INTEGRATIVE OMICS REVEALS INSIGHTS INTO HUMAN LIVER DEVELOPMENT, DISEASE ETIOLOGY, AND PRECISION MEDICINE

by

Zhipeng Liu

A Dissertation

Submitted to the Faculty of Purdue University

In Partial Fulfillment of the Requirements for the degree of

Doctor of Philosophy

Department of Medicinal Chemistry and Molecular Pharmacology

West Lafayette, Indiana

December 2019

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF COMMITTEE APPROVAL

Dr. Wanqing Liu, Co-Chair

Department of Medicinal Chemistry and Molecular Pharmacology

Dr. Tony Hazbun, Co-Chair

Department of Medicinal Chemistry and Molecular Pharmacology

Dr. Pete Pascuzzi

Department of Information and Library Science

Dr. Min Zhang

Department of Statistics

Approved by:

Dr. Andy Hudmon

Dr. Jason Cannon

ACKNOWLEDGMENTS

First and foremost, I would like to thank my advisor Dr. Wanqing Liu for his guidance and support during my whole PhD study, especially the patience and support to allow me to explore the questions that I am truly interested in. Dr. Liu is not only an excellent scientist who always motivates me to think how to make a bigger impact in science, but also an intelligent and trustful friend who guides me to face the challenges in life.

I would also thank my committee for the support from all the aspects. Dr. Tony Hazbun is always supportive and giving me valuable suggestions on improving my project. My very first R programming knowledge was learnt from Dr. Pete Pascuzzi's class, and I am also much appreciated for his help on building up the Shiny application and the NGS support. Last but not least, I would thank Dr. Min Zhang for her guidance in the field of statistical genetics. My initial interest and knowledge about GWAS were gained from Dr. Zhang's class in the first semester of PhD study.

The goal of this project cannot be achieved without help from the collaborators. I thank Dr. Yang Zhang for his excellent work in the animal experiments, Dr. Sarah Graham for her contrition in analyzing the UK biobank data, and Drs. ChienWei (Jack) Chiang and Lang Li in the collaboration of drug metabolism data. I would also thank all the lab mates in Dr. Liu's lab: Dr. Rongrong Wei, Dr. Shaminie Athinarayanan, Dr. Omaima Ali, Mr. Songyao Ma, Dr. Xiaokun Wang, Dr. Chunna Guo, Dr. Samaa Shama, Dr. Defeng Cai, Mr. Zheyun Peng, and Mr. Ze Long. Also, I would particularly thank Dr. Jean-Christophe (Chris) Rochet and his group: Dr. Paola Montenegro, Mr. Sayan Dutta, Ms. Aswathy Chandran, Ms. Chandnee Chandrasekaran, and Ms.

Jennifer Hensel for the great time being together. I also thank my friends in Purdue, especially my roommate Mr. Saeed Akhand for being always supportive.

Lastly, I would like to dedicate this dissertation to my parents: Mr. Xinfeng Liu and Mrs. Qiaomian Duan. I can never achieve thus far without their unconditional love and support.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Transcriptomic regulation of human liver is a tightly controlled and highly dynamic process. Genetic and environmental exposures to this process play pivotal roles in the development of multiple liver disorders. Despite accumulating knowledge have gained through large-scale genomics studies in the developed adult livers, the contributing factors to the interindividual variability in the pediatric livers remain largely uninvestigated. In the first two chapters of the present study, we addressed this question through an integrative analysis of both genetic variations and transcriptome-wide RNA expression profiles in a pediatric human liver cohort with different developmental stages ranging from embryonic to adulthood. Our systematic analysis revealed a transcriptome-wide transition from stem-cell-like to liver-specific profiles during the course of human liver development. Moreover, for the first time, we observed different genetic control of hepatic gene expression in different developmental stages. Motivated by the critical roles of genetics variations and development in regulating hepatic gene expression, we constructed robust predictive models to impute the virtual liver gene expression using easily available genotype and demographic information. Our model is promising in improving both PK/PD modeling and disease diagnosis for pediatric patients.

In the last two chapters of the study, we analyzed the genomics data in a more liver disease-related context. Specifically, in the third chapter, we identified Macrophage migration inhibitory factor (MIF) and its related pathways as potential targets underlying human liver fibrosis through an integrative omics analysis. In the last chapter, utilizing the largest-to-date publicly available GWAS summary data, we dissected the causal relationships among three important and clinically related metabolic diseases: non-alcoholic fatty liver disease (NAFLD), type 2 diabetes

(T2D), and obesity. Our analysis suggested new subtypes and provided insights into the precision treatment or prevention for the three complex diseases.

Taken together, through integrative analysis of multiple levels of genomics information, we improved the current understanding of human liver development, the pathogenesis of liver disorders, and provided implications to precision medicine.

CHAPTER 1. TRANSCRIPTOME-WIDE REGULATION OF HUMAN PEDIATRIC LIVER DEVELOPMENT

1.1 Introduction

The development of human liver is a tightly controlled process during which a variety of hepatic cell types were matured and organized as a specialized functional organ (Gordillo, Evans, & Gouon-Evans, 2015; Ober & Lemaigre, 2018). This process is highly dynamic with numerous molecular and morphological changes (Si-Tayeb, Lemaigre, & Duncan, 2010). A better understanding of this dynamic process would provide insights into liver regeneration and the pathogenesis of liver diseases.

Accumulating evidence has shown that the dysregulation of the precise temporal control of the transcriptome is critical in the pathogenesis of multiple liver disorders (Ober & Lemaigre, 2018). Despite the numerous knowledge about the transcriptome regulation gained from cell culture or animal models (T. T. Li et al., 2009; Schrem, Klempnauer, & Borlak, 2002; Q. Zhou et al., 2017), recent studies have shown that human beings display specific regulatory features (Cardoso-Moreira et al., 2019; Y. Yu et al., 2010). Consequently, it's challenging and sometimes misleading to understand human liver development using model systems. Better approach including using human liver tissues would help solve these issues, however, due to the availability of the human samples, especially in the prenatal stage, our understanding of human liver development is still far from complete.

To address this knowledge gap, we sampled human liver tissues covering different developmental stages from embryonic to adulthood and aimed to explore the transcriptome

dynamics at both gene and isoform level. Next, to understand the cellular composition changes across different developmental stages, we deconvoluted the cell type proportion using the published liver single cell signatures (MacParland et al., 2018). Moreover, to investigate the effects of predisposed genetic risk on transcriptome regulation, especially at the prenatal stage, we calculated the polygenetic scores of 5 important metabolic diseases (Khera et al., 2018; Khera et al., 2019) and identified candidate risk genes and pathways for the corresponding diseases. A workflow summarizing the key question and analyses performed in this chapter was listed in Figure 1.1.

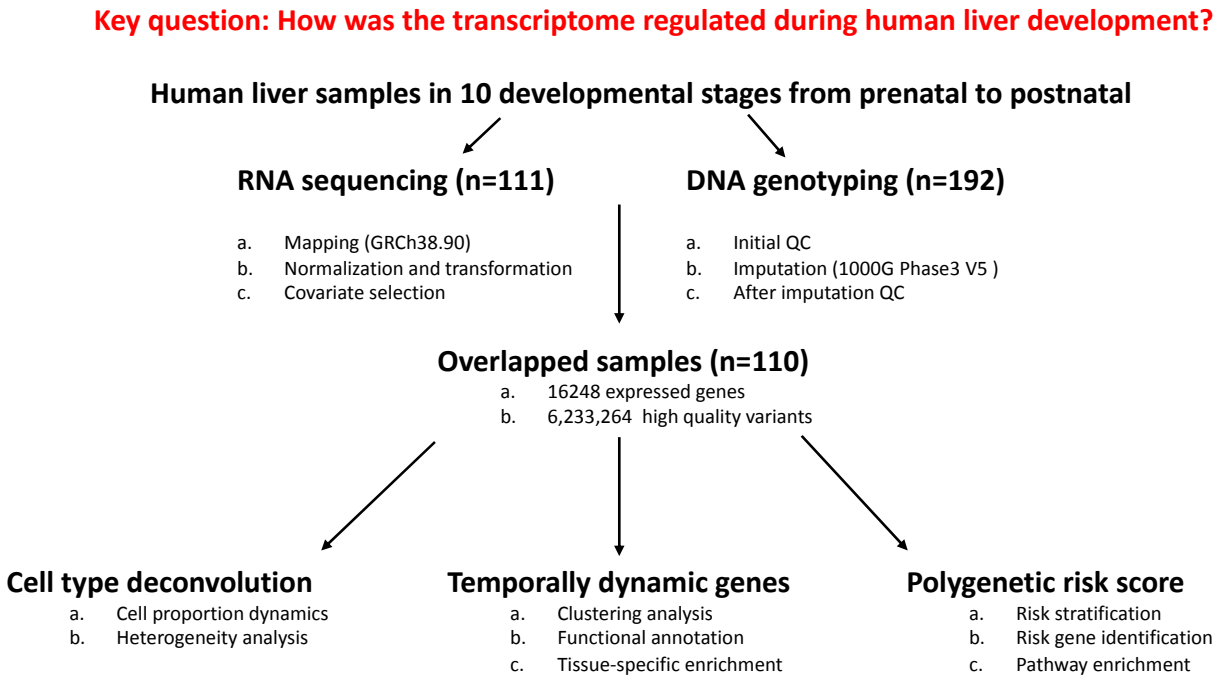


Figure 1.1 Analysis workflow

1.2 Materials and methods

1.2.1 RNA sequencing

The RNA samples were extracted from the liver tissues using Qiagen AllPrep DNA/RNA/miRNAUniversal Kit, followed by DNase I treatment. Libraries were prepared with Illumina TruSeq Stranded mRNA HT (RS-122-21) kit following the manufacturer's instruction. Paired-end sequencing (2 x 76bp) was performed on Illumina HiSeq 4000 platform in the Indiana University genomics core.

