpression modules was assessed using the first principal component of each module to represent the expression of that module, followed by 3-way, 2-way, 1-way ANOVAs to identify differential modules (DMs) at FDR <0.05 that are influenced by the various sex-biasing factors. Each module was annotated with canonical pathways from GO and KEGG. Modules without pathway annotations did not show significant enrichment for genes in any pathways tested. Colors correspond to the statistical significance of the effects of sex factors on modules in the form of -log10(FDR).

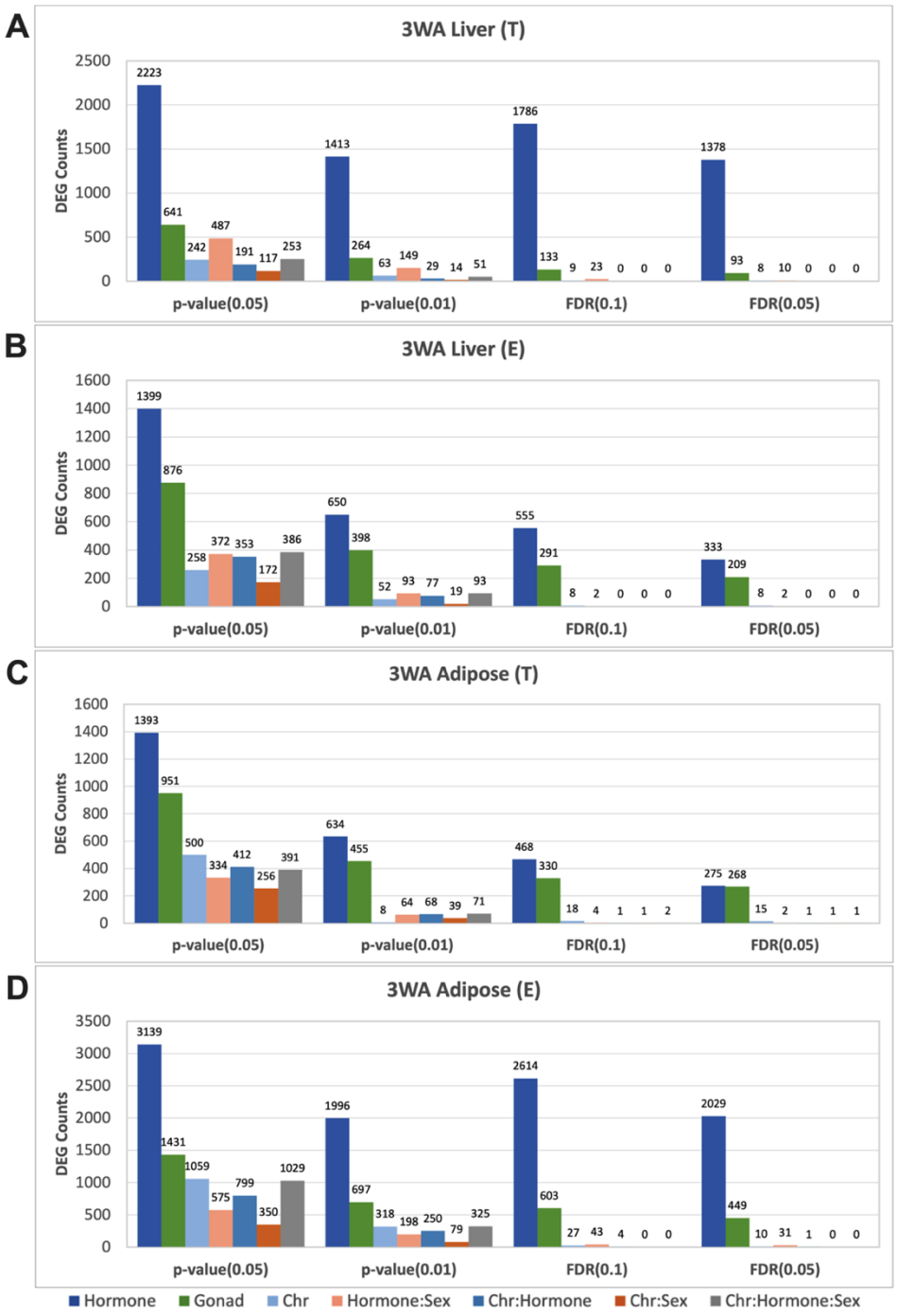

Figure 6.4. Bar graphs representing the 3-way ANOVA DEG numbers across various statistical cutoffs. Across the cutoffs there is a consistent trend of which sex biasing factor produces greater influence on DEG numbers: hormones (testosterone T or estradiol E) > gonad > sex chromosome. (A) Liver 3WA DEG counts when Testosterone (T) and blank groups are considered. (B). Liver 3WA DEG counts when Estradiol (E) and blank groups are considered. (C) Adipose 3WA DEG counts when Testosterone (T) and blank groups are considered. (D) Adipose 3WA DEG counts when Estradiol (E) and blank groups are considered.

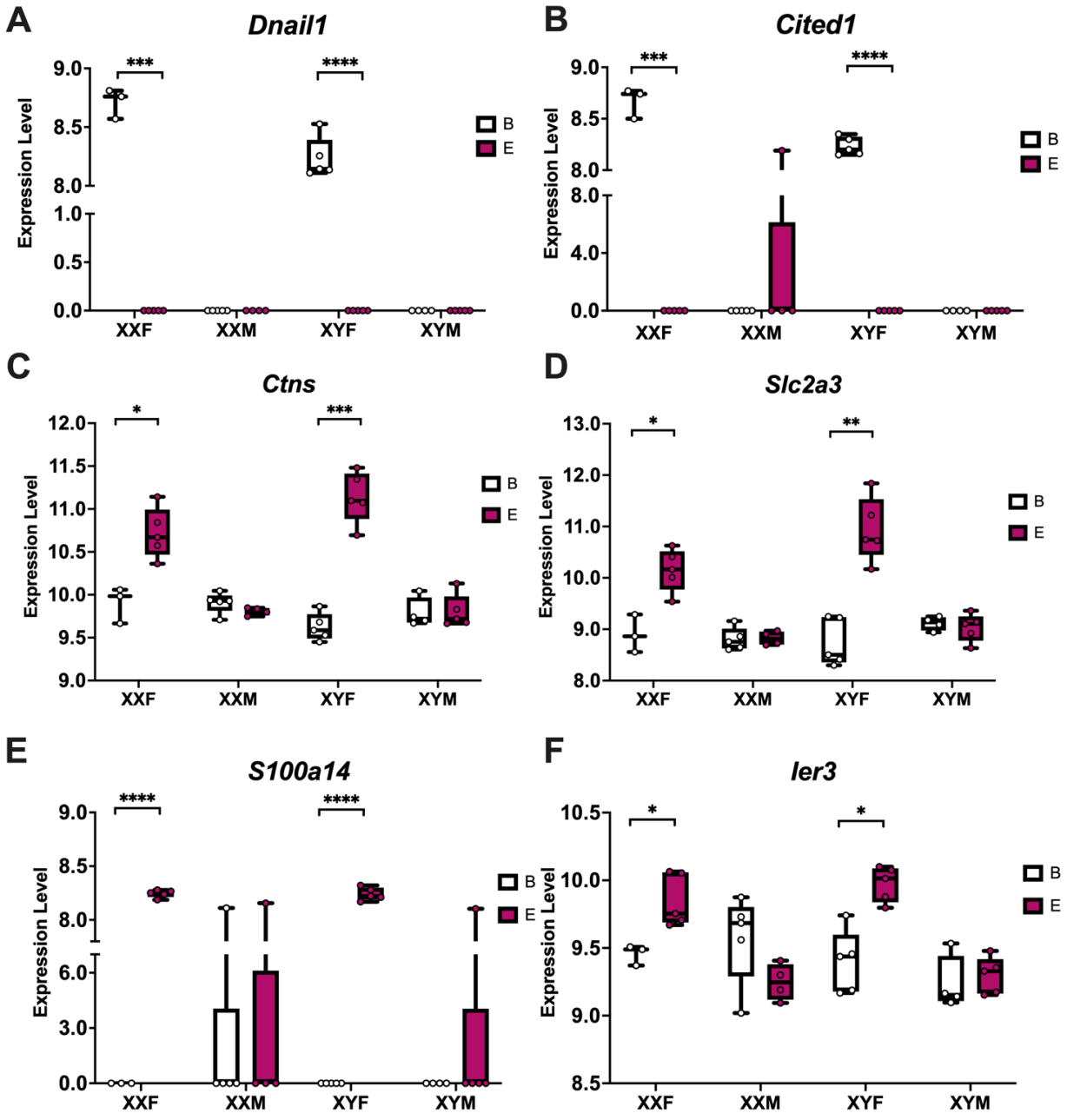

Figure 6.5. Bar Plots highlighting genes that showcase significant interactions between estradiol and gonad type in adipose tissue. T-test was used to calculate within genotype statistical significance. FDR<0.0001****, FDR<0.001***, FDR<0.01**, FDR<0.05*

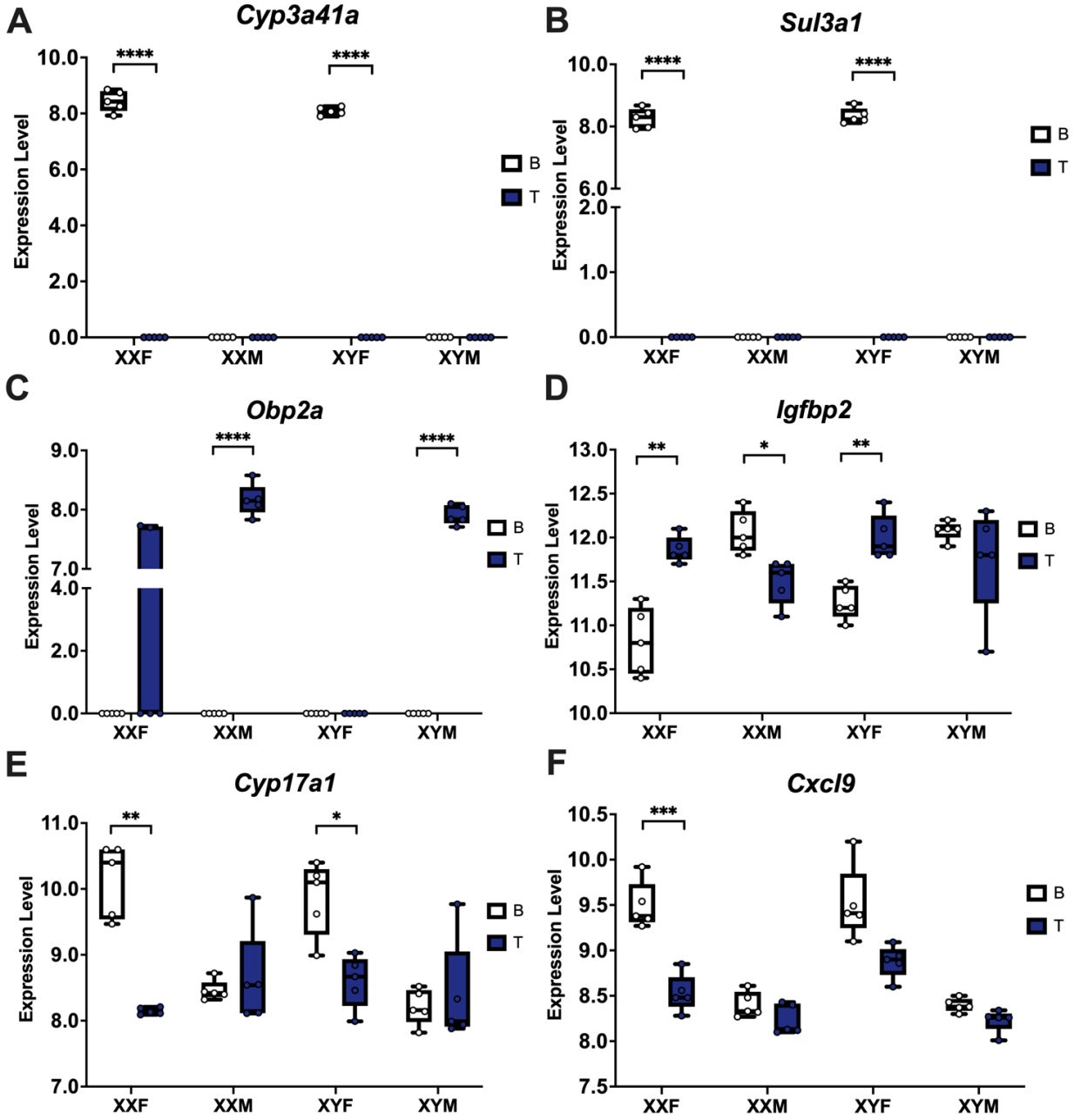

Figure 6.6. Bar Plots highlighting genes that showcase significant interactions between testosterone and genotypes in liver. T-test was used to calculate within genotype statistical significance. FDR<0.0001****, FDR<0.001***, FDR<0.01**, FDR<0.05*