1.2.2 Genotyping

DNA samples processing and genotyping were performed on Vanderbilt Technologies for Advanced Genomics (VANTAGE). In brief, 192 liver DNA samples, two experimental replicates, and six unrelated HapMap controls were processed following the Infinium LCG Assay protocol and then genotyped through the Infinium Human MEGA Ex Vanderbilt Beadchips. The reproducibility between duplicates was above 99%. Genome Studio v2.0.2.3 were used to analyze the scanned data and prepare the PLINK files for subsequent analysis.

1.2.3 Imputation and quality control

To maximum the coverage of genotyping and increase the power of the subsequent analyses, we performed genotype imputation through the Michigan Imputation Server (https://imputationserver.sph.umich.edu) (Das et al., 2016). Before the imputation, initial QC was performed by PLINK1.9 to remove low-quality and rare genotyped variants. In specific, variants with genotyping call rate < 0.9, Hardy-Weinberg p value < 1e-3, and minor allele frequency (MAF) < 0.05, individuals with call rate less than 0.9 were filtered out. There are 190 samples and 717,098 variants passed the initial QC and were converted to VCF files for

imputation. Haplotype phasing and genotype imputation were performed using minimac3 (Das et al., 2016) and Eagle v2.3 (Loh, Palamara, & Price, 2016) with the 1000G Phase3 V5 reference panel (Genomes Project et al., 2015). Imputed variants with accuracy (R^2) > 0.8 were kept and went through the post-imputation QC using the same criteria as the initial QC. In total, 190 samples and 6,233,264 variants passed the accuracy and QC threshold. The genomic coordinates of the high-quality variants were lifted from hg19 to hg38 using LiftOver (https://genome.ucsc.edu/cgi-bin/hgLiftOver) for subsequent analysis.

1.2.4 RNA-Seq data processing

The quality of the Fastq files was assessed through FastQC v0.11.8 (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Then low-quality reads were trimmed using Trimmomatic v.0.36 (Bolger, Lohse, & Usadel, 2014) with the parameters: *SLIDINGWINDOW:4:20 MINLEN:25*.

The trimmed reads were aligned to the GRCh38.90 reference genome (Schneider et al., 2017) using STAR v.2.7.0c (Dobin & Gingeras, 2015). To accommodate the requirement of the subsequent isoform level quantification, the bam files were generated in both genomic and transcriptome coordinates. Moreover, to reduce the reference mapping bias for the subsequent allele specific expression (ASE) analysis, WASP (van de Geijn, McVicker, Gilad, & Pritchard, 2015) filtering tags were added based on the genotyping files generated in **section 1.2.3**. More details on ASE analysis are discussed in the later sections. An example mapping parameters were listed as follows:

$STAR --runMode alignReads --runThreadN 5 --genomeDir $GRCh38 --readFilesIn R1_001.merged.fastq.gz R2_001.merged.fastq.gz --readFilesCommand zcat --outFileNamePrefix unique.wasp. --quantMode TranscriptomeSAM GeneCounts --outSAMtype BAM SortedByCoordinate --outSAMstrandField intronMotif --outFilterMultimapNmax 1 --outSAMattributes All --waspOutputMode SAMtag --varVCFfile $vcf

The uniquely mapped reads were quantified for gene- and isoform-level expressions using RSEM v.1.3.1 (B. Li & Dewey, 2011). The alignment quality metrics were generated by picardmetrics (https://github.com/slowkow/picardmetrics) using Picard tools (https://github.com/broadinstitute/picard).

1.2.5 RNA-Seq data normalization

The resulted gene- and isoform-level counts from RSEM were further processed to adjust technical variates. Firstly, we removed gene and transcripts with counts less than 10 in more than 50% of the samples. We also removed mitochondrial, immunoglobulin (IG), and T cell receptor (TR) genes were removed due to their extensive interindividual variability. Secondly, we calculated the normalization factors to correct the GC content and gene/isoform length bias using CQN. Next, DESeq2 (Love, Huber, & Anders, 2014) was used to output the normalized read counts using the normalization factors calculated by CQN (Hansen, Irizarry, & Wu, 2012). Finally, the vst function of DESeq2 was used to stabilize the variance and transform the normalized reads into log2 scale.

1.2.6 Outlier detection and quality control

Firstly, we inspected the potential outliers through PCA plot. The outliers were defined as the samples deviated 3 standard deviations (SD) from the mean of PC1 or PC2. We didn't detect any outliers using this threshold. Secondly, we assessed if there were any mislabeling in terms of the sex of the samples. We examined the sample sex by a scatter plot of XIST gene expression (female only) and the average expression of genes located in the non-pseudoautosomal region (PAR) of Y chromosome (male only). We identified 6 samples with potential mislabeling (Figure 1.2 B). Four of them were prenatal samples and two of them were child samples. We used the sex inferred from the gene expression pattern for the subsequent analysis (Figure 1.2 C, D). Finally, we checked the concordance between DNA and RNA samples using verifyBamID (Jun et al., 2012). We didn't identify any sample swaps.

1.2.7 Covariates selection

To adjust the confounding effects of the technical factors (e.g. sequencing depth, mapping quality) on interindividual variability of human liver expression, we performed a principal component analysis (PCA) using 90 RNA-Seq metrics collected from picardmetrics (https://github.com/slowkow/picardmetrics). The first principal component explained more than 98% of the total variance, indicating the technique factors affect the gene expression in a uniform way. Therefore, the covariates included in the subsequent analyses were the first PC of the RNA-Seq metrics (RNAseq PC1) and other known confounding factors including sex, RNA integrity number (RIN), library batch, and the first three genetic PCs. Age was not adjusted when identifying the temporally dynamic genes but was included as a covariate when evaluating the genetic effects on gene expression.

1.2.8 Identification of the temporally dynamic genes

Based on the reported post conception days (PCD), we characterized the liver samples into 10 developmental stages (Table 1.1). Then DESeq2 was used to identify differentially expressed genes (DEG) between two adjacent stages. Covariates including RNAseq PC1, sex, RIN, library batch, and the first three genetic PCs were adjusted in the differential gene expression analysis. We required the significant DEGs to have at least a 2-fold change and Benjamini-Hochberg FDR adjusted p value less than 0.05. The temporally dynamic genes were defined as the significant DEGs in at least one comparison between two adjacent developmental stages.

1.2.9 K-means clustering of the temporally dynamic genes

To understand the expression pattern of the temporally dynamic genes, we performed k-means clustering analysis to group all the temporally dynamic genes into six clusters. We tested a range of different numbers of clusters from 1 to 30, then the best number of cluster (k=6) was determined as the elbow point of the total within sum of squares plot (**Figure 1.5**). The clusterProfiler package (G. Yu, Wang, Han, & He, 2012) was used to perform KEGG pathway enrichment analysis in each of the clusters.

1.2.10 Functional enrichment of the temporally dynamic genes

To understand the relevance of the temporally dynamic genes to several important biological functions, we tested the enrichment in a list of gene sets including pharmacogenomics (PGx) genes, GWAS related genes, RNA binding proteins (RBP), human transcription factors (TF), and human epigenetic factors. In specific, the list of PGx genes was obtained from a study in which the authors manually curated a list of PGx gene based on their functions in drug metabolism (Wei et al., 2012). GWAS genes were downloaded from the GWAS Catalog v.1.0.1 (Welter et

al., 2014). The list of RBP genes was obtained from a review study listed all the established RBPs (Gerstberger, Hafner, & Tuschl, 2014). The List of human TF was downloaded from the AnimalTFDB v.3.0 database (H. Hu et al., 2019). The list of human epigenetic factors was obtained from the EpiFactors (v.1.7.3) database (Medvedeva et al., 2015). Two-sided Fisher's exact test was used to determine the significance of the enrichment and depletion. The effect size of the enrichment or depletion was calculated as the odds ratio (OR) in log2 scale.

1.2.11 Tissue-specific enrichment of the temporally dynamic genes

To examine the expression pattern of the temporally dynamic genes in different human tissues, we performed tissue-specific enrichment analysis using the TissueEnrich package (Jain & Tuteja, 2019). The tissue-specific genes were defined as genes whose expression levels in a particular tissue are 5-fold higher than all the other tissues. The gene expression data of different human tissues were collected from the Human Protein Atlas (HPA) and the GTEx project.

1.2.12 Characterization of the temporally dynamic long non-coding RNAs (lncRNAs)

As the human reference genome was updated actively, we first checked the most recent genome build (GRCh38.97) to annotate the expressed LncRNAs in the pediatric livers. Then we assessed both *cis* and *trans* regulating effects of the dynamic LncRNA on the candidate target genes respectively.

For the *cis* effects, dynamic LncRNAs were assigned to their closest protein coding genes (target PCG) based on the distance between TSS. Each target PCG was then matched with its closest neighboring protein coding gene (control PCG) to control the shared *cis*-regulatory effect. Each dynamic LncRNA was also paired with a randomly selected protein coding gene (random PCG)

to generate the null distribution of the LncRNA-mRNA correlations. We calculated the Pearson's correlations for LncRNA-target PCG, target PCG-control PCG, and LncRNA-random PCG using all the liver samples (n=110). We identified candidate *cis*-coexpressed protein coding genes if the adjusted p value (LncRNA-target PCG) < 0.05 and p value (LncRNA-target PCG) < p value (target PCG-control PCG). The KEGG pathway enrichment analysis was then performed to understand the functions of the target protein coding genes.

For the *trans* effects, we calculated the Pearson's correlation between each of the dynamic LncRNAs and each of the dynamic protein coding genes. We considered the absolute value of the correlation coefficients larger than 0.9 and adjusted p value less than 0.05 as the candidate target genes. The KEGG pathway enrichment analysis was then performed to understand the functions of the target genes. The correlations between LncRNA and the target genes were visualized using the igraph package (http://igraph.org).

1.2.13 Cell type deconvolution

To explore the cell type composition of liver tissue during development, we decomposed the bulk RNA-Seq data into cell type proportions using the CIBERSORT method (Gentles et al., 2015). The performance of the deconvolution relies on the cell type signature, which are the expression levels of the cell-type specific gene markers. We constructed the cell type signature from a single cell RNA-Seq (scRNAseq) study examining 8,444 cells from five human adult livers (MacParland et al., 2018). The study identified 20 different cell populations, but four of them were contributed by only one donor liver, indicating an individual-specific population or technical issues. Therefore, we only included the other 16 cell populations that were contributed by at least two livers. Then the cell type signature and the bulk RNA-Seq data were imported

into CIBERSORT for cell proportion estimation. We further removed 9 negligible populations (less than 1% in more than 80% of the samples) and re-calculated the relative proportions of the abundant cell types, including hepatocytes I, hepatocytes II, hepatic stellate cells (HSC), erythroid cells, T cells, cholangiocytes, mature B cells, liver sinusoidal endothelial cells (LSEC), non-inflammatory macrophages (Kuffer cells), inflammatory macrophages, and plasma cells. The cell type heterogeneity (entropy) was calculated as $-\sum_{i=1}^{n} \log(p_i) * p_i$, where p_i is the proportion of the i_{th} cell type and n is the total number of cell types.

1.2.14 Genome-wide polygenic score (GPS) calculation and risk stratification

GPS is defined as the weighted sum of the variants contributed to the risk for complex traits/diseases (Chatterjee, Shi, & Garcia-Closas, 2016), such as body mass index (BMI) and Type 2 diabetes (T2D). We calculated the GPS of 5 metabolic diseases, including BMI, T2D, Atrial fibrillation (AF), Coronary artery disease (CAD), and Inflammatory bowel disease (IBD), using the genotype of the liver samples and weights from previous studies (Khera et al., 2018; Khera et al., 2019). The genotype available in the liver set is highly overlapped with the reported weights (mean overlapping rate > 98%). As the ancestry is a major confounding factor of GPS (Figure 1.14), we calculated the corrected GPS as the residuals of regressing the first three genetic PCs on the raw GPS. The corrected GPS didn't have a significant correlation with ancestry any more (Figure 1.15). Then we stratified the individuals based on their corrected GPS of each disease. The individuals with the highest 10% GPS were considered as high risk, while individuals with the lowest 10% GPS were classified as low risk. The rest of the individuals were considered as medium risk.

To examine the effects of genetic predisposition on transcriptomic signatures of the human liver at the early stage of development. For each of the diseases, DESeq2 was used to identify differentially expressed genes in prenatal samples with high (top 10%) and low (tail 10%) risk of disease. Post-conception days, RNAseq PC1, sex, RIN, library batch, and the first three genetic PCs were controlled in the analysis. Genes with FDR adjusted $p < 0.05$ and fold change > 2 were considered as significant. To capture the overall pattern of differentially expressed genes, we adopted a liberal threshold (nominal $p < 0.05$ and fold change > 2) to perform the KEGG pathway enrichment analysis. Pathways with FDR adjusted $p < 0.05$ were considered as significant.

1.3 Results

1.3.1 Characteristics of the human liver samples

A total of 110 human pediatric liver samples passed the stringent quality control of both genotyping and mRNA sequencing (see section 1.2.3 and 1.2.4 for details) and were analyzed in the study. Based on the post-conception days (PCD) of each sample, we characterized the liver samples into 10 developmental stages from embryonic to adulthood (Table 1.1, Figure 1.2 A).

Table 1.1 Developmental stages

Stage	Name	Description	Sample size
Stage 1	Embryonicand Early_fetal	4 < Age <= 10 PCW	12
Stage 2	Early fetal_	10 < Age <= 13 PCW	11
Stage 3	Early mid_fetal	13 < Age <= 16 PCW	12
Stage 4	Mid fetal	16 < Age <= 17 PCW	7
Stage 5	Infancy	0 < Age <= 12M	11
Stage 6	Toddlerhood	1 < Age <= 2Y	12
Stage 7	Early childhood	2 < Age <= 7Y	9
Stage 8	Middle childhood	7 < Age <= 13Y	8
Stage 9	Adolescence	13 < Age <= 20Y	16
Stage 10	Adulthood	Age > 20Y	12

PCW: post-conception weeks.

The sample sizes of adjacent stages are relatively similar (the difference between adjacent stages ranging from 1 to 8). There were no power analyses to determine the same size before sampling. The sample collection was not performed in a randomized and blind way. Due to the availability of the samples, prenatal liver with post-conception weeks less than 4 weeks and greater than 17 weeks were missing in the study. Due to the incomplete demographic information and potential mislabeling, we determined the sex of the samples through a scatter plot of XIST gene expression (female only) and the average expression of genes located in the non-pseudoautosomal region (PAR) of Y chromosome (male only). As shown in Figure 1.2 B, two clear clusters of samples separated by the sex-specific gene expression. We then used the sex inferred from the gene expression pattern for the subsequent analysis (Figure 1.2 C, D). The ancestries of the samples were diverse, in which most of the prenatal samples were black and most postnatal samples were white. Our RNA samples have good integrity (>6 for all the samples, and the mean value is 7.4), ensuring good quality of subsequent RNA sequencing (Figure 1.2 F).

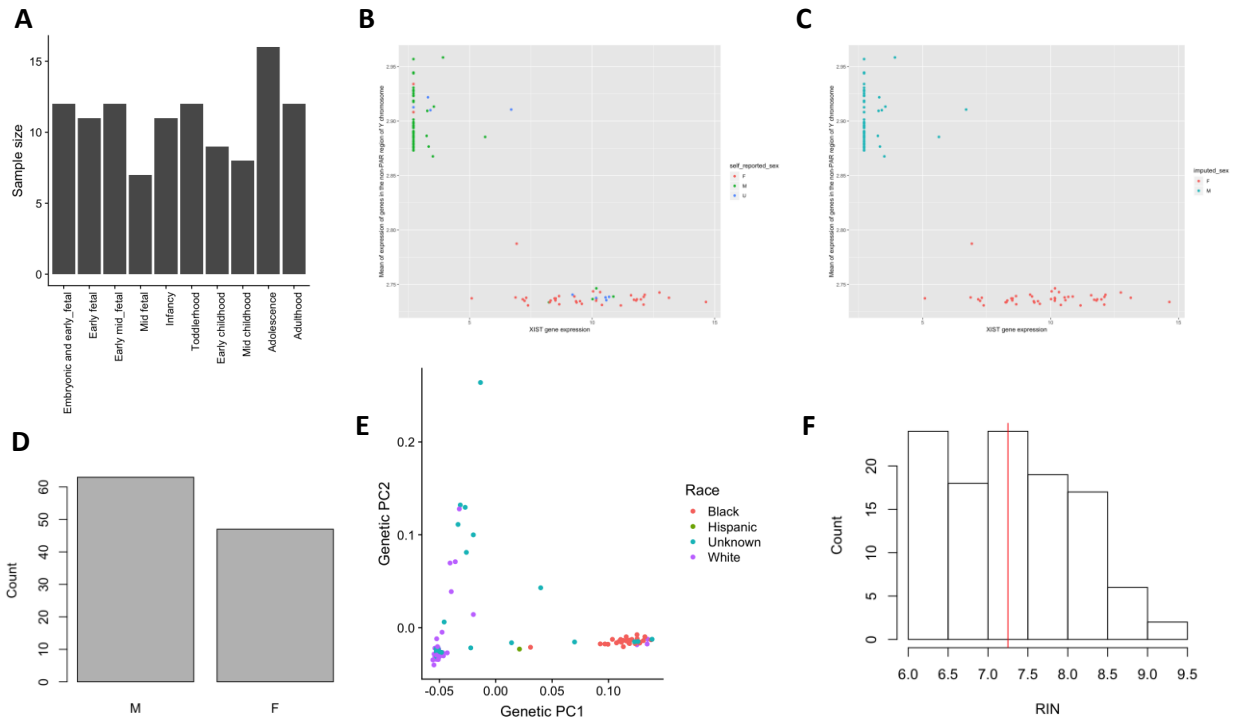


Figure 1.2 Characteristics of the samples

A. The number of sample size of each developmental stage; B. The scatter plot of XIST gene expression (female only) and the average expression of genes located in the non-pseudoautosomal region (PAR) of Y chromosome (male only). The color represents the reported sex; C. The scatter plot of XIST and non-PAR expression. The color represents the sex inferenced from the expression pattern; D. The number of male and female samples in the study; E. PCA plot of the genetics background of the samples; F. The distribution of RIN values of the samples. Red line indicates the mean value.

1.3.2 RNA-Seq quality control metrics and covariates selection

A total of 3.2 billion read pairs (29.3 ± 6.3 million per sample) were uniquely mapped to the human GRCh38.90 reference genome (Figure 1.3 A). The majority of the reads were from mRNA (coding region and UTR regions) (Figure 1.3 B). The examination of the relative converge of the transcript indicated underrepresentation in 5 prime and overrepresentation in 3 prime (Figure 1.3 C), which is not uncommon in mRNA sequencing (Shanker et al., 2015) and will be corrected in the subsequent analyses.

To systematically adjust the confounding effects of technique factors on gene expressions, we collected 90 RNA-Seq quality control metrics and summarized them through a principle component analysis (PCA). The first PC captured 98.8% of the total variance, and samples were not separated into obvious groups by the first two PCs (Figure 1.3 D). Therefore, we used the first PC to represent the technical effects of RNA-Seq.