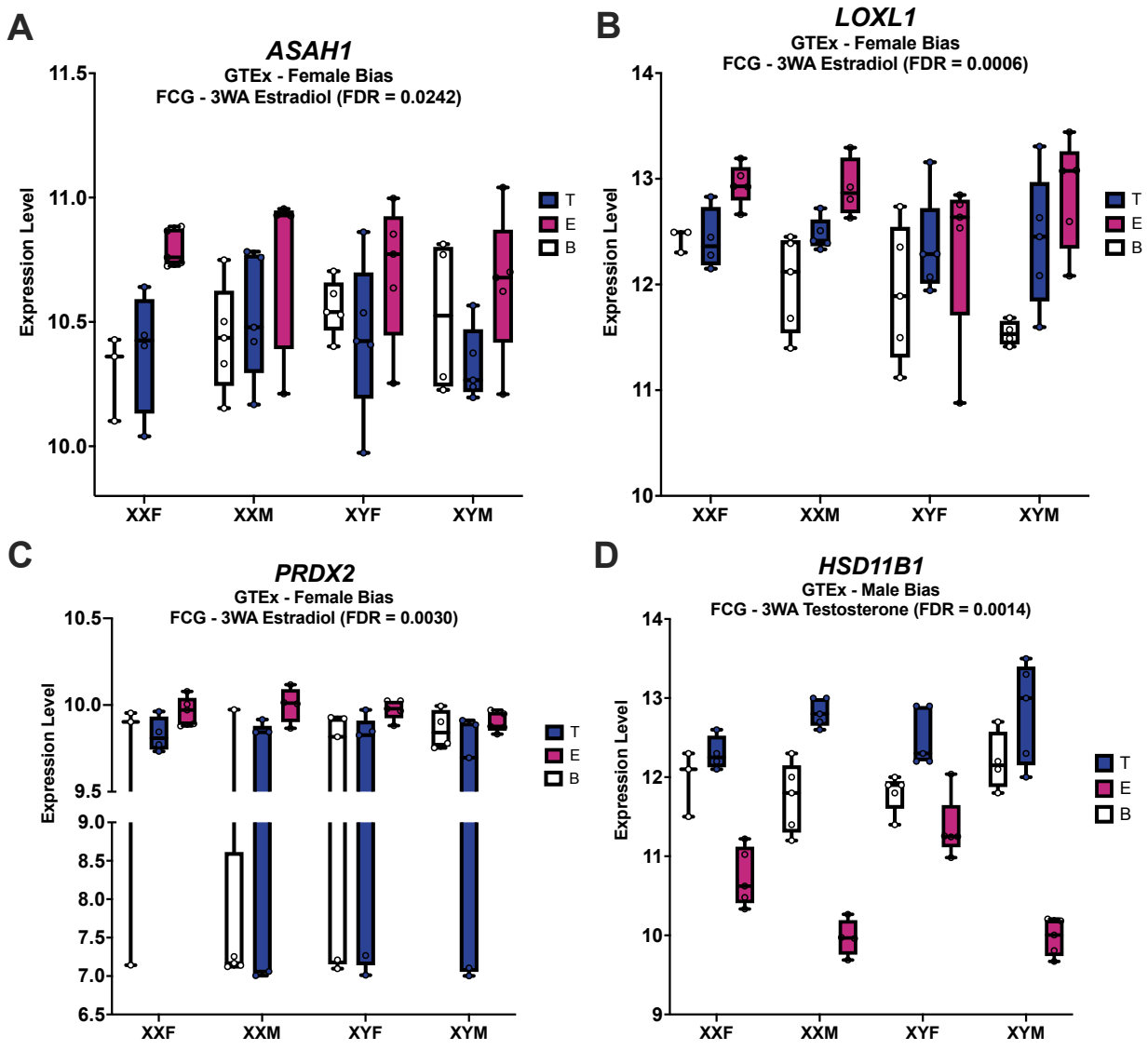

Figure 6.7. Bar Plots highlighting genes that showcase sex bias in the human GTEx study (Oliva et al. 2020) which also show a matched sex bias through one of the sex biasing factors in adipose tissue. We highlight in the heading of each panel the statistical test conducted and their associated FDR value.

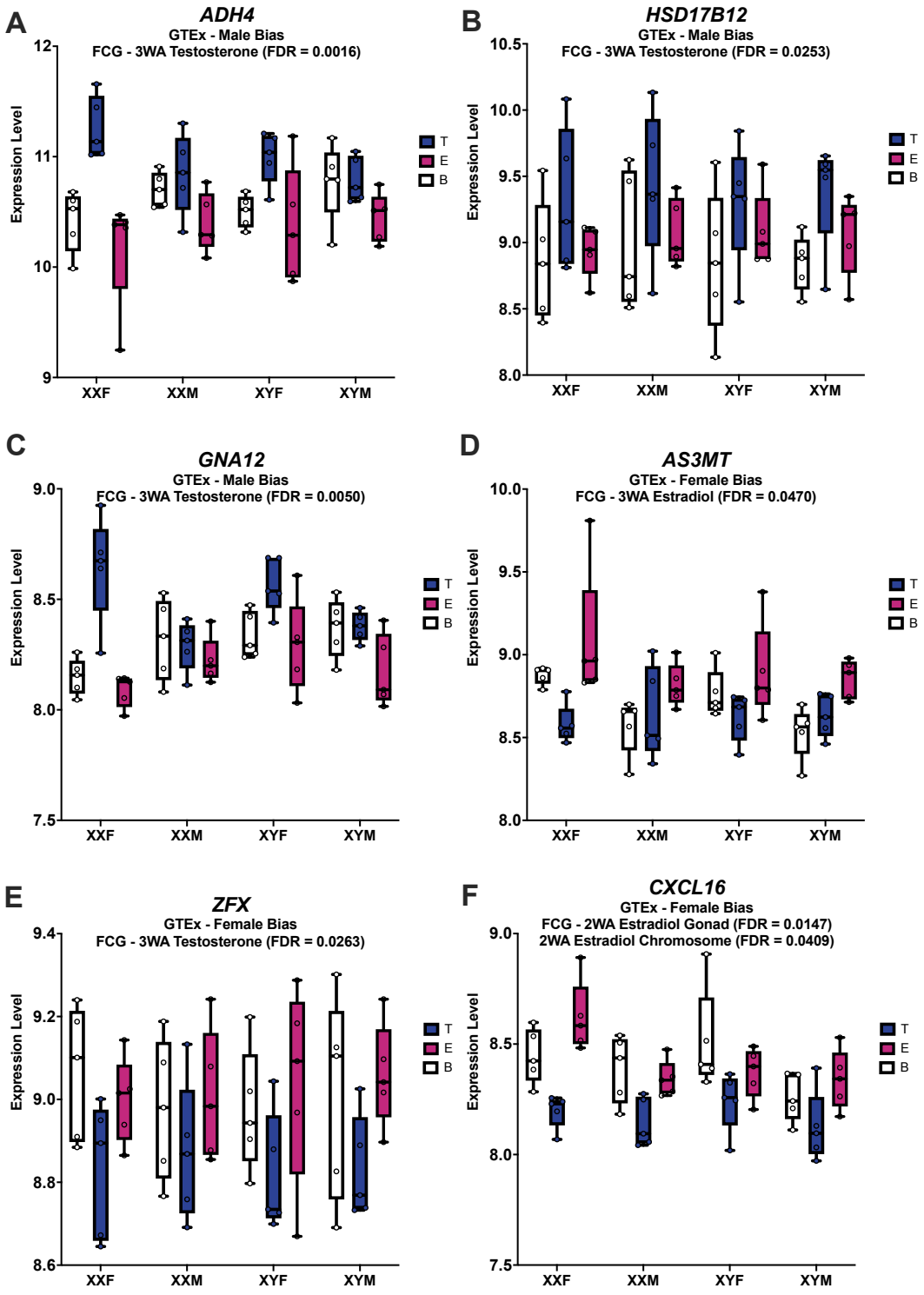

Figure 6.8. Bar Plots highlighting genes that showcase sex bias in the human GTEx study (Oliva et al. 2020) which also show a matched sex bias through one of the sex biasing factors in liver tissue. We highlight in the heading of each panel the statistical test conducted and their associated FDR value.

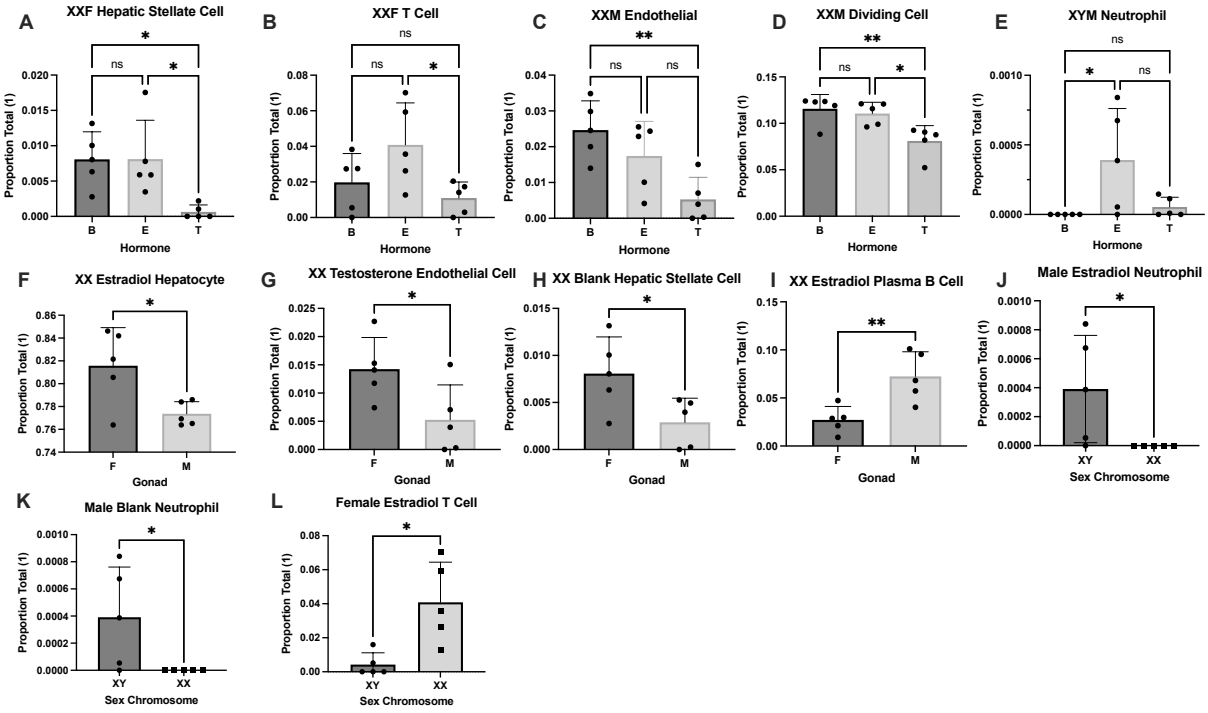

Figure 6.9. Deconvolution results for the liver, highlighting cell types that showed a statistical difference in cell type proportion between hormone treatments (A-E), gonads (F-I), and sex chromosome (J-L). * FDR<0.05, ** FDR <0.01, ns. not significant. For hormonal comparison we used the 1 Way ANOVA followed by post hoc analysis and for gonad and chromosome we used a T-test to calculate statistics.

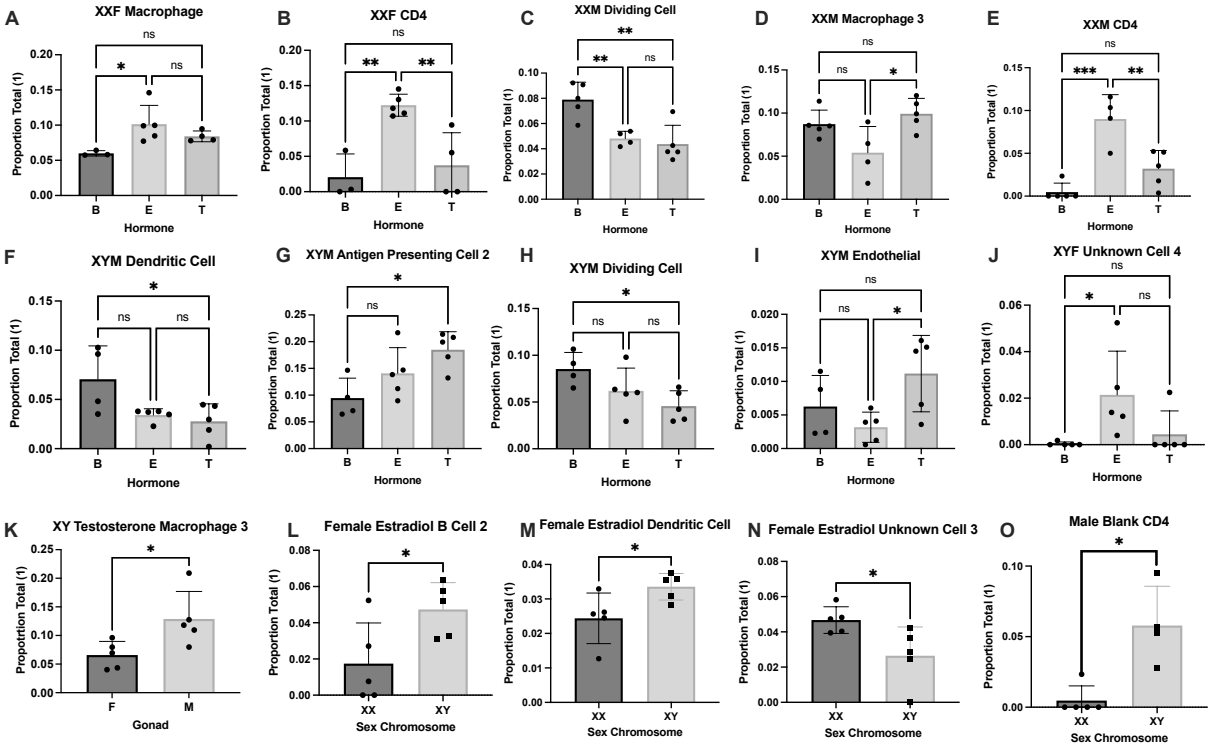

Figure 6.10. Deconvolution results for the adipose tissue, highlighting cell types that showed a statistical difference in cell type proportion between hormone treatments (A-J), gonads (K), and sex chromosome (L-O). * FDR<0.05, ** FDR <0.01, ns. not significant. 1 Way ANOVA followed by post hoc analysis to calculate statistics for those with three groups and T-test was used for those with two groups.