To further understand the effects of biological and technical confounding factors on gene expression, we examined the pair-wise correlation between the first 10 PCs of gene expression profiling and factors including sex, age (PCD), the first 3 genetics PCs, RIN, library batch, and the first PC of RNA-Seq metrics (Figure 1.3 E). We found that age has the most significant correlation with the first PC of gene expression, suggesting the existence of a broad age-related effect. Other factors including sex, RIN, library batch, RNAseq PC1, and the first three GPCs will be adjusted as covariates.

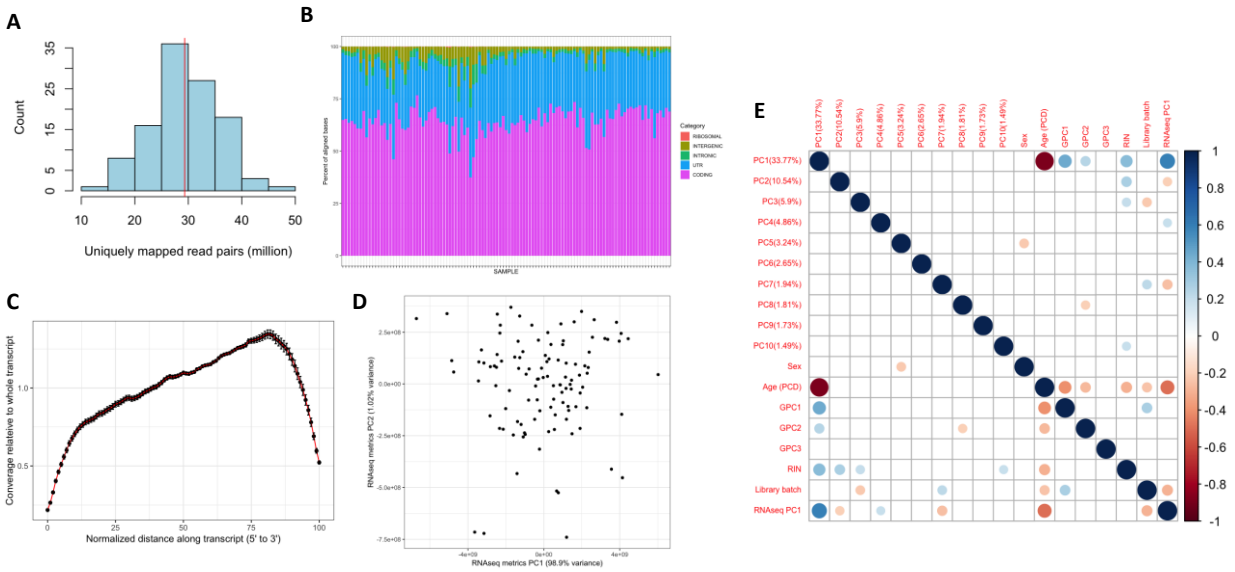


Figure 1.3 Characteristics of the RNA sequencing

A. Distribution of the number of uniquely mapped read pairs (million). Red line indicated the mean value; B. The proportion of the reads mapped to the different regions of the genome. Each bar represents one sample; C. The normalized coverage of the transcript. Black bars represent 95 confidence intervals; D. The PCA plot of the 90 quality control metrics collected during RNA sequencing; E. Correlation matrix of the first 10 PCs of the gene expression and covariates including sex, age (post-conceptions days), the first three genetic PCs, RIN, library batch, and the first PC of RNA-Seq QC metrics. Scale bar represents the Pearson correlation coefficient. Only correlations with $p < 0.05$ were plotted.

1.3.3 Identification and characterization of temporally dynamic genes

A systematic survey of temporally regulated genes would provide insights into human liver development. As shown in Figure 1.3 E and Figure 1.4 A, there are dramatic differences in the gene expression patterns of different developmental stages. To curate a complete list of genes under temporal control, we performed differential gene expression analyses between each pair of adjacent stages. We found 34.9% (5669/16248) of expressed genes were dynamic during development (Figure 1.4 B). The most significant difference happens between stages of mid fetal (16 < age <= 17PCW) and infancy (0 < age <= 12M), which is consistent with findings in other human tissues such as the brain. K-means clustering analysis of the temporally dynamic genes indicated that there are mainly two big clusters: early expressed (highly expressed in prenatal stages) and late expressed (highly expressed in postnatal stages) genes. Based on the gene

expression levels, these two big clusters can be further dissected into six sub-clusters: early low, early medium, early high, late low, late medium, late high (Figure 1.6).

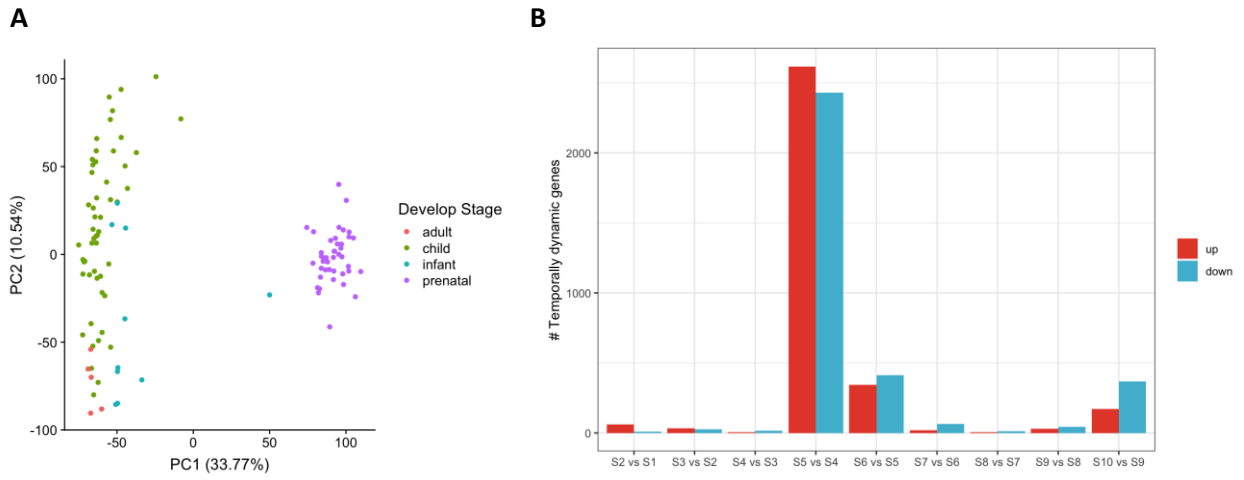


Figure 1.4 Identification of temporally dynamic genes

A. PCA plot the gene expression pattern of liver samples; B. The number of significantly differentially expression genes in each of the comparison between two adjacent stages.

K-means clustering analysis (Figure 1.5) and the KEGG pathway enrichment analysis indicated that pathways related to cell proliferation (e.g. cell cycle, DNA replication) and autoimmune disease (e.g. systemic lupus erythematosus) were overrepresented in the early expressed genes, while pathways involved in liver-specific functions such as biosynthesis, drug metabolism, and fatty acid metabolism were enriched in the late expressed genes (Figure 1.6).

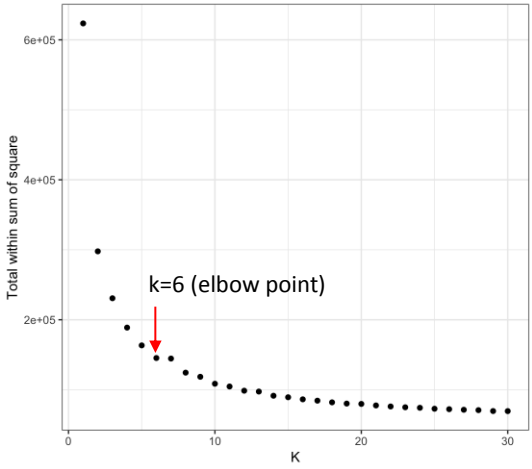


Figure 1.5 K-means clustering analysis

A list of different k from 1 to 30 were tested. The best k = 6 was picked up at the elbow point.

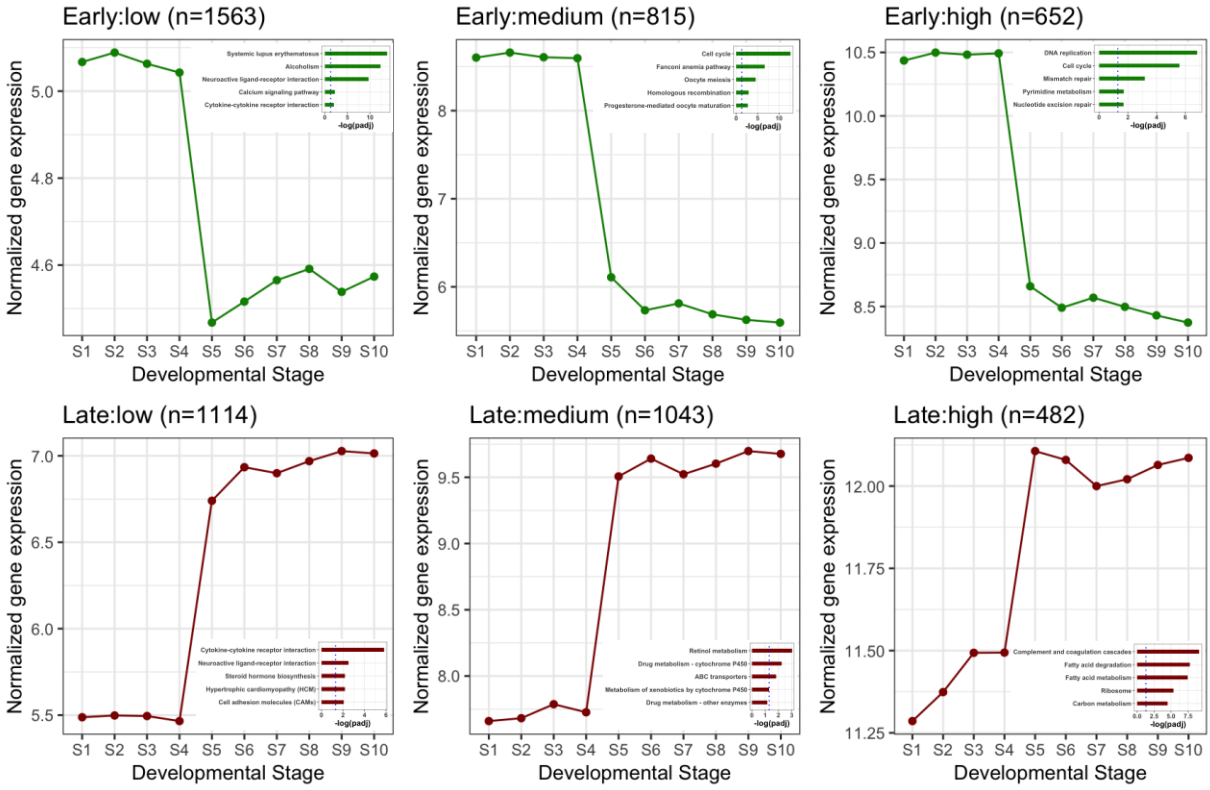


Figure 1.6 K-means clustering analysis of the dynamic genes.