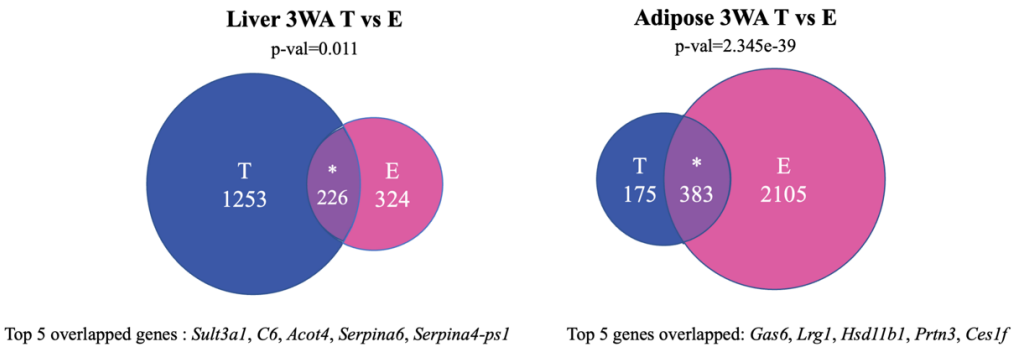

Figure 6.11*:* Venn diagrams representing 3-way ANOVA DEG comparison between testosterone (T)- and estradiol (E)-treated group in liver and inguinal adipose tissue.

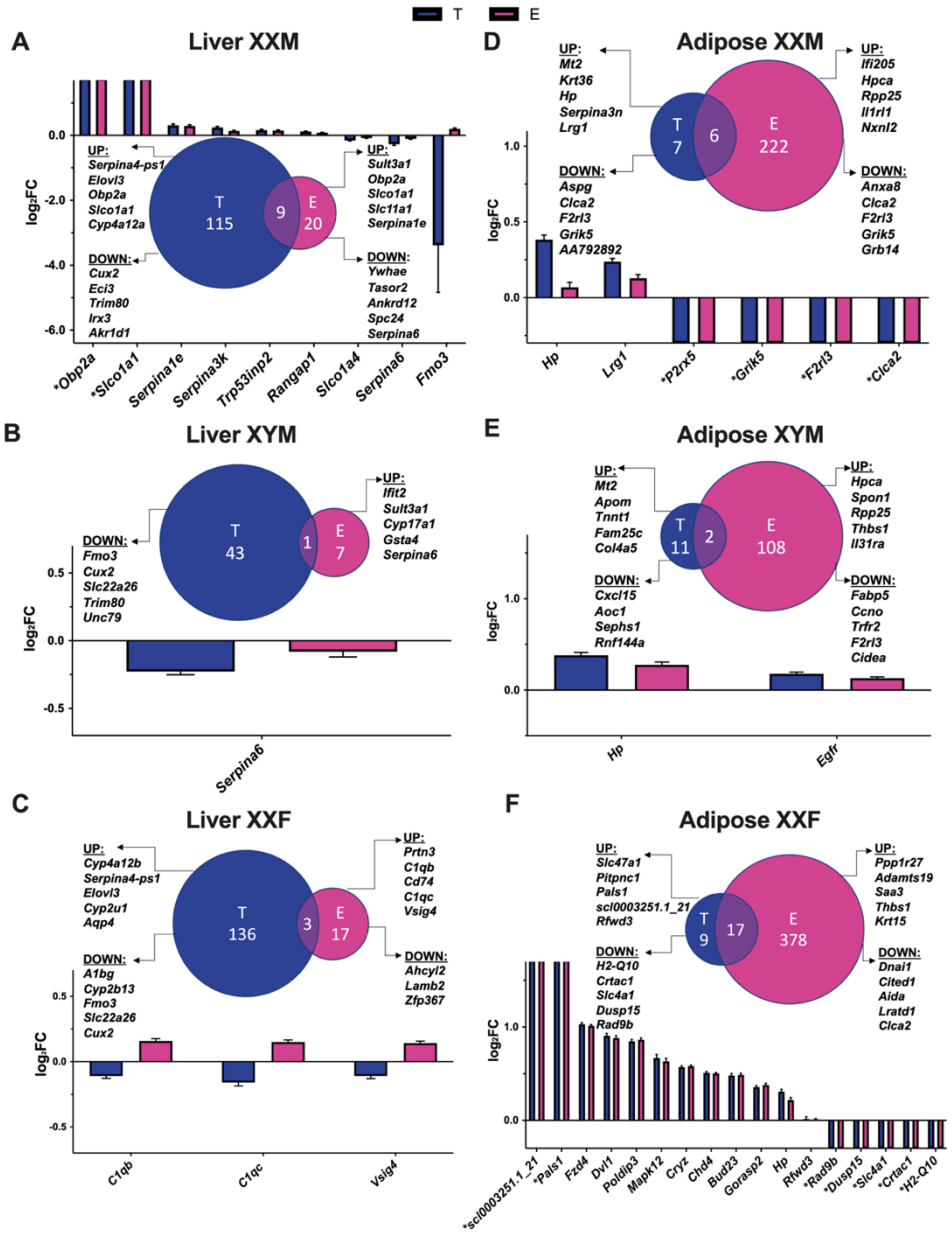

Figure 6.12. Venn Diagrams of DEG comparisons and Bar Graphs of overlapping DEGs between estradiol (E vs blank, abbreviated as E) and testosterone (T vs blank, abbreviated as T) treatment for each genotype in liver and adipose. **A**. Liver XXM. **B**. Liver XYM. **C**. Liver XXF. **D**. Adipose XXM. **E**. Adipose XYM. **F**. Adipose XXF. The bar graphs focused on the DEGs that passed an FDR < 0.05 and were overlapping between testosterone and estradiol treatment for each genotype and tissue. To understand the effects of each hormone, we plotted the $\log_2$ fold change (log2FC) of the hormonal effects. The Venn diagrams showcase comparison of DEGs of T effect vs E effect, as well as the top 5 up and down regulated genes for T or E in liver or adipose tissue for each genotype. There was no statistically significant overlap between any comparisons in the Venn diagrams. *represents genes that are not expressed in one of the comparison groups and thus have infinite log2FC values.

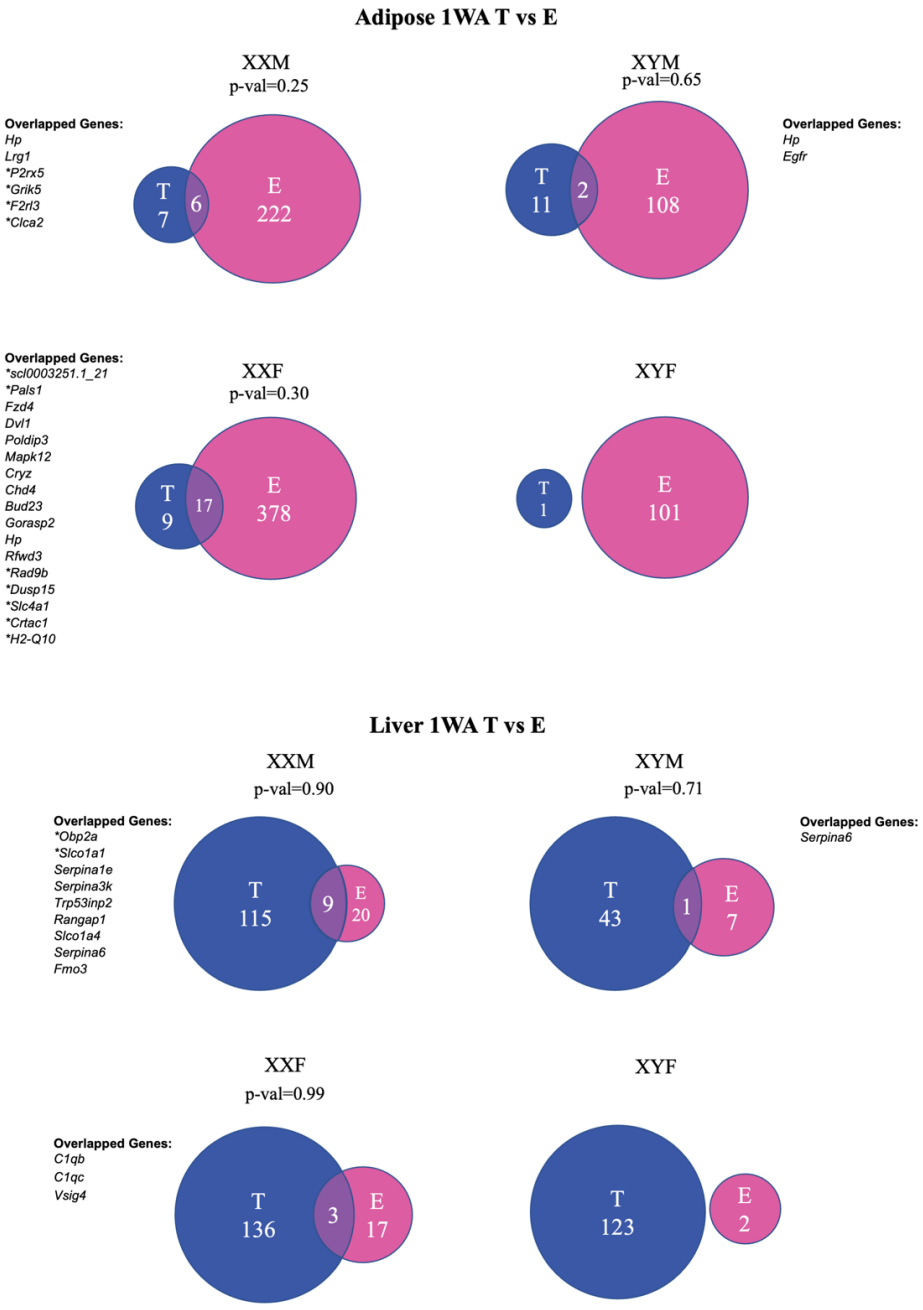

Figure 6.13. Venn diagrams representing 1-way ANOVA DEG comparison between testosterone (T)- and estradiol (E)-treated group across all genotypes in liver and inguinal adipose tissue.

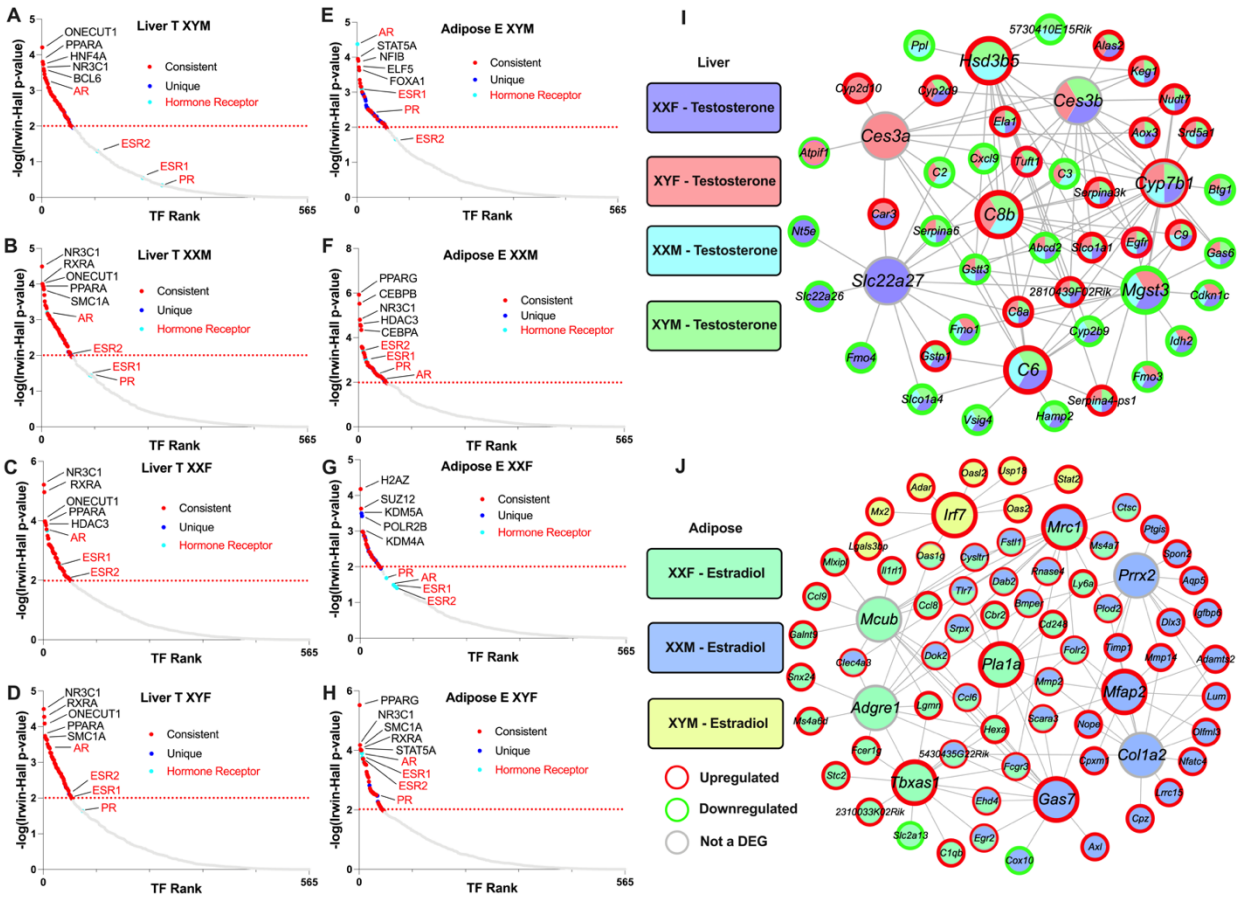

Figure 6.14. Transcription factor analysis (A-H) and key driver analysis (I-J) of DEGs informed by estradiol and testosterone treatment in liver and adipose. **A-D** represents TF analysis for liver. **E-H** represents TF analysis for adipose. For the TF network we utilized DEGs (FDR <0.05) from our 1WA for testosterone and estradiol treatment analysis using the BART tool, where a TF was considered significant by an Irwin-Hall p<0.01 analogous to -log10(p-value) = 2. Red color signifies the TF is present in at least one other genotype and the blue color signifies if the TF is only present in the given genotype. Turquoise color and red font denote a hormonal receptor relevant to testosterone and estradiol. Labelled TFs showcase the Top 5 by rank and additional hormonal receptors. **I.** Liver gene regulatory network (GRN). **J.** Adipose GRN. For GRN construction we overlaid DEGs (FDR <0.05) from our post hoc 1WA for testosterone and estradiol treatment onto our previously built adipose and liver Bayesian networks utilizing a KDA analysis from the Mergeomics package. We visualized the top 5 KDs for the testosterone or estradiol DEGs from each genotype group. KDs are labeled as larger nodes and DEGs as smaller nodes. Direction of DEGs is annotated with red or green borders for upregulation or downregulation, respectively.