K-means clustering analysis of the identified dynamic genes. The pattern of gene expresion was annotated by the KEGG analysis. Green color represents early expressed genes, while red color represents late expressed genes.

To further interpret the relevance to several important liver functions, we tested the enrichment of the temporally dynamic genes in the gene sets including pharmacogenomics (PGx) genes, GWAS related genes, RNA binding proteins (RBP), and human transcription factors (TF). The overall human TFs were slightly overrepresented in early expressed genes and significantly underrepresented in the late expressed genes (Figure 1.7 A). However, the liver- enriched transcription factors (LETFs) such as HNF4A, FOXA2, HNF1A, CEBPA, and CEBPB were mostly expressed in the postnatal stages (Figure 1.7 B). RBPs were enriched in the late highly expressed genes but underrepresented in the other clusters, suggesting the critical roles of posttranscriptional regulation during the developmental processes and differentiation of tissue identity (Baralle & Giudice, 2017; Brinegar et al., 2017). Consistent with the KEGG analysis, we found liver-specific functional genes (PGx genes) and disease-related genes (GWAS genes) were enriched in the late expressed genes.

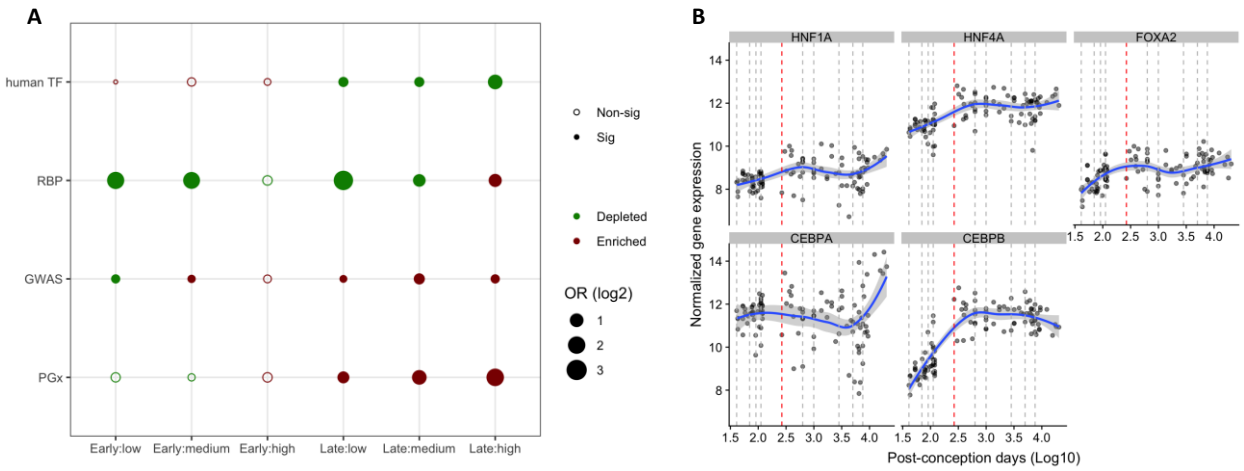


Figure 1.7 Functional annotation of the temporally dynamic genes

A. Enrichment analysis for the 6 clusters of genes in human transcription factors (TF), RNA-binding proteins (RBP), GWAS genes, and PGx genes. Two-sided Fisher exact test was used to determine the significance of the enrichment or depletion. FDR adjusted p value < 0.05 was considered as significant. The extent of enrichment or depletion was expressed as log2 odds ratio; B. Expression trajectory of liver enriched transcription factors. Loess regression was used to fit the trend line. Shades represent 95% CI. The red line indicates the birth day, grey dash lines represent the different developmental stages.

Moreover, we explored the epigenetic regulation of human liver development by testing the enrichment of epigenetic factors (histone and histone modification genes) in each of the dynamic clusters. We found that the epigenetic regulators tend to be up regulated in the prenatal stages (Figure 1.8), suggesting the substantial epigenetic modifications occur in the early development of human livers.

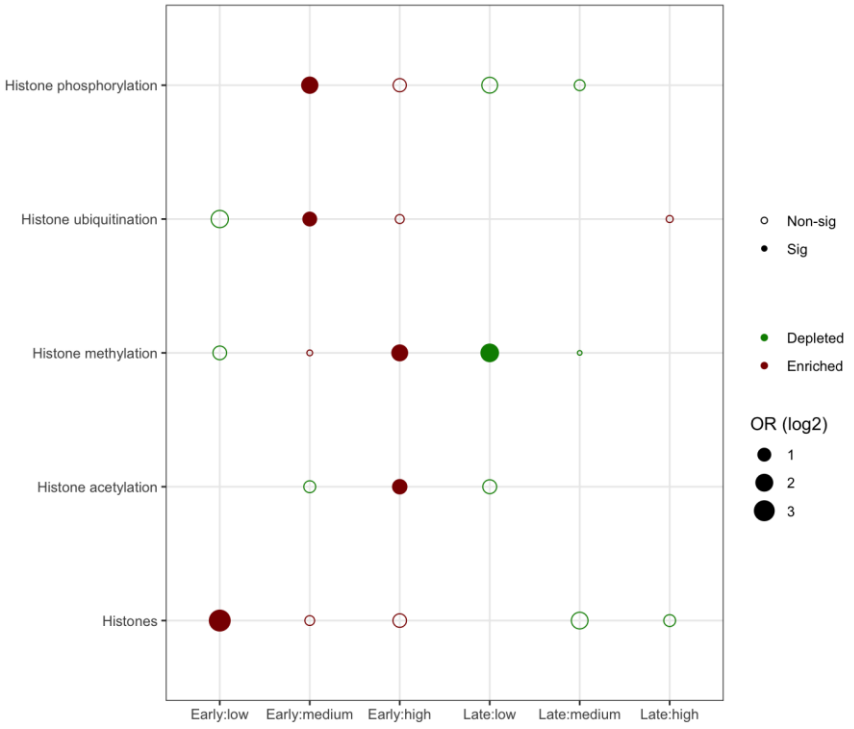


Figure 1.8 Enrichment analysis for the 6 clusters of genes in epigenetic regulators

Two-sided Fisher exact test was used to determine the significance of the enrichment or depletion. FDR adjusted p value < 0.05 was considered as significant. The extent of enrichment or depletion was expressed as log2 odds ratio.

1.3.4 Tissue-specific enrichment indicated a transition from a stem-cell like to liver-specific expression profiling during development

Next, we performed the tissue-specific enrichment analysis to see the expression pattern of the dynamic genes in multiple human tissues. Interestingly, we found the early expressed genes displayed a stem-cell like transcriptomic signature, in which actively proliferating tissues such as bone marrow and lymph node were the most significantly enriched (Figure 1.9 A). The late

expressed genes were predominantly enriched in liver-specific expressed gene set (Figure 1.9 B). The closest tissues to liver are adipose and small intestine, which is consistent with their related similar expression patterns (Consortium et al., 2017). These findings suggested the major function of liver was transited from cell growth in the prenatal stage to liver-specific metabolic processes in the postnatal stage.

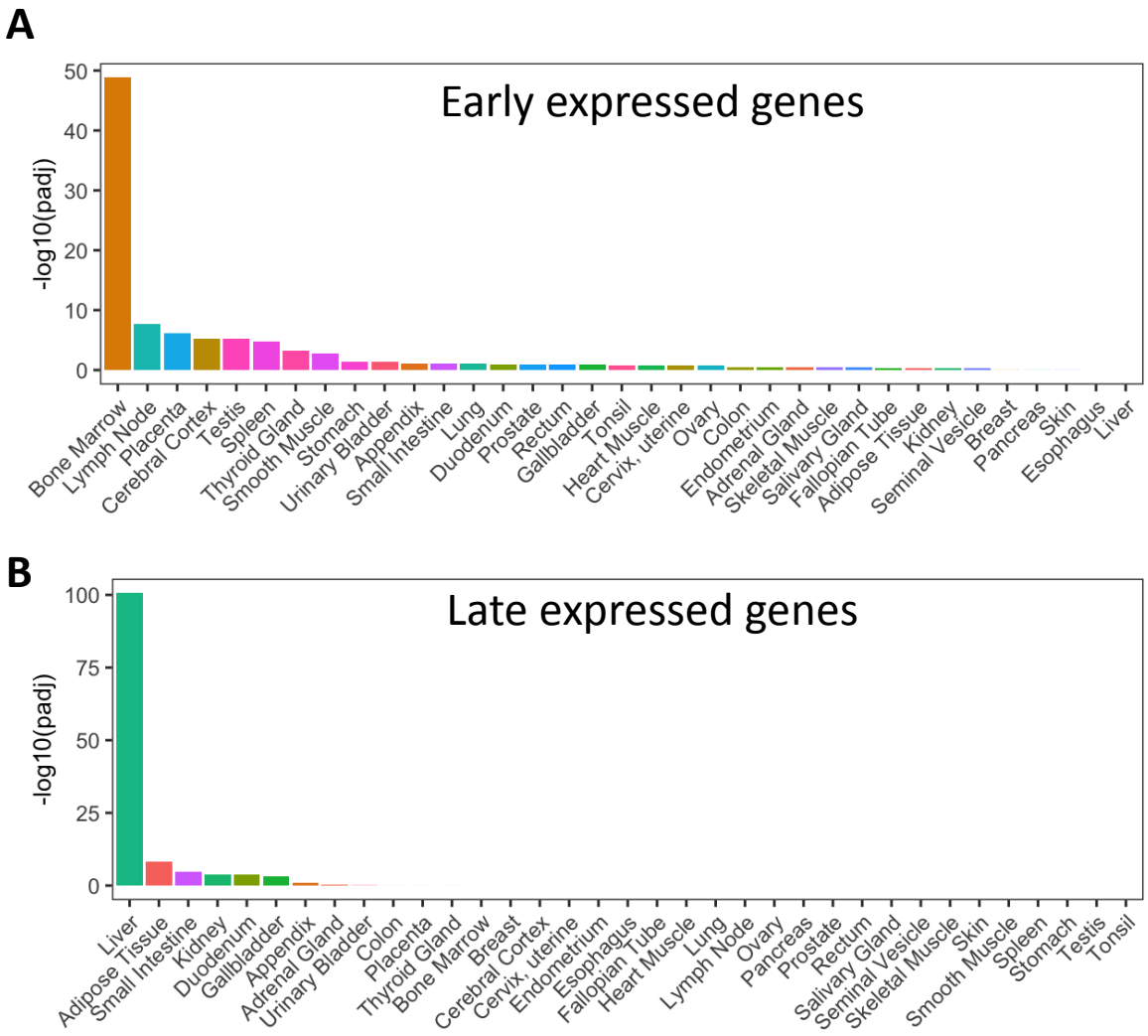


Figure 1.9 Tissue-specific expression analysis

A. Enrichment of the early expressed (low, medium, and high) genes in tissue-specific expression signatures; B. Enrichment of the late expressed (low, medium, and high) genes in tissue-specific expression signatures. Fisher exact test was used to determine significance, followed by FDR multiple testing adjustment.

1.3.5 Temporally dynamic isoforms showed similar patterns as dynamic genes, but with larger effect sizes

We further analyzed the temporally regulation of gene expression at the isoform level. Similar to the gene level, we observed significantly differentially expressed isoforms between stages of prenatal and postnatal (Figure 1.10). The majority (4399 out of 6811) of dynamic isoforms were dynamic at the gene level as well (Figure 1.11 A), but with larger effect sizes (Figure 1.11 B, $p < 2.2e-16$ by Kolmogorov–Smirnov test). Pathway enrichment analysis indicated that the overlapped dynamic genes were mostly involved in steroid hormone synthesis, systemic lupus erythematosus, and drug metabolism (Figure 1.11 C), which are similar to the enriched pathways at the gene level.

We also identified a group of isoform-specific temporally dynamic genes (n=2412). These genes were mainly enriched in pathways regarding mRNA splicing (Figure 1.11 C), indicating the temporal control of alternative splicing in liver was regulated in an isoform-specific way. Consistent with the KEGG pathway analysis, RBPs were significantly overrepresented in the isoform-specific dynamic genes (Figure 1.11 D).

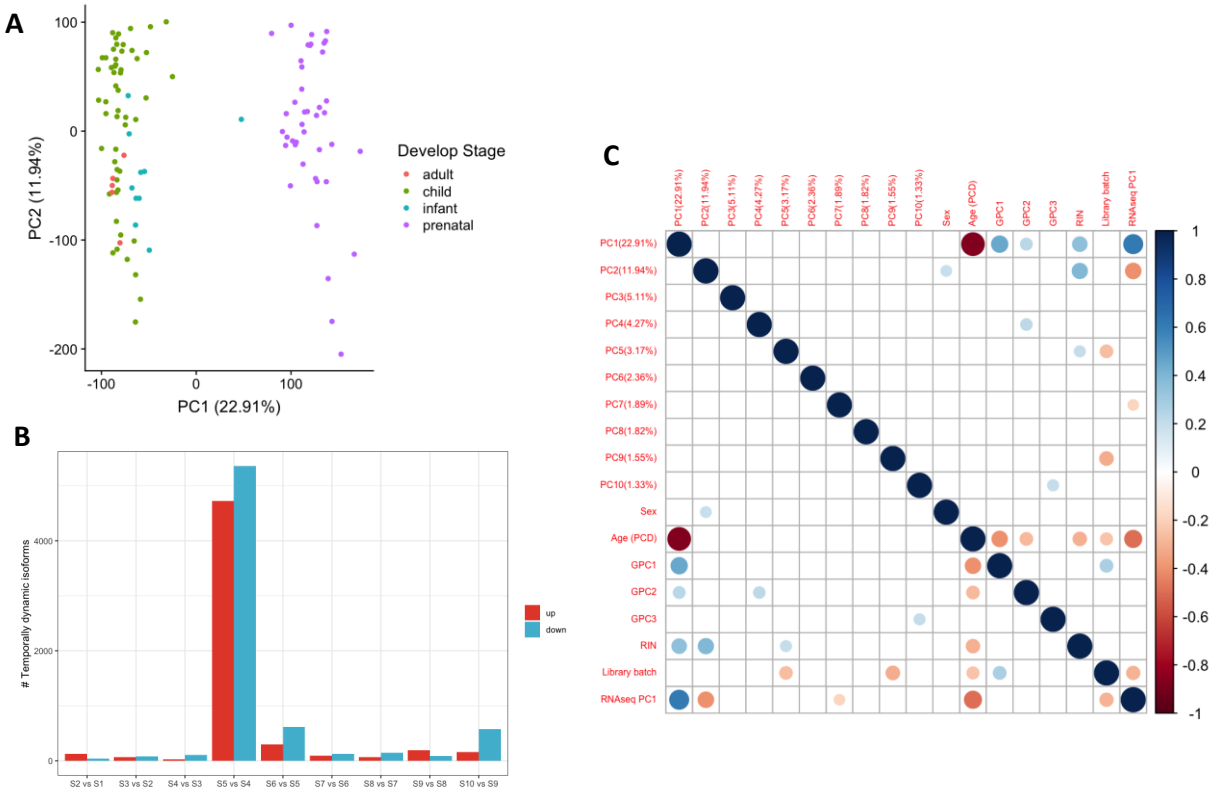


Figure 1.10 Temporally dynamic isoforms

A. PCA plot the isoform expression pattern of liver samples; B. The number of significantly differentially expression isoforms in each of the comparison between two adjacent stages; C. Correlation metrics of the first 10 PCs of isoform expression and the covariates.