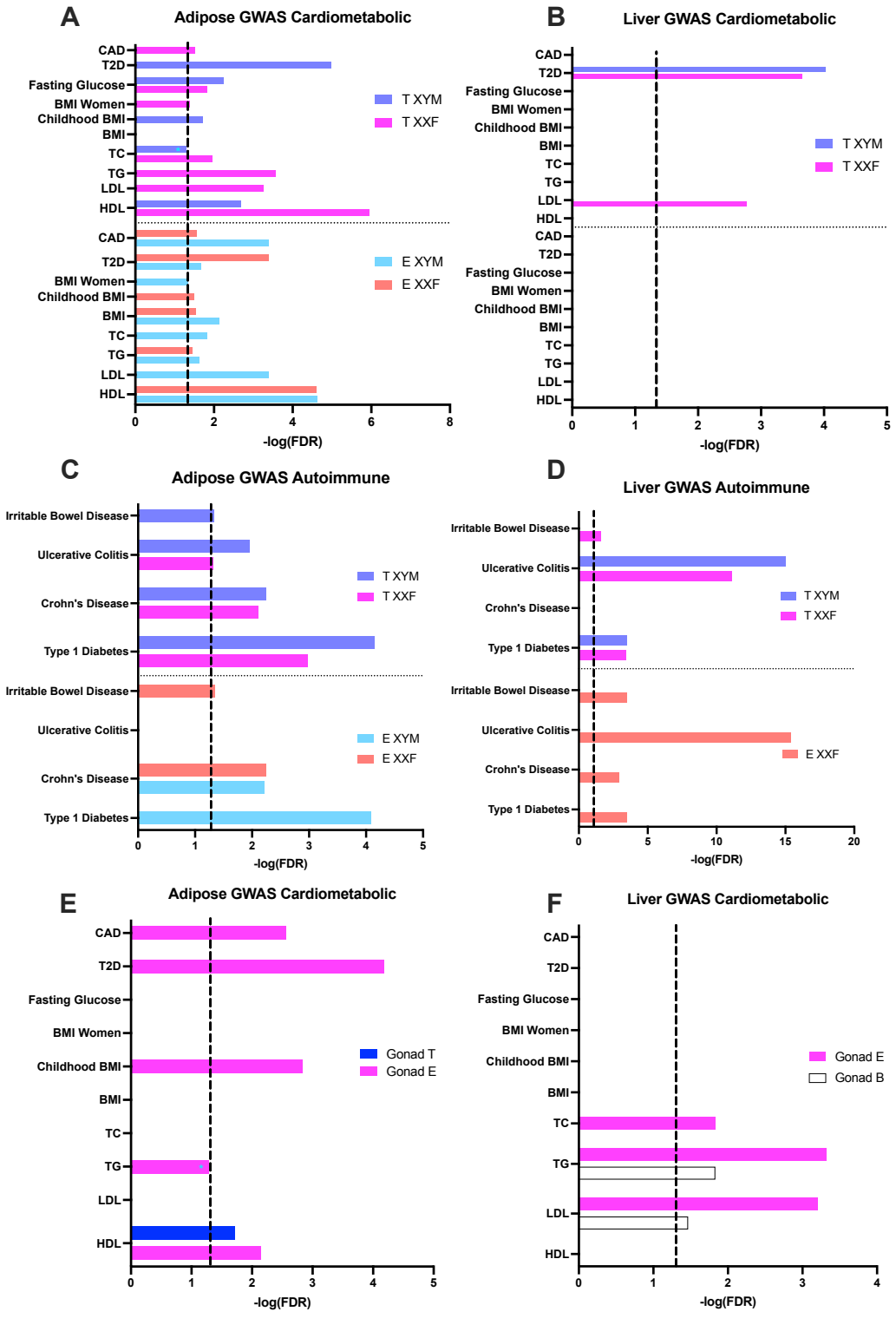

Figure 6.15. Bar graphs showing enrichment of the hormone DEGs (A-D) and gonadal DEGs (E-F) for known cardiometabolic and autoimmune diseases based on MSEA analysis. The

cardiometabolic category included Coronary Artery Disease (CAD), Type 2 Diabetes (T2D), fasting glucose level, BMI in women, BMI during childhood, BMI, total cholesterol (TC), triglyceride (TG), low-density lipoprotein (LDL) cholesterol, and high-density lipoprotein (HDL) cholesterol **(Figure 3.15A-B)**. The autoimmune category included Irritable Bowel Disease (IBD), Ulcerative Colitis (UC), Crohn's Disease (CD), and Type 1 Diabetes (T1D) (**Figure 3.15C; 3.15D**). **A.** Association of adipose testosterone (T) and estradiol (E) DEGs with cardiometabolic diseases/traits. **B.** Association of liver T and E DEGs with cardiometabolic diseases/traits. **C.** Association of adipose T and E DEGs with autoimmune diseases. **D.** Association of liver T and E DEGs with autoimmune diseases. **E.** Association of adipose gonadal DEGs with cardiometabolic diseases/traits. **F.** Association of liver gonadal DEGs for cardiometabolic diseases/traits. **A-D**. Hormone DEGs at an FDR <0.05 derived from the posthoc one-way ANOVA were tested against genetic association signals with cardiometabolic and autoimmune diseases and traits. **E** and **F** Gonadal DEGs at an FDR < 0.05 from two-way ANOVA were tested against genetic association signals with cardiometabolic diseases. Dotted line signifies FDR <0.05 and *denotes enrichment minimally below the FDR< 0.05 cutoff.

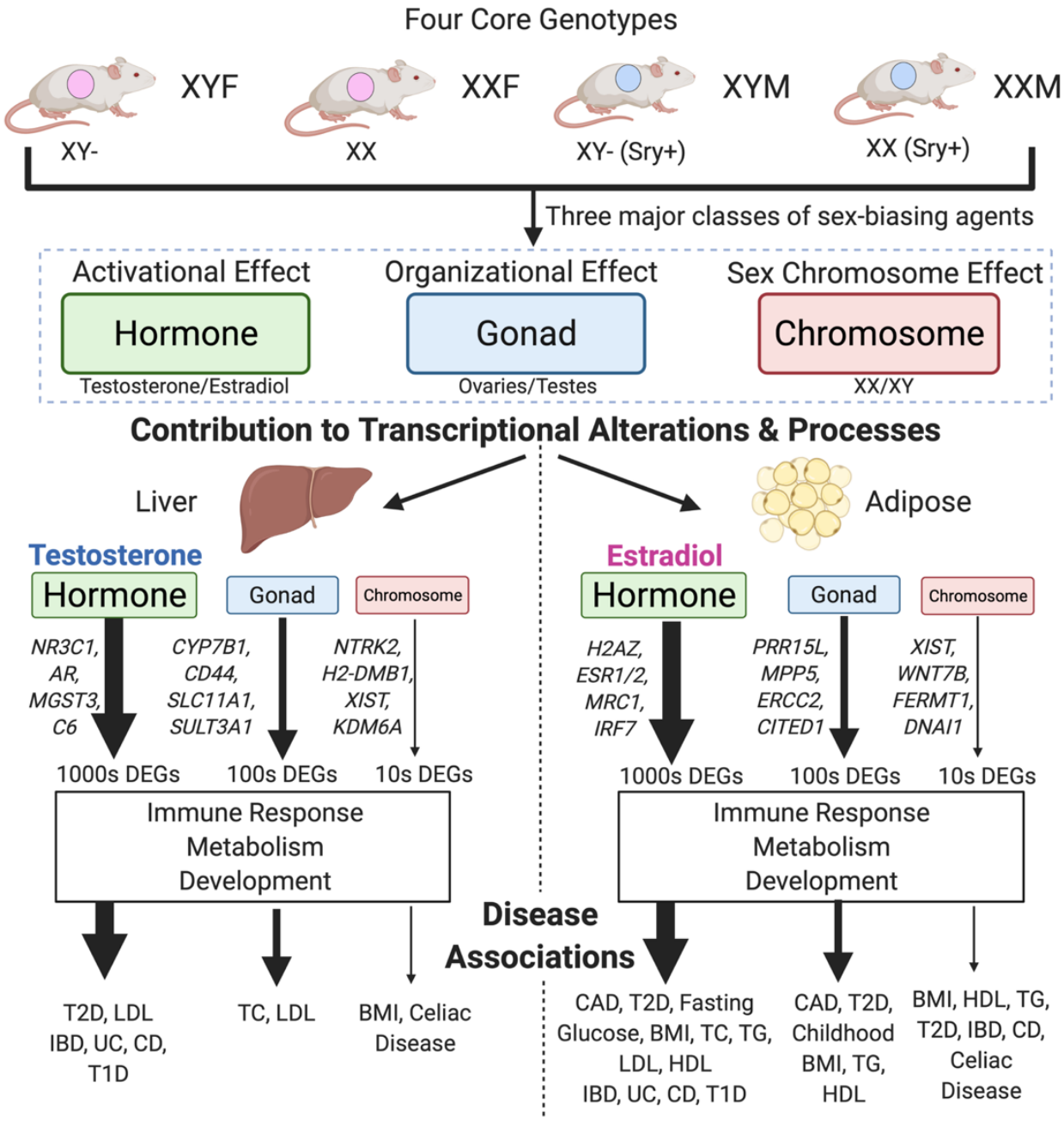

Figure 6.16. Study Summary. Utilizing the FCG model, we separated the effects of three major classes of sex-biasing agents and uncovered their relative contribution to transcriptional alterations in the liver and adipose tissue, the resulting biological processes enriched and finally the diseases associated.

Chapter 7. Conclusion and Future Directions

This dissertation work, had three major aims, which included the development of a user-friendly bioinformatic tool (Aim 1), an investigation into the role of sex factors in gene regulation in tissues important in MetDs (Aim 2), and examining the gene networks underlying MetDs and drug candidates for treatment (Aim 3). These aims fully exploit the power of systems biology through computational and experimental approaches to achieve a better understanding of complex MetDs and the role of sex factors.

As can be seen from this body of work, MetDs are highly complex and have a tremendous number of contributing factors, which include different omics layers interacting with one another, multiple tissues that cross talk to contribute to disease outcome, sex differences through the three sex biasing factors, as well as comorbidities, which can play off one another to exacerbate each metabolic outcome. As has been highlighted in our work, we believe this complexity should not be considered through one unique layer at a time e.g. genomics alone, but where possible to integrate multiple data types (omics, tissues, sex information) to create a complete picture, which is more likely to help explain a multifaceted complex disease.

We have tried to achieve this through a multi-tissue multi-omics systems biology approach, which takes into consideration multiple omics types and tissue contributions. With the belief that this approach will unravel some of the complexities behind diseases such as MetDs, we developed a far more accessible web server Mergeomics 2.0 to allow the wider scientific community access to this approach with their own datasets or even sample datasets to highlight the biology that can be uncovered as well as pinpoint molecular targets, which may have been previously untapped for a role in disease or a given trait. In this dissertation, we highlighted just as previous studies have before that our approach is robust in pinpointing novel genes with validation of coagulation factor

II (*F2*) for a role in lipid transport and metabolism in adipose tissue. We have additionally, highlighted numerous targets in liver fibrosis, CAD, T2D and T1D, which in the near future will be tested through in vivo experiments for their potential to mitigate disease. Another area, which we have been exploring is using our newly built drug repositioning tool PharmOmics to target disease network genes with matching drug target genes. Through this process we have been trying to prioritize drugs that are already FDA approved for another indication to repurpose them for our given trait or disease of interest. We are now working with electronic health records from the UK Biobank, and UCLA hospital, as a form of in silico validation to see if any of the candidate drugs have evidence for improving our phenotype of interest prior to experimental validation.

Another layer, which is important in disease and treatment is sex differences. In terms of understanding where sex differences arise from in health, disease or treatment is a major challenge. It is particularly challenging given that there are three major sex biasing factors: sex hormones, gonads and sex chromosomes. In human populations it is not possible to tease apart their individual contributions and specific interactions, therefore using a unique mouse model the FCG with GDX and subsequent hormone replacement, becomes critical to uncover their relative role in gene regulation and disease associations. Thus, the analysis we conducted in Chapter 3 now provides a rich resource to complement any study of disease, where sex differences are known to play a role. For example, as highlighted in the previous Chapter 5 there are known sex differences in disease progression and outcome in NASH. We were able to use the FCG data, to help show where the sex differences in disease development may be attributed through in terms of hormones, gonads or chromosomes. In the future, it will be important to conduct a similar sex difference analysis across a broader range of tissues to get a better idea of the effects across the whole system and even more importantly to do this at a higher resolution i.e. at the single cell level to understand the

interactions that are occurring within a given tissue, which will help us to be more pinpointed with future therapies.

The idea of being more granular is not just limited to the studies of sex differences as mentioned above but also to move in general away from bulk tissue data, where signals typically are dominated by the most abundant cell type. We are now moving closer to having access to a broad array of single cell eQTL datasets, which will help us be more specific when linking causal inference from GWAS datasets through Mergeomics, which we aim to visualize in single cell gene regulatory networks for hub gene prioritization. In addition, it has been generally challenging to incorporate the full array of omics datasets for integration for a given phenotype due to data inaccessibility or imperfect population/phenotype matching. In the near future, with more omics datasets being collected, I would aim to integrate more layers and explore how to best utilize datasets such as microbiome and spatial transcriptomics as options which currently are not including in our integrative pipelines.

Overall, the dissertation work highlights the extreme complexity in MetDs as well as the importance of a holistic approach to unravel mechanistic insight and gene targets for therapy. As more omics resources become available our ability to integrate through novel approaches in a higher resolution manner becomes crucial and will eventually open the door to more precise targeting for disease treatment.

Appendix

References

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. Blencowe, M., et al., *Network Modeling Approaches and Applications to Unravelling Non-Alcoholic Fatty Liver Disease.* Genes (Basel), 2019. **10**(12).
4. Hunter, D.J., *Gene-environment interactions in human diseases.* Nat Rev Genet, 2005. **6**(4): p. 287-98.
5. Dai, X., et al., *Genetics of coronary artery disease and myocardial infarction.* World journal of cardiology, 2016. **8**(1): p. 1-23.
6. Sookoian, S. and C.J. Pirola, *Genetic predisposition in nonalcoholic fatty liver disease.* Clin Mol Hepatol, 2017. **23**(1): p. 1-12.
7. Marbach, D., et al., *Wisdom of crowds for robust gene network inference.* Nat Methods, 2012. **9**(8): p. 796-804.
8. 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.
9. 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.
10. Rhinn, H., et al., *Integrative genomics identifies APOE epsilon4 effectors in Alzheimer's disease.* Nature, 2013. **500**(7460): p. 45-50.

11. 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.

12. 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.

13. 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.

14. 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.

15. 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.

16. 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.

17. 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.

18. 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.

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

20. 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.

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

22. 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.

23. 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.

24. Yang, X., *Multitissue Multiomics Systems Biology to Dissect Complex Diseases.* Trends Mol Med, 2020. **26**(8): p. 718-728.

25. Boyle, E.A., Y.I. Li, and J.K. Pritchard, *An Expanded View of Complex Traits: From Polygenic to Omnigenic.* Cell, 2017. **169**(7): p. 1177-1186.

26. Subramanian, I., et al., *Multi-omics Data Integration, Interpretation, and Its Application.* Bioinformatics and biology insights, 2020. **14**: p. 1177932219899051-1177932219899051.

27. Graw, S., et al., *Multi-omics data integration considerations and study design for biological systems and disease.* Mol Omics, 2020.

28. Huang, S., K. Chaudhary, and L.X. Garmire, *More Is Better: Recent Progress in Multi-Omics Data Integration Methods.* Front Genet, 2017. **8**: p. 84.

29. Shi, Q., et al., *Pattern fusion analysis by adaptive alignment of multiple heterogeneous omics data.* Bioinformatics, 2017. **33**(17): p. 2706-2714.

30. Wang, B., et al., *Similarity network fusion for aggregating data types on a genomic scale.* Nat Methods, 2014. **11**(3): p. 333-7.

31. Yuan, Y., R.S. Savage, and F. Markowetz, *Patient-specific data fusion defines prognostic cancer subtypes.* PLoS Comput Biol, 2011. **7**(10): p. e1002227.

32. Shen, R., et al., *Integrative subtype discovery in glioblastoma using iCluster.* PLoS One, 2012. **7**(4): p. e35236.

33. Lock, E.F. and D.B. Dunson, *Bayesian consensus clustering.* Bioinformatics, 2013. **29**(20): p. 2610-6.

34. de Tayrac, M., et al., *Simultaneous analysis of distinct Omics data sets with integration of biological knowledge: Multiple Factor Analysis approach.* BMC Genomics, 2009. **10**: p. 32.

35. Le Cao, K.A., I. Gonzalez, and S. Dejean, *integrOmics: an R package to unravel relationships between two omics datasets.* Bioinformatics, 2009. **25**(21): p. 2855-6.

36. Rohart, F., et al., *mixOmics: An R package for 'omics feature selection and multiple data integration.* PLoS Comput Biol, 2017. **13**(11): p. e1005752.

37. Vaske, C.J., et al., *Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM.* Bioinformatics, 2010. **26**(12): p. i237-45.

38. Koh, H.W.L., et al., *iOmicsPASS: network-based integration of multiomics data for predictive subnetwork discovery.* NPJ Syst Biol Appl, 2019. **5**: p. 22.

39. Zoppi, J., et al., *MiBiOmics: an interactive web application for multi-omics data exploration and integration.* BMC Bioinformatics, 2021. **22**(1): p. 6.

40. Bonnet, E., L. Calzone, and T. Michoel, *Integrative multi-omics module network inference with Lemon-Tree.* PLoS Comput Biol, 2015. **11**(2): p. e1003983.

41. Hernandez-de-Diego, R., et al., *PaintOmics 3: a web resource for the pathway analysis and visualization of multi-omics data.* Nucleic Acids Res, 2018. **46**(W1): p. W503-W509.

42. Dimitrakopoulos, C., et al., *Network-based integration of multi-omics data for prioritizing cancer genes.* Bioinformatics, 2018. **34**(14): p. 2441-2448.

43. Zhou, Y., et al., *Metascape provides a biologist-oriented resource for the analysis of systems-level datasets.* Nat Commun, 2019. **10**(1): p. 1523.

44. Huang, S., K. Chaudhary, and L.X. Garmire, *More Is Better: Recent Progress in Multi-Omics Data Integration Methods.* Frontiers in Genetics, 2017. **8**(84).

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

46. 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.

47. Chella Krishnan, K., et al., *Integration of Multi-omics Data from Mouse Diversity Panel Highlights Mitochondrial Dysfunction in Non-alcoholic Fatty Liver Disease.* Cell Syst, 2018. **6**(1): p. 103-115 e7.

48. Chen, L., et al., *Integrative genomic analysis identified common regulatory networks underlying the correlation between coronary artery disease and plasma lipid levels.* BMC Cardiovasc Disord, 2019. **19**(1): p. 310.

49. Hartman, R.J.G., et al., *Sex-Stratified Gene Regulatory Networks Reveal Female Key Driver Genes of Atherosclerosis Involved in Smooth Muscle Cell Phenotype Switching.* Circulation, 2021. **143**(7): p. 713-726.

50. Liu, Y., et al., *Spatiotemporal Gene Coexpression and Regulation in Mouse Cardiomyocytes of Early Cardiac Morphogenesis.* J Am Heart Assoc, 2019. **8**(15): p. e012941.

51. 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.

52. Zhao, Y., et al., *Multi-omics integration reveals molecular networks and regulators of psoriasis.* BMC Syst Biol, 2019. **13**(1): p. 8.

53. Jung, S.M., K.S. Park, and K.J. Kim, *Deep phenotyping of synovial molecular signatures by integrative systems analysis in rheumatoid arthritis.* Rheumatology (Oxford), 2020.

54. Drake, J., et al., *Assessing the Role of Long Noncoding RNA in Nucleus Accumbens in Subjects With Alcohol Dependence.* Alcohol Clin Exp Res, 2020. **44**(12): p. 2468-2480.

55. 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.

56. Min, H.K., et al., *Integrated systems analysis of salivary gland transcriptomics reveals key molecular networks in Sjogren's syndrome.* Arthritis Res Ther, 2019. **21**(1): p. 294.

57. Diamante, G., et al., *Systems toxicogenomics of prenatal low-dose BPA exposure on liver metabolic pathways, gut microbiota, and metabolic health in mice.* Environ Int, 2021. **146**: p. 106260.

58. Zhang, G., et al., *Differential metabolic and multi-tissue transcriptomic responses to fructose consumption among genetically diverse mice.* Biochim Biophys Acta Mol Basis Dis, 2020. **1866**(1): p. 165569.

59. Shu, L., et al., *Prenatal Bisphenol A Exposure in Mice Induces Multitissue Multiomics Disruptions Linking to Cardiometabolic Disorders.* Endocrinology, 2019. **160**(2): p. 409-429.

60. Blencowe, M., et al., *Gene networks and pathways for plasma lipid traits via multitissue multiomics systems analysis.* J Lipid Res, 2021. **62**: p. 100019.

61. Zhao, Y., et al., *Integrative Genomics Analysis Unravels Tissue-Specific Pathways, Networks, and Key Regulators of Blood Pressure Regulation.* Front Cardiovasc Med, 2019. **6**: p. 21.

62. Hui, S.T., et al., *The Genetic Architecture of Diet-Induced Hepatic Fibrosis in Mice.* Hepatology, 2018. **68**(6): p. 2182-2196.

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

64. Chen, Y.-W., et al., *PharmOmics: A Species- and Tissue-specific Drug Signature Database and Online Tool for Toxicity Prediction and Drug Repurposing.* bioRxiv, 2019: p. 837773.

65. Lamb, J., *The Connectivity Map: a new tool for biomedical research.* Nat Rev Cancer, 2007. **7**(1): p. 54-60.

66. Subramanian, A., et al., *A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles.* Cell, 2017. **171**(6): p. 1437-1452.e17.

67. Cheng, F., et al., *Network-based approach to prediction and population-based validation of in silico drug repurposing.* Nat Commun, 2018. **9**(1): p. 2691.

68. Consortium, G.T., *The GTEx Consortium atlas of genetic regulatory effects across human tissues.* Science, 2020. **369**(6509): p. 1318-1330.

69. Auton, A., et al., *A global reference for human genetic variation.* Nature, 2015. **526**(7571): p. 68-74.

70. Xu, J., et al., *EWAS: epigenome-wide association study software 2.0.* Bioinformatics, 2018. **34**(15): p. 2657-2658.

71. Consortium, E.P., *The ENCODE (ENCyclopedia Of DNA Elements) Project.* Science, 2004. **306**(5696): p. 636-40.

72. Liberzon, A., et al., *The Molecular Signatures Database (MSigDB) hallmark gene set collection.* Cell Syst, 2015. **1**(6): p. 417-425.

73. Kanehisa, M., et al., *KEGG for linking genomes to life and the environment.* Nucleic Acids Res, 2008. **36**(Database issue): p. D480-4.

74. Fabregat, A., et al., *The Reactome Pathway Knowledgebase.* Nucleic Acids Res, 2018. **46**(D1): p. D649-D655.

75. *BioCarta.* Biotech Software & Internet Report, 2001. **2**(3): p. 117-120.

76. The Gene Ontology, C., *The Gene Ontology Resource: 20 years and still GOing strong.* Nucleic Acids Res, 2019. **47**(D1): p. D330-D338.

77. Slenter, D.N., et al., *WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research.* Nucleic Acids Res, 2018. **46**(D1): p. D661-D667.

78. Huang, R., et al., *The NCATS BioPlanet - An Integrated Platform for Exploring the Universe of Cellular Signaling Pathways for Toxicology, Systems Biology, and Chemical Genomics.* Front Pharmacol, 2019. **10**: p. 445.

79. Song, W.M. and B. Zhang, *Multiscale Embedded Gene Co-expression Network Analysis.* PLoS Comput Biol, 2015. **11**(11): p. e1004574.

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

81. 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.

82. Szklarczyk, D., et al., *The STRING database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets.* Nucleic Acids Res, 2021. **49**(D1): p. D605-d612.

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

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

85. Roberson, E.D., et al., *A subset of methylated CpG sites differentiate psoriatic from normal skin.* J Invest Dermatol, 2012. **132**(3 Pt 1): p. 583-92.

86. Gu, X., et al., *Correlation between Reversal of DNA Methylation and Clinical Symptoms in Psoriatic Epidermis Following Narrow-Band UVB Phototherapy.* J Invest Dermatol, 2015. **135**(8): p. 2077-2083.

87. Buniello, A., et al., *The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019.* Nucleic Acids Res, 2019. **47**(D1): p. D1005-D1012.

88. van der Fits, L., et al., *Imiquimod-induced psoriasis-like skin inflammation in mice is mediated via the IL-23/IL-17 axis.* J Immunol, 2009. **182**(9): p. 5836-45.