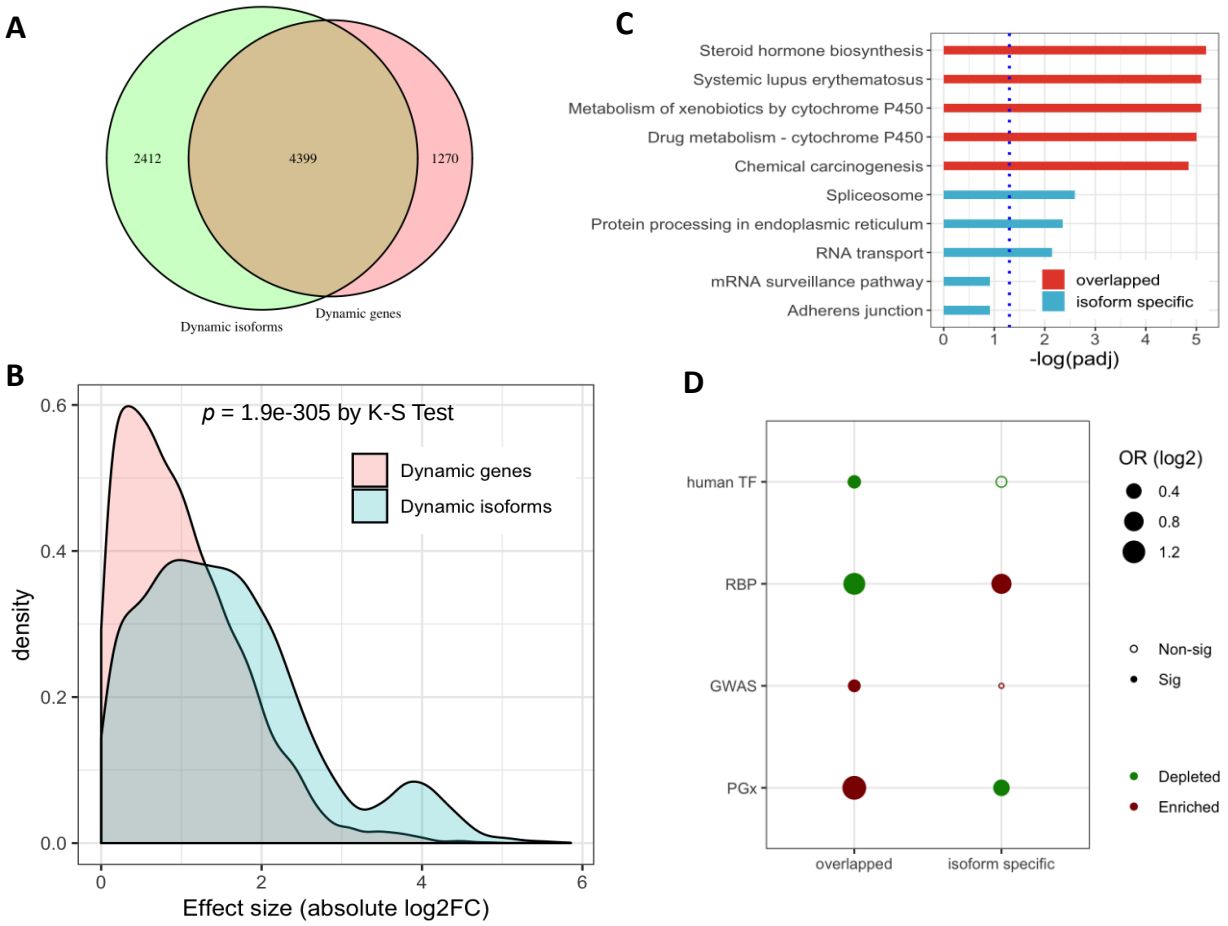


Figure 1.11 Comparison between dynamic genes and dynamic isoforms

A. Venn diagram of dynamic genes and isoforms; B. The distribution densities of the effect size of dynamic genes and isoforms. K-S test was used to determine the significance; C. KEGG pathway enrichment analysis of overlapped genes and isoform specific genes; D. Enrichment analysis of the overlapped and isoform specific genes in functional gene set including human transcription factors (TF), RNA-binding proteins (RBP), GWAS genes, and PGx genes. Two-sided Fisher exact test was used to determine the significance of the enrichment or depletion. FDR adjusted p value < 0.05 was considered as significant. The extent of enrichment or depletion was expressed as log2 odds ratio.

1.3.6 Dynamic LncRNAs exert broad effects on human liver development

LncRNAs, which are long transcripts (more than 200bp) without detectable protein coding potentials, have been identified to be co-expressed with developmental regulators in adult tissues. To assess the contribution of lncRNAs to human liver development, we examined the expressed lncRNAs in the pediatric livers and investigated their cis- and trans-regulating effects on the candidate target genes.

Based on the most recent human reference genome (GRCh38.97), we identified 1412 lncRNAs expressed in the pediatric liver set. Consistent with previous findings, lncRNAs tend to have lower expression levels (Figure 1.12 A) and shorter transcript length (Figure 1.12B) as compared to the protein coding genes (both p values < 2.2e-16 by Mann-Whitney U test). We identified 596 dynamic lncRNAs out of the total 1412 expressed lncRNAs. The dynamic genes were significantly enriched in lncRNAs (OR=1.41, p value = 8.8e-10 by Fisher's exact test). More specifically, dynamic lncRNAs tend to be highly expressed in the early stages of the development (Figure 1.12 C).

To systematically examine the cis-regulating effects of dynamic lncRNAs, we calculated the Pearson's correlation between each of the dynamic lncRNAs and its most adjacent protein coding gene (target PCG). As the correlation between lncRNA and target PCG might be confounded by the shared cis regulatory effects, we further calculated the correlation between target PCG and its immediately neighboring protein coding gene as a control (control PCG). Also, we calculated the correlation between each dynamic lncRNA and a randomly selected protein coding gene (random PCG) to generate the null distribution of the lncRNA-PCG correlations (Figure 1.12 D). As shown in Figure 1.12 E, dynamic lncRNAs displayed strong positive correlations with the target PCGs as compared with the control PCGs (K-S test, p = 6.5e-16), suggesting the existence of the cis-regulating effects of dynamic lncRNAs. We also observed a bias to the positive target PCG-control PCG correlation (K-S test, p = 3.8e-3), indicating the shared cis regulatory effects. Furthermore, we identified 258 candidate target genes with adjusted p value (lncRNA-target PCG) less than 0.05 and p value (lncRNA-target PCG) less than p value (target PCG-control PCG). Pathway analysis of the candidate target

genes indicated strong enrichments in development and metabolism related pathways (Figure 1.12 F).

Moreover, we investigated the trans regulation effects of the dynamic lncRNAs as increasing evidence have shown that lncRNA might regulate remote gene expression through affecting chromatin states, influencing nuclear structure, and interacting with other transcriptional regulators (Kopp & Mendell, 2018). To that end, we calculated the Pearson's correlation between each of the dynamic lncRNAs and each of the dynamic protein coding genes. Correlations with coefficients larger than 0.9 and adjusted p values less than 0.05 were considered as candidate target genes. We identified lncRNA AC073349.1 has the most (n=229) correlated protein coding genes. Pathway analysis indicated that the candidate gene were significantly enriched in porphyrin metabolism (adjusted p = 1.02e-5) and cell cycle (adjusted p = 0.019) pathways. The co-expressed relationships between AC073349.1 and its target genes in porphyrin metabolism and cell cycle pathways were shown in Figure 1.12 G.

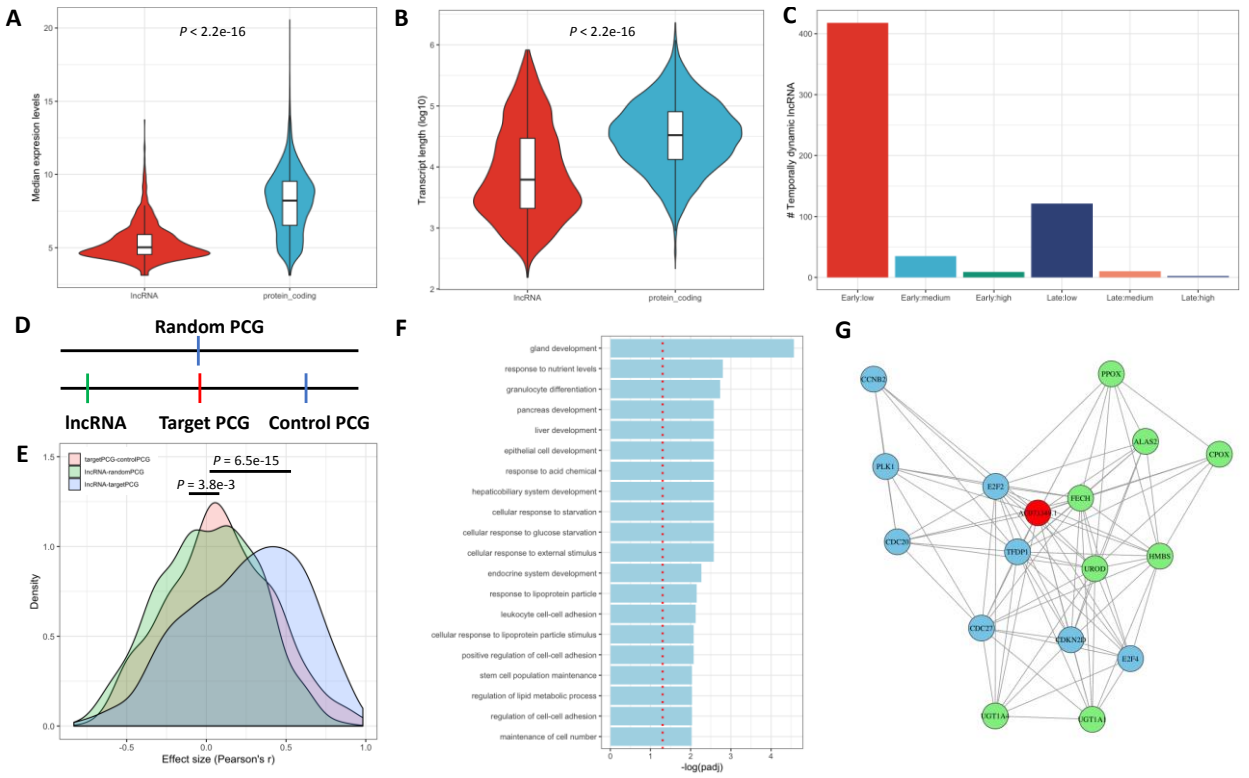


Figure 1.12 Effects of dynamic lncRNAs on transcriptional regulation.

A. Comparison of expression levels between lncRNAs and protein coding genes. Mann–Whitney U test was used to determine the significance; B. Comparison of transcripts length (in log10 scale) between lncRNAs and protein coding genes. Mann–Whitney U test was used to determine the significance; C. The number of dynamic lncRNAs in each of the expression clusters; D. Diagram of the tested correlations. lncRNA: dynamic lncRNA; target PCG : the most closet protein coding gene to the lncRNA; control PCG: the immediate neighboring protein coding gene to the target PCG; random PCG: randomly selected protein coding gene; F. Pathway analysis of the candidate cis-regulating targets genes of the dynamic lncRNAs; G. Correlations between lncRNA AC073349.1and its candidate trans-regulating target genes. Nodes were colored by the gene functions: red: lncRNA, blue: cell cycle pathways; green: porphyrin metabolism pathways.

1.3.7 Cell type heterogeneity decreased during human liver development

To understand the temporal change in cell type composition, we utilized the single-cell data of adult human livers from a published study (MacParland et al., 2018) to deconvolve our bulk RNA-Seq data. After removing negligible cell populations (see section 1.2.12 for details), we estimated the relative proportions of 9 cell types in the liver. Hepatocytes, which are the building blocks of liver and the most abundant cells, increased during embryonic and fetal development (Figure 1.13 A). The change in hepatocytes proportion correlated closely with the expression

level of HNF4A, which is the master TF controlling hepatocytes differentiation (Kamiya, Inoue, & Gonzalez, 2003) (Figure 1.13 C). Consistent with the dynamic gene analysis, the proportions of proliferating active cells such as erythroid cells and T cells were decreased over the development. The composition of other major liver cell types including Kuffer cells and hepatic stellate cells (HSC) seemed to remain stable over time, but this could due to the limited sensitivity of deconvolution for less abundant cells.