89. Menter, A., et al. *A phase 2b trial of baricitinib, an oral JAK inhibitor, in patients with moderate to severe psoriasis*. in *JOURNAL OF THE AMERICAN ACADEMY OF DERMATOLOGY*. 2014. MOSBY-ELSEVIER 360 PARK AVENUE SOUTH, NEW YORK, NY 10010-1710 USA.

90. Martin, G., et al., *Updates on Psoriasis and Cutaneous Oncology: Proceedings from the 2016 MauiDerm Meeting based on presentations by.* J Clin Aesthet Dermatol, 2016. **9**(9 Suppl 1): p. S5-s29.

91. McLaughlin, F. and N.B. La Thangue, *Histone deacetylase inhibitors in psoriasis therapy.* Curr Drug Targets Inflamm Allergy, 2004. **3**(2): p. 213-9.

92. Kwatra, S.G., et al., *JAK inhibitors in psoriasis: a promising new treatment modality.* J Drugs Dermatol, 2012. **11**(8): p. 913-8.

93. Rendon, A. and K. Schäkel, *Psoriasis Pathogenesis and Treatment.* International journal of molecular sciences, 2019. **20**(6): p. 1475.

94. Marioni, R.E., et al., *GWAS on family history of Alzheimer's disease.* Translational Psychiatry, 2018. **8**(1): p. 99.

95. Middeldorp, C.M., et al., *A Genome-Wide Association Meta-Analysis of Attention-Deficit/Hyperactivity Disorder Symptoms in Population-Based Pediatric Cohorts.* J Am Acad Child Adolesc Psychiatry, 2016. **55**(10): p. 896-905 e6.

96. Olfson, E. and L.J. Bierut, *Convergence of genome-wide association and candidate gene studies for alcoholism.* Alcohol Clin Exp Res, 2012. **36**(12): p. 2086-94.

97. Locke, A.E., et al., *Genetic studies of body mass index yield new insights for obesity biology.* Nature, 2015. **518**(7538): p. 197-206.

98. Rashkin, S.R., et al., *Pan-cancer study detects genetic risk variants and shared genetic basis in two large cohorts.* Nat Commun, 2020. **11**(1): p. 4423.

99. Nikpay, M., et al., *A comprehensive 1000 Genomes–based genome-wide association meta-analysis of coronary artery disease.* Nature Genetics, 2015. **47**(10): p. 1121-1130.

100. Manning, A.K., et al., *A genome-wide approach accounting for body mass index identifies genetic variants influencing fasting glycemic traits and insulin resistance.* Nat Genet, 2012. **44**(6): p. 659-69.

101. Shah, S., et al., *Genome-wide association and Mendelian randomisation analysis provide insights into the pathogenesis of heart failure.* Nat Commun, 2020. **11**(1): p. 163.

102. Willer, C.J., et al., *Discovery and refinement of loci associated with lipid levels.* Nat Genet, 2013. **45**(11): p. 1274-1283.

103. Coleman, J.R.I., et al., *Genome-wide gene-environment analyses of major depressive disorder and reported lifetime traumatic experiences in UK Biobank.* Mol Psychiatry, 2020. **25**(7): p. 1430-1446.

104. Timmers, P.R., et al., *Genomics of 1 million parent lifespans implicates novel pathways and common diseases and distinguishes survival chances.* Elife, 2019. **8**.

105. Blauwendraat, C., et al., *Parkinson's disease age at onset genome-wide association study: Defining heritability, genetic loci, and alpha-synuclein mechanisms.* Mov Disord, 2019. **34**(6): p. 866-875.

106. Nair, R.P., et al., *Genome-wide scan reveals association of psoriasis with IL-23 and NF-kappaB pathways.* Nat Genet, 2009. **41**(2): p. 199-204.

107. Pairo-Castineira, E., et al., *Genetic mechanisms of critical illness in COVID-19.* Nature, 2020.

108. Schizophrenia Working Group of the Psychiatric Genomics, C., *Biological insights from 108 schizophrenia-associated genetic loci.* Nature, 2014. **511**(7510): p. 421-7.

109. Hahn, J., et al., *Genetic loci associated with prevalent and incident myocardial infarction and coronary heart disease in the Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Consortium.* PLoS One, 2020. **15**(11): p. e0230035.

110. Wang, Y.F., et al., *Identification of 38 novel loci for systemic lupus erythematosus and genetic heterogeneity between ancestral groups.* Nat Commun, 2021. **12**(1): p. 772.

111. Fuchsberger, C., et al., *The genetic architecture of type 2 diabetes.* Nature, 2016. **536**(7614): p. 41-47.

112. Kupers, L.K., et al., *Meta-analysis of epigenome-wide association studies in neonates reveals widespread differential DNA methylation associated with birthweight.* Nat Commun, 2019. **10**(1): p. 1893.

113. Sammallahti, S., et al., *Maternal anxiety during pregnancy and newborn epigenome-wide DNA methylation.* Mol Psychiatry, 2021.

114. Rijlaarsdam, J., et al., *Epigenetic profiling of social communication trajectories and co-occurring mental health problems: a prospective, methylome-wide association study.* Dev Psychopathol, 2021: p. 1-10.

115. Boyle, A.P., et al., *Annotation of functional variation in personal genomes using RegulomeDB.* Genome Res, 2012. **22**(9): p. 1790-7.

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

117. Derry, J.M., et al., *Identification of genes and networks driving cardiovascular and metabolic phenotypes in a mouse F2 intercross.* PLoS One, 2010. **5**(12): p. e14319.

118. Wang, S.S., et al., *Identification of pathways for atherosclerosis in mice: integration of quantitative trait locus analysis and global gene expression data.* Circ Res, 2007. **101**(3): p. e11-30.

119. Yang, X., et al., *Tissue-specific expression and regulation of sexually dimorphic genes in mice.* Genome Res, 2006. **16**(8): p. 995-1004.

120. Schadt, E.E., et al., *Mapping the genetic architecture of gene expression in human liver.* PLoS Biol, 2008. **6**(5): p. e107.

121. Austin, M.A., *Plasma Triglyceride and Coronary Heart-Disease.* Arteriosclerosis and Thrombosis, 1991. **11**(1): p. 2-14.

122. Reitz, C., et al., *Relation of plasma lipids to Alzheimer disease and vascular dementia.* Archives of Neurology, 2004. **61**(5): p. 705-714.

123. Di Paolo, G. and T.W. Kim, *Linking lipids to Alzheimer's disease: cholesterol and beyond.* Nature Reviews Neuroscience, 2011. **12**(5): p. 284-296.

124. Muoio, D.M. and C.B. Newgard, *Molecular and metabolic mechanisms of insulin resistance and β-cell failure in type 2 diabetes.* Nature reviews Molecular cell biology, 2008. **9**(3): p. 193.

125. Zhang, F. and G. Du, *Dysregulated lipid metabolism in cancer.* World journal of biological chemistry, 2012. **3**(8): p. 167.

126. Zhang, B.B., G.C. Zhou, and C. Li, *AMPK: An Emerging Drug Target for Diabetes and the Metabolic Syndrome.* Cell Metabolism, 2009. **9**(5): p. 407-416.

127. Libby, P., P.M. Ridker, and G.K. Hansson, *Progress and challenges in translating the biology of atherosclerosis.* Nature, 2011. **473**(7347): p. 317-325.

128. Tabas, I. and C.K. Glass, *Anti-Inflammatory Therapy in Chronic Disease: Challenges and Opportunities.* Science, 2013. **339**(6116): p. 166-172.

129. Heller, D.A., et al., *Genetic and Environmental-Influences on Serum-Lipid Levels in Twins.* New England Journal of Medicine, 1993. **328**(16): p. 1150-1156.

130. Kathiresan, S., et al., *Six new loci associated with blood low-density lipoprotein cholesterol, high-density lipoprotein cholesterol or triglycerides in humans.* Nature Genetics, 2008. **40**(2): p. 189-197.

131. van Dongen, J., et al., *Heritability of metabolic syndrome traits in a large population-based sample.* Journal of Lipid Research, 2013. **54**(10): p. 2914-23.

132. Teslovich, T.M., et al., *Biological, clinical and population relevance of 95 loci for blood lipids.* Nature, 2010. **466**(7307): p. 707-13.

133. Kathiresan, S., et al., *A genome-wide association study for blood lipid phenotypes in the Framingham Heart Study.* Bmc Medical Genetics, 2007. **8**.

134. Willer, C.J., et al., *Newly identified loci that influence lipid concentrations and risk of coronary artery disease.* Nature Genetics, 2008. **40**(2): p. 161-169.

135. Willer, C.J., et al., *Discovery and refinement of loci associated with lipid levels.* Nature Genetics, 2013. **45**(11): p. 1274-83.

136. Klarin, D., et al., *Genetics of blood lipids among~ 300,000 multi-ethnic participants of the Million Veteran Program.* Nature genetics, 2018. **50**(11): p. 1514.

137. Hoffmann, T.J., et al., *A large electronic-health-record-based genome-wide study of serum lipids.* Nature genetics, 2018. **50**(3): p. 401.

138. Lonsdale, J., et al., *The genotype-tissue expression (GTEx) project.* Nature genetics, 2013. **45**(6): p. 580.

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

140. MacNeil, L.T. and A.J.M. Walhout, *Gene regulatory networks and the role of robustness and stochasticity in the control of gene expression.* Genome Research, 2011. **21**(5): p. 645-657.

141. Wang, K., M.Y. Li, and M. Bucan, *Pathway-based approaches for analysis of genomewide association studies.* American Journal of Human Genetics, 2007. **81**(6): p. 1278-1283.

142. Zhong, H., et al., *Liver and Adipose Expression Associated SNPs Are Enriched for Association to Type 2 Diabetes.* Plos Genetics, 2010. **6**(5).

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

144. Baranzini, S.E., et al., *Pathway and network-based analysis of genome-wide association studies in multiple sclerosis.* Human Molecular Genetics, 2009. **18**(11): p. 2078-2090.

145. Jia, P.L., et al., *Common variants conferring risk of schizophrenia: A pathway analysis of GWAS data.* Schizophrenia Research, 2010. **122**(1-3): p. 38-42.

146. Welter, D., et al., *The NHGRI GWAS Catalog, a curated resource of SNP-trait associations.* Nucleic acids research, 2013. **42**(D1): p. D1001-D1006.

147. Zhang, K.L., 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**: p. W90-W95.

148. Goh, K.I., et al., *The human disease network.* Proceedings of the National Academy of Sciences of the United States of America, 2007. **104**(21): p. 8685-8690.

149. Makinen, V.P., et al., *Integrative Genomics Reveals Novel Molecular Pathways and Gene Networks for Coronary Artery Disease.* Plos Genetics, 2014. **10**(7).

150. 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 genetics, 2017. **13**(9): p. e1007040.

151. Zhao, Y., et al., *Integrative Genomics Analysis Unravels Tissue-Specific Pathways, Networks, and Key Regulators of Blood Pressure Regulation.* Frontiers in cardiovascular medicine, 2019. **6**: p. 21.

152. Zhao, Y., et al., *Multi-omics integration reveals molecular networks and regulators of psoriasis.* BMC systems biology, 2019. **13**(1): p. 8.

153. 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. **6**(1): p. 103-115. e7.

154. Hewing, B. and U. Landmesser, *LDL, HDL, VLDL, and CVD prevention: lessons from genetics?* Current cardiology reports, 2015. **17**(7): p. 56.

155. Santos-Gallego, C.G., *HDL: quality or quantity?* Atherosclerosis, 2015. **243**(1): p. 121-123.

156. McGillicuddy, F.C., et al., *Interferon γ attenuates insulin signaling, lipid storage, and differentiation in human adipocytes via activation of the JAK/STAT pathway.* Journal of Biological Chemistry, 2009. **284**(46): p. 31936-31944.

157. Parker, B.L., et al., *An integrative systems genetic analysis of mammalian lipid metabolism.* Nature, 2019. **567**(7747): p. 187.

158. Zizola, C., et al., *Cellular retinol-binding protein type I (CRBP-I) regulates adipogenesis.* Molecular and cellular biology, 2010. **30**(14): p. 3412-3420.

159. Schäffler, A. and J. Schölmerich, *Innate immunity and adipose tissue biology.* Trends in immunology, 2010. **31**(6): p. 228-235.

160. Vance, J.E., *MAM (mitochondria-associated membranes) in mammalian cells: lipids and beyond.* Biochimica et Biophysica Acta (BBA)-Molecular and Cell Biology of Lipids, 2014. **1841**(4): p. 595-609.

161. Liu, J.P., et al., *Cholesterol involvement in the pathogenesis of neurodegenerative diseases.* Molecular and Cellular Neuroscience, 2010. **43**(1): p. 33-42.

162. Jeong, H., et al., *Lethality and centrality in protein networks.* Nature, 2001. **411**(6833): p. 41-42.

163. Zhu, J., et al., *Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks.* Nature genetics, 2008. **40**(7): p. 854-861.

164. Wang, I.M., et al., *Systems analysis of eleven rodent disease models reveals an inflammatome signature and key drivers.* Molecular Systems Biology, 2012. **8**.

165. Kurt, Z., et al., *Tissue-specific pathways and networks underlying sexual dimorphism in non-alcoholic fatty liver disease.* Biology of sex differences, 2018. **9**(1): p. 46.

166. Knox, C., et al., *DrugBank 3.0: a comprehensive resource for 'Omics' research on drugs.* Nucleic Acids Research, 2011. **39**: p. D1035-D1041.

167. Yue, W.W. and U. Oppermann, *High-throughput structural biology of metabolic enzymes and its impact on human diseases.* Journal of Inherited Metabolic Disease, 2011. **34**(3): p. 575-581.

168. Maitland-van der Zee, A.H., et al., *The effect of nine common polymorphisms in coagulation factor genes (F2, F5, F7, F12 and F13) on the effectiveness of statins: the GenHAT study.* Pharmacogenetics and Genomics, 2009. **19**(5): p. 338-344.

169. Ference, B.A., et al., *Effect of Long-Term Exposure to Lower Low-Density Lipoprotein Cholesterol Beginning Early in Life on the Risk of Coronary Heart Disease A Mendelian Randomization Analysis.* Journal of the American College of Cardiology, 2012. **60**(25): p. 2631-2639.

170. Guilherme, A., et al., *Adipocyte dysfunctions linking obesity to insulin resistance and type 2 diabetes.* Nat Rev Mol Cell Biol, 2008. **9**(5): p. 367-77.

171. Voight, B.F., et al., *Plasma HDL cholesterol and risk of myocardial infarction: a mendelian randomisation study.* Lancet, 2012. **380**(9841): p. 572-580.

172. Memisogullari, R. and E. Bakan, *Levels of ceruloplasmin, transferrin, and lipid peroxidation in the serum of patients with Type 2 diabetes mellitus.* Journal of Diabetes and Its Complications, 2004. **18**(4): p. 193-197.

173. Volkmar, M., et al., *DNA methylation profiling identifies epigenetic dysregulation in pancreatic islets from type 2 diabetic patients.* Embo Journal, 2012. **31**(6): p. 1405-1426.

174. Blencowe, M., et al., *Network Modeling Approaches and Applications to Unravelling Non-Alcoholic Fatty Liver Disease.* Genes, 2019. **10**(12): p. 966.

175. Chen, L., et al., *Integrative genomic analysis identified common regulatory networks underlying the correlation between coronary artery disease and plasma lipid levels.* BMC Cardiovascular Disorders, 2019. **19**(1): p. 1-10.

176. Lamina, C., et al., *A genome-wide association meta-analysis on apolipoprotein A-IV concentrations.* Human molecular genetics, 2016. **25**(16): p. 3635-3646.

177. Willer, C.J., et al., *Discovery and refinement of loci associated with lipid levels.* Nature Genetics, 2013. **45**(11): p. 1274-+.

178. Joshi-Tope, G., et al., *Reactome: a knowledgebase of biological pathways.* Nucleic Acids Research, 2005. **33**: p. D428-D432.

179. Ogata, H., et al., *KEGG: Kyoto Encyclopedia of Genes and Genomes.* Nucleic Acids Research, 1999. **27**(1): p. 29-34.

180. TA, H.L.S.P.J.H.R.E.M.J.C.F.M., *Potential etiologic and functional implications of genome-wide association loci for human diseases and traits.* Proc Natl Acad Sci U S A, 2009. **106**(23): p. 9362-7.

181. Emilsson, V., et al., *Genetics of gene expression and its effect on disease.* Nature, 2008. **452**(7186): p. 423-U2.

182. Derry, J.M.J., et al., *Identification of Genes and Networks Driving Cardiovascular and Metabolic Phenotypes in a Mouse F2 Intercross.* Plos One, 2010. **5**(12).

183. Schadt, E.E., et al., *Mapping the genetic architecture of gene expression in human liver.* Plos Biology, 2008. **6**(5): p. 1020-1032.

184. Greenawalt, D.M., et al., *A survey of the genetics of stomach, liver, and adipose gene expression from a morbidly obese cohort.* Genome Research, 2011. **21**(7): p. 1008-1016.

185. Fehrmann, R.S.N., et al., *Trans-eQTLs Reveal That Independent Genetic Variants Associated with a Complex Phenotype Converge on Intermediate Genes, with a Major Role for the HLA.* PLoS genetics, 2011. **7**(8).

186. Wang, S.S., et al., *Identification of pathways for atherosclerosis in mice - Integration of quantitative trait locus analysis and global gene expression data.* Circulation Research, 2007. **101**(3): p. E11-E30.

187. Yang, X., et al., *Tissue-specific expression and regulation of sexually dimorphic genes in mice.* Genome Research, 2006. **16**(8): p. 995-1004.

188. Tu, Z.D., et al., *Integrative Analysis of a Cross-Loci Regulation Network Identifies App as a Gene Regulating Insulin Secretion from Pancreatic Islets.* PLoS genetics, 2012. **8**(12).

189. Nica, A.C., et al., *The Architecture of Gene Regulatory Variation across Multiple Human Tissues: The MuTHER Study.* PLoS genetics, 2011. **7**(2).

190. Romanoski, C.E., et al., *Network for Activation of Human Endothelial Cells by Oxidized Phospholipids.* Circulation Research, 2011. **109**(5): p. E27-U52.

191. Romanoski, C.E., et al., *Network for activation of human endothelial cells by oxidized phospholipids: a critical role of heme oxygenase 1.* Circulation Research, 2011. **109**(5): p. e27-41.

192. Dixon, A.L., et al., *A genome-wide association study of global gene expression.* Nature genetics, 2007. **39**(10): p. 1202-7.

193. Fehrmann, R.S., et al., *Trans-eQTLs reveal that independent genetic variants associated with a complex phenotype converge on intermediate genes, with a major role for the HLA.* Plos Genetics, 2011. **7**(8): p. e1002197.

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

195. Montgomery, S.B., et al., *Transcriptome genetics using second generation sequencing in a Caucasian population.* Nature, 2010. **464**(7289): p. 773-U151.

196. Stranger, B.E., et al., *Patterns of Cis Regulatory Variation in Diverse Human Populations.* PLoS genetics, 2012. **8**(4): p. 272-284.

197. Stranger, B.E., et al., *Population genomics of human gene expression.* Nature genetics, 2007. **39**(10): p. 1217-1224.

198. Dimas, A.S., et al., *Common Regulatory Variation Impacts Gene Expression in a Cell Type-Dependent Manner.* Science, 2009. **325**(5945): p. 1246-1250.

199. Duan, S., et al., *Genetic architecture of transcript-level variation in humans.* American Journal of Human Genetics, 2008. **82**(5): p. 1101-1113.

200. Maher, B., *ENCODE: The human encyclopaedia.* Nature, 2012. **489**(7414): p. 46-8.

201. Shu, L., et al., *Mergeomics: integration of diverse genomics resources to identify pathogenic perturbations to biological systems.* bioRxiv, 2016: p. 036012.

202. 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. **57**(1): p. 289-300.

203. Yang, X., et al., *Systematic genetic and genomic analysis of cytochrome P450 enzyme activities in human liver.* Genome Research, 2010. **20**(8): p. 1020-1036.

204. Ye, J., et al., *Primer-BLAST: a tool to design target-specific primers for polymerase chain reaction.* BMC bioinformatics, 2012. **13**(1): p. 134.

205. Livak, K.J. and T.D. Schmittgen, *Analysis of relative gene expression data using real-time quantitative PCR and the 2(T)(-Delta Delta C) method.* Methods, 2001. **25**(4): p. 402-408.

206. Folch, J., M. Lees, and G.S. Stanley, *A simple method for the isolation and purification of total lipides from animal tissues.* Journal of biological chemistry, 1957. **226**(1): p. 497-509.

207. Diamante, G., et al., *Systems toxicogenomics of prenatal low-dose BPA exposure on liver metabolic pathways, gut microbiota, and metabolic health in mice.* Environment International, 2021. **146**: p. 106260.

208. Jurgens, C.A., et al., *β-Cell Loss and β-Cell Apoptosis in Human Type 2 Diabetes Are Related to Islet Amyloid Deposition.* The American Journal of Pathology, 2011. **178**(6): p. 2632-2640.

209. Hull, R.L., et al., *Islet Amyloid: A Critical Entity in the Pathogenesis of Type 2 Diabetes.* The Journal of Clinical Endocrinology & Metabolism, 2004. **89**(8): p. 3629-3643.

210. Kahn, S.E., S. Andrikopoulos, and C.B. Verchere, *Islet amyloid: a long-recognized but underappreciated pathological feature of type 2 diabetes.* Diabetes, 1999. **48**(2): p. 241-253.

211. Cooper, G.J., et al., *Purification and characterization of a peptide from amyloid-rich pancreases of type 2 diabetic patients.* Proceedings of the National Academy of Sciences, 1987. **84**(23): p. 8628-8632.

212. Janson, J., et al., *The mechanism of islet amyloid polypeptide toxicity is membrane disruption by intermediate-sized toxic amyloid particles.* Diabetes, 1999. **48**(3): p. 491-8.

213. O'Brien, T.D., et al., *Islet amyloid polypeptide: a review of its biology and potential roles in the pathogenesis of diabetes mellitus.* Vet Pathol, 1993. **30**(4): p. 317-32.

214. Halban, P.A., et al., *beta-cell failure in type 2 diabetes: postulated mechanisms and prospects for prevention and treatment.* J Clin Endocrinol Metab, 2014. **99**(6): p. 1983-92.

215. Haffner, S.M., *Epidemiology of type 2 diabetes: risk factors.* Diabetes Care, 1998. **21 Suppl 3**: p. C3-6.

216. Barker, D.J., et al., *Type 2 (non-insulin-dependent) diabetes mellitus, hypertension and hyperlipidaemia (syndrome X): relation to reduced fetal growth.* Diabetologia, 1993. **36**(1): p. 62-7.

217. Sasaki, H., et al., *Associations of birthweight and history of childhood obesity with beta cell mass in Japanese adults.* Diabetologia, 2020.

218. Costes, S., et al., *beta-Cell failure in type 2 diabetes: a case of asking too much of too few?* Diabetes, 2013. **62**(2): p. 327-35.

219. Janson, J., et al., *Spontaneous diabetes mellitus in transgenic mice expressing human islet amyloid polypeptide.* Proc Natl Acad Sci U S A, 1996. **93**(14): p. 7283-8.

220. Vilchez, D., I. Saez, and A. Dillin, *The role of protein clearance mechanisms in organismal ageing and age-related diseases.* Nat Commun, 2014. **5**: p. 5659.

221. Hogan, M.F., et al., *RNA-seq-based identification of Star upregulation by islet amyloid formation.* Protein Eng Des Sel, 2019. **32**(2): p. 67-76.

222. Huang, C.J., et al., *Induction of endoplasmic reticulum stress-induced beta-cell apoptosis and accumulation of polyubiquitinated proteins by human islet amyloid polypeptide.* Am J Physiol Endocrinol Metab, 2007. **293**(6): p. E1656-62.

223. Cahill, K.M., et al., *Improved identification of concordant and discordant gene expression signatures using an updated rank-rank hypergeometric overlap approach.* Sci Rep, 2018. **8**(1): p. 9588.

224. Eguchi, K. and R. Nagai, *Islet inflammation in type 2 diabetes and physiology.* J Clin Invest, 2017. **127**(1): p. 14-23.

225. Hess, D.A., et al., *MIST1 Links Secretion and Stress as both Target and Regulator of the Unfolded Protein Response.* Mol Cell Biol, 2016. **36**(23): p. 2931-2944.

226. Riopel, M., et al., *Chronic fractalkine administration improves glucose tolerance and pancreatic endocrine function.* J Clin Invest, 2018. **128**(4): p. 1458-1470.

227. Keefe, M.D., et al., *beta-catenin is selectively required for the expansion and regeneration of mature pancreatic acinar cells in mice.* Dis Model Mech, 2012. **5**(4): p. 503-14.

228. Tschen, S.I., et al., *Age-dependent decline in beta-cell proliferation restricts the capacity of beta-cell regeneration in mice.* Diabetes, 2009. **58**(6): p. 1312-20.

229. Young, A., *Inhibition of insulin secretion.* Adv Pharmacol, 2005. **52**: p. 173-92.

230. Ferreira, A., *Calpain dysregulation in Alzheimer's disease.* ISRN Biochem, 2012. **2012**: p. 728571.

231. Gurlo, T., et al., *beta Cell-specific increased expression of calpastatin prevents diabetes induced by islet amyloid polypeptide toxicity.* JCI Insight, 2016. **1**(18): p. e89590.

232. Ji, J., L. Su, and Z. Liu, *Critical role of calpain in inflammation.* Biomed Rep, 2016. **5**(6): p. 647-652.

233. Eldor, R., et al., *Conditional and specific NF-kappaB blockade protects pancreatic beta cells from diabetogenic agents.* Proc Natl Acad Sci U S A, 2006. **103**(13): p. 5072-7.

234. Jones, S.V. and I. Kounatidis, *Nuclear Factor-Kappa B and Alzheimer Disease, Unifying Genetic and Environmental Risk Factors from Cell to Humans.* Front Immunol, 2017. **8**: p. 1805.

235. Lilienbaum, A. and A. Israel, *From calcium to NF-kappa B signaling pathways in neurons.* Mol Cell Biol, 2003. **23**(8): p. 2680-98.

236. Reichenbach, N., et al., *Inhibition of Stat3-mediated astrogliosis ameliorates pathology in an Alzheimer's disease model.* EMBO Mol Med, 2019. **11**(2).

237. Grivennikov, S.I. and M. Karin, *Dangerous liaisons: STAT3 and NF-kappaB collaboration and crosstalk in cancer.* Cytokine Growth Factor Rev, 2010. **21**(1): p. 11-9.

238. Zhou, Z., et al., *Estrogen receptor alpha protects pancreatic beta-cells from apoptosis by preserving mitochondrial function and suppressing endoplasmic reticulum stress.* J Biol Chem, 2018. **293**(13): p. 4735-4751.

239. Kato, N., *Insights into the genetic basis of type 2 diabetes.* Journal of diabetes investigation, 2013. **4**(3): p. 233-244.

240. Rege, N.K., et al., *Evolution of insulin at the edge of foldability and its medical implications.* Proc Natl Acad Sci U S A, 2020. **117**(47): p. 29618-29628.

241. Prentki, M., F.M. Matschinsky, and S.R. Madiraju, *Metabolic signaling in fuel-induced insulin secretion.* Cell Metab, 2013. **18**(2): p. 162-85.

242. Ebrahimi, A.G., et al., *Beta cell identity changes with mild hyperglycemia: Implications for function, growth, and vulnerability.* Mol Metab, 2020. **35**: p. 100959.

243. Marselli, L., et al., *Persistent or Transient Human beta Cell Dysfunction Induced by Metabolic Stress: Specific Signatures and Shared Gene Expression with Type 2 Diabetes.* Cell Rep, 2020. **33**(9): p. 108466.

244. Rivera, J.F., et al., *Autophagy defends pancreatic beta cells from human islet amyloid polypeptide-induced toxicity.* J Clin Invest, 2014. **124**(8): p. 3489-500.

245. Dobin, A., et al., *STAR: ultrafast universal RNA-seq aligner.* Bioinformatics, 2013. **29**(1): p. 15-21.

246. Anders, S., P.T. Pyl, and W. Huber, *HTSeq--a Python framework to work with high-throughput sequencing data.* Bioinformatics, 2015. **31**(2): p. 166-9.

247. Fadista, J., et al., *Global genomic and transcriptomic analysis of human pancreatic islets reveals novel genes influencing glucose metabolism.* Proc Natl Acad Sci U S A, 2014. **111**(38): p. 13924-9.

248. Law, C.W., et al., *voom: Precision weights unlock linear model analysis tools for RNA-seq read counts.* Genome Biol, 2014. **15**(2): p. R29.

249. Ogata, H., et al., *Computation with the KEGG pathway database.* Biosystems, 1998. **47**(1-2): p. 119-28.

250. Langfelder, P. and S. Horvath, *Eigengene networks for studying the relationships between co-expression modules.* BMC Syst Biol, 2007. **1**: p. 54.

251. Franzen, O., L.M. Gan, and J.L.M. Bjorkegren, *PanglaoDB: a web server for exploration of mouse and human single-cell RNA sequencing data.* Database (Oxford), 2019. **2019**.

252. Shen, L., *GeneOverlap: An R package to test and visualize gene overlaps.* R Package, 2014.

253. Newman, A.M., et al., *Determining cell type abundance and expression from bulk tissues with digital cytometry.* Nature Biotechnology, 2019. **37**(7): p. 773-782.

254. Chen, E.Y., et al., *Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool.* BMC Bioinformatics, 2013. **14**: p. 128.

255. Lonardo, A., et al., *Sex differences in nonalcoholic fatty liver disease: state of the art and identification of research gaps.* Hepatology, 2019. **70**(4): p. 1457-1469.

256. Hui, S.T., et al., *The genetic architecture of diet-induced hepatic fibrosis in mice.* Hepatology, 2018. **68**(6): p. 2182-2196.

257. 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.

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

259. Emilsson, V., et al., *Genetics of gene expression and its effect on disease.* Nature, 2008. **452**(7186): p. 423-428.

260. Schadt, E., et al., *Mapping the Genetic Architecture of Gene Expression in Human Liver.* PLoS Biology, 2008. **6**.

261. Zhong, H., et al., *Liver and adipose expression associated SNPs are enriched for association to type 2 diabetes.* PLoS Genet, 2010. **6**(5): p. e1000932.

262. Mardinoglu, A., et al., *Genome-scale metabolic modelling of hepatocytes reveals serine deficiency in patients with non-alcoholic fatty liver disease.* Nature Communications, 2014. **5**(1): p. 3083.

263. Yilmaz, Y., et al., *Hepatic expression and serum levels of syndecan 1 (CD138) in patients with nonalcoholic fatty liver disease.* Scand J Gastroenterol, 2012. **47**(12): p. 1488-93.

264. Zhao, R.R., et al., *Targeting Chondroitin Sulfate Glycosaminoglycans to Treat Cardiac Fibrosis in Pathological Remodeling.* Circulation, 2018. **137**(23): p. 2497-2513.

265. Povero, D., et al., *Lipid-induced toxicity stimulates hepatocytes to release angiogenic microparticles that require Vanin-1 for uptake by endothelial cells.* Sci Signal, 2013. **6**(296): p. ra88.

266. Fan, N., et al., *Rapid Two-Photon Fluorescence Imaging of Monoamine Oxidase B for Diagnosis of Early-Stage Liver Fibrosis in Mice.* Analytical Chemistry, 2021. **93**(18): p. 7110-7117.

267. Kaludercic, N., et al., *Monoamine oxidase B prompts mitochondrial and cardiac dysfunction in pressure overloaded hearts.* Antioxid Redox Signal, 2014. **20**(2): p. 267-80.

268. Clayton, J.A. and F.S. Collins, *Policy: NIH to balance sex in cell and animal studies.* Nature News, 2014. **509**(7500): p. 282.

269. Arnold, A.P., *Promoting the understanding of sex differences to enhance equity and excellence in biomedical science*. 2010, BioMed Central.

270. Rask-Andersen, M., et al., *Genome-wide association study of body fat distribution identifies adiposity loci and sex-specific genetic effects.* Nature communications, 2019. **10**(1): p. 339-339.

271. Arnold, A.P., *The organizational–activational hypothesis as the foundation for a unified theory of sexual differentiation of all mammalian tissues.* Hormones and behavior, 2009. **55**(5): p. 570-578.

272. Arnold, A.P., *The end of gonad-centric sex determination in mammals.* Trends in genetics, 2012. **28**(2): p. 55-61.

273. Arnold, A.P., et al., *Cell-autonomous sex determination outside of the gonad.* Developmental Dynamics, 2013. **242**(4): p. 371-379.

274. Schaafsma, S.M. and D.W. Pfaff, *Etiologies underlying sex differences in autism spectrum disorders.* Frontiers in neuroendocrinology, 2014. **35**(3): p. 255-271.

275. Arnold, A.P., *A general theory of sexual differentiation.* Journal of neuroscience research, 2017. **95**(1-2): p. 291-300.

276. Arnold, A.P., *Rethinking sex determination of non-gonadal tissues.* Curr Top Dev Biol, 2019. **134**: p. 289-315.

277. Link, J.C., et al., *X chromosome dosage of histone demethylase KDM5C determines sex differences in adiposity.* J Clin Invest, 2020. **130**(11): p. 5688-5702.

278. Chen, X., et al., *The number of x chromosomes causes sex differences in adiposity in mice.* PLoS Genet, 2012. **8**(5): p. e1002709.

279. Link, J.C., et al., *Increased High-Density Lipoprotein Cholesterol Levels in Mice With XX Versus XY Sex Chromosomes.* Arteriosclerosis, Thrombosis, and Vascular Biology, 2015. **35**(8): p. 1778-1786.

280. Mode, A. and J.A. Gustafsson, *Sex and the liver - a journey through five decades.* Drug Metab Rev, 2006. **38**(1-2): p. 197-207.

281. Waxman, D.J. and M.G. Holloway, *Sex differences in the expression of hepatic drug metabolizing enzymes.* Mol Pharmacol, 2009. **76**(2): p. 215-28.

282. Sugathan, A. and D.J. Waxman, *Genome-wide analysis of chromatin states reveals distinct mechanisms of sex-dependent gene regulation in male and female mouse liver.* Mol Cell Biol, 2013. **33**(18): p. 3594-610.

283. Zheng, D., et al., *Genomics of sex hormone receptor signaling in hepatic sexual dimorphism.* Mol Cell Endocrinol, 2018. **471**: p. 33-41.

284. van Nas, A., et al., *Elucidating the role of gonadal hormones in sexually dimorphic gene coexpression networks.* Endocrinology, 2009. **150**(3): p. 1235-49.

285. De Vries, G.J., et al., *A model system for study of sex chromosome effects on sexually dimorphic neural and behavioral traits.* Journal of Neuroscience, 2002. **22**(20): p. 9005-9014.

286. Burgoyne, P.S. and A.P. Arnold, *A primer on the use of mouse models for identifying direct sex chromosome effects that cause sex differences in non-gonadal tissues.* Biology of sex differences, 2016. **7**(1): p. 68.

287. Berletch, J.B., et al., *Escape from X inactivation varies in mouse tissues.* PLoS Genet, 2015. **11**(3): p. e1005079.

288. Itoh, Y., et al., *The X-linked histone demethylase Kdm6a in CD4+ T lymphocytes modulates autoimmunity.* J Clin Invest, 2019. **129**(9): p. 3852-3863.

289. Golden, L.C., et al., *Parent-of-origin differences in DNA methylation of X chromosome genes in T lymphocytes.* Proceedings of the National Academy of Sciences, 2019. **116**(52): p. 26779-26787.

290. Oliva, M., et al., *The impact of sex on gene expression across human tissues.* Science, 2020. **369**(6509): p. eaba3066.

291. Anderson, S.T. and G.A. FitzGerald, *Sexual dimorphism in body clocks.* Science, 2020. **369**(6508): p. 1164-1165.

292. Anderson, W.D., et al., *Sex differences in human adipose tissue gene expression and genetic regulation involve adipogenesis.* Genome Research, 2020. **30**(10): p. 1379-1392.

293. Matthews, B.J. and D.J. Waxman, *Impact of 3D genome organization, guided by cohesin and CTCF looping, on sex-biased chromatin interactions and gene expression in mouse liver.* Epigenetics & Chromatin, 2020. **13**(1): p. 30.

294. Adams, J.M., et al., *Somatostatin is essential for the sexual dimorphism of GH secretion, corticosteroid-binding globulin production, and corticosterone levels in mice.* Endocrinology, 2015. **156**(3): p. 1052-1065.

295. Kuroda, M., et al., *Interferon regulatory factor 7 mediates obesity-associated MCP-1 transcription.* Plos one, 2020. **15**(5): p. e0233390.

296. Link, J.C. and K. Reue, *Genetic Basis for Sex Differences in Obesity and Lipid Metabolism.* Annu Rev Nutr, 2017. **37**: p. 225-245.

297. Voskuhl, R.R. and S.M. Gold, *Sex-related factors in multiple sclerosis susceptibility and progression.* Nat Rev Neurol, 2012. **8**(5): p. 255-63.

298. Norheim, F., et al., *Gene-by-Sex Interactions in Mitochondrial Functions and Cardio-Metabolic Traits.* Cell Metab, 2019. **29**(4): p. 932-949 e4.

299. Varlamov, O., et al., *Androgen Effects on Adipose Tissue Architecture and Function in Nonhuman Primates.* Endocrinology, 2012. **153**(7): p. 3100-3110.

300. Leung, K.-C., et al., *Estrogen Regulation of Growth Hormone Action.* Endocrine Reviews, 2004. **25**(5): p. 693-721.

301. Shen, M. and H. Shi, *Sex hormones and their receptors regulate liver energy homeostasis.* International journal of endocrinology, 2015. **2015**.

302. Kaneko, S. and X. Li, *X chromosome protects against bladder cancer in females via a KDM6A-dependent epigenetic mechanism.* Sci Adv, 2018. **4**(6): p. eaar5598.

303. Davis, E.J., et al., *A second X chromosome contributes to resilience in a mouse model of Alzheimer's disease.* Sci Transl Med, 2020. **12**(558).

304. Gévry, N., et al., *Histone H2A.Z is essential for estrogen receptor signaling.* Genes & development, 2009. **23**(13): p. 1522-1533.

305. Mauvais-Jarvis, F., et al., *Sex and gender: modifiers of health, disease, and medicine.* The Lancet, 2020. **396**(10250): p. 565-582.

306. Lin, W.-J., H.-M. Hsueh, and J.J. Chen, *Power and sample size estimation in microarray studies.* BMC Bioinformatics, 2010. **11**(1): p. 48.

307. Pawitan, Y., et al., *False discovery rate, sensitivity and sample size for microarray studies.* Bioinformatics, 2005. **21**(13): p. 3017-3024.

308. Team, R.C., *R: A language and environment for statistical computing.* 2020.

309. Song, W.-M. and B. Zhang, *Multiscale embedded gene co-expression network analysis.* PLoS computational biology, 2015. **11**(11): p. e1004574.

310. Langfelder, P. and S. Horvath, *WGCNA: an R package for weighted correlation network analysis.* BMC bioinformatics, 2008. **9**(1): p. 559.

311. Emilsson, V., et al., *Genetics of gene expression and its effect on disease.* Nature, 2008. **452**(7186): p. 423-428.

312. Wang, S.S., et al., *Identification of pathways for atherosclerosis in mice: integration of quantitative trait locus analysis and global gene expression data.* Circulation research, 2007. **101**(3): p. e11-e30.

313. Schadt, E.E., et al., *Mapping the genetic architecture of gene expression in human liver.* PLoS Biol, 2008. **6**(5): p. e107.

314. 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.

315. Shannon, P., et al., *Cytoscape: a software environment for integrated models of biomolecular interaction networks.* Genome research, 2003. **13**(11): p. 2498-2504.

316. Wang, Z., et al., *BART: a transcription factor prediction tool with query gene sets or epigenomic profiles.* Bioinformatics, 2018. **34**(16): p. 2867-2869.

317. MacArthur, J., et al., *The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog).* Nucleic acids research, 2017. **45**(D1): p. D896-D901.

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