Next, we calculated the entropy of cell type proportions to investigate the change in cell type heterogeneity during development. As shown in Figure 1.13 B, the heterogeneity decreased till the end of infancy. The increasing uniformness corroborated the finding of tissue-specific enrichment analysis, in which liver transited from a stem-cell like tissue to a more functional organ during development.

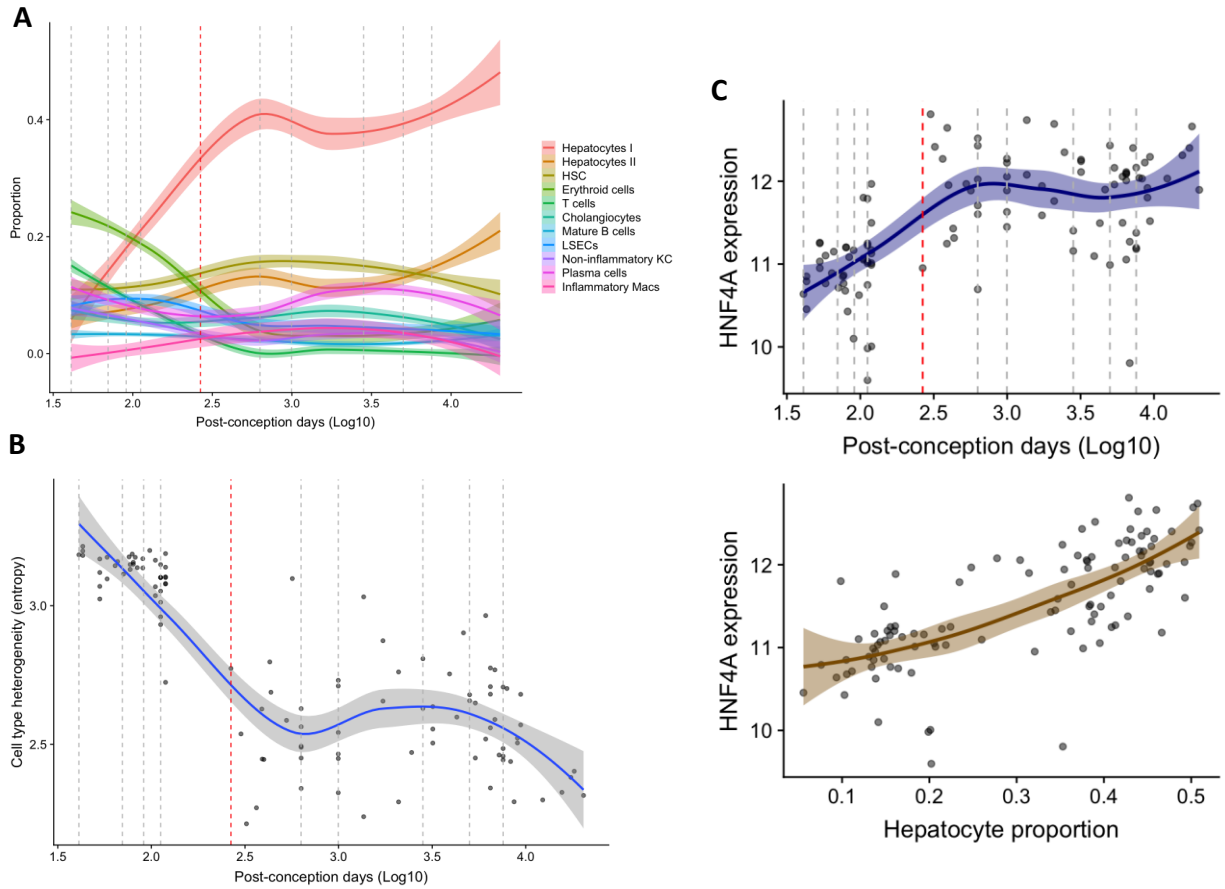


Figure 1.13 Cell type deconvolution

A. Relative cell type proportions over the course of developmental stages; C. Cell type heterogeneity; C. Expression trajectory of HNF4A; D. Correlation between HNF4A expression and hepatocyte proportion. Loess regression was used to fit the trend line. Shades represent 95% CI. The red vertical line indicates the birth day, grey vertical dash lines represent the different developmental stages.

1.3.8 Genome-wide polygenic score (GPS) analysis identified risk genes for metabolic diseases

Recent studies have shown that the risk of complex disease could be better predicted through modeling the genetic variants at the genome-wide level (Richardson, Harrison, Hemani, & Davey Smith, 2019). GPS, which is the weighted sum of multiple variants based on their associations with the disease, has been shown to display similar performance as a monogenetic mutation in predicting complex diseases such as Type 2 diabetes (T2D) (Khera et al., 2018). Therefore, GPS has becoming a powerful approach to stratify the general population into different risk groups. To understand the predisposed genetics effects on disease pathogenesis at

the early stages of development, we compared the transcriptomic profiles of the prenatal livers with high (top 10% GPS score) and low (tail 10% GPS score) risk for 5 metabolic diseases (Figure 1.17).

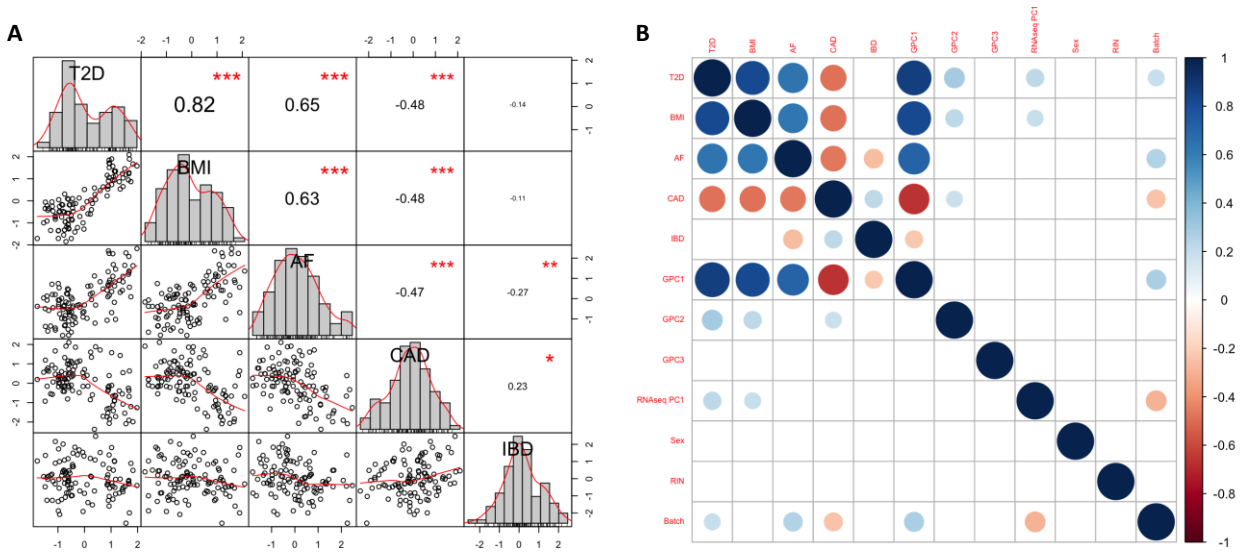


Figure 1.14 Correlation matrix of the raw polygenic scores

A. The distribution and pair-wise correlation of the polygenic scores of 5 metabolic diseases; B. Correlation matrix between the polygenic scores of 5 metabolic diseases and covariates. Scale bar represents the Pearson correlation coefficient, and only correlation with $p < 0.05$ was plotted. T2D: type 2 diabetes, BMI: body mass index, AF: Atrial fibrillation, CAD: Coronary artery disease, IBD: Inflammatory bowel disease.

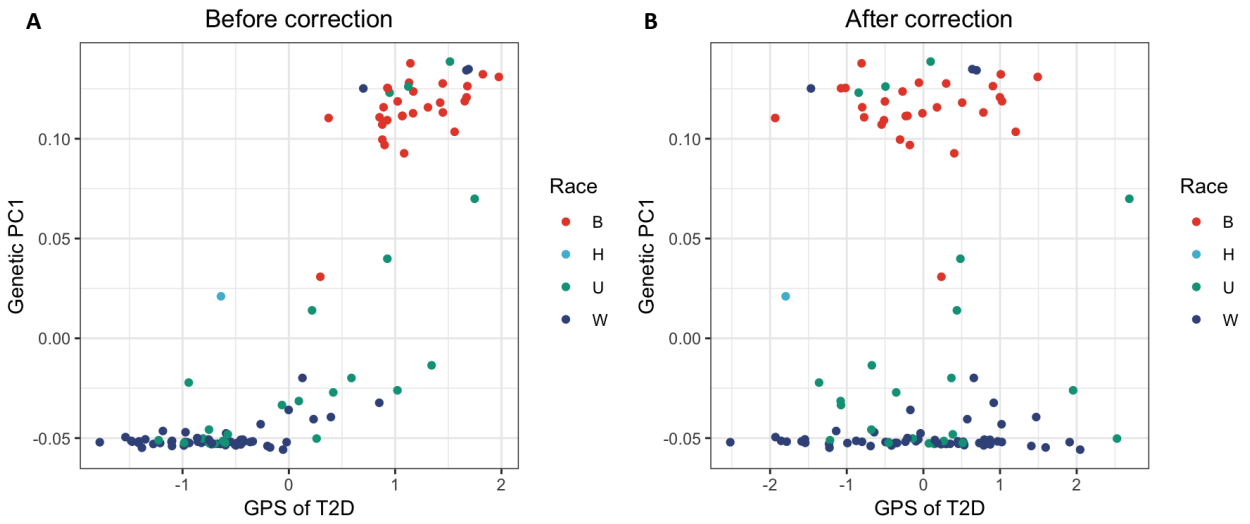


Figure 1.15 Ancestry corrected polygenic scores.

A. Scatter plot of the raw polygenic score of T2D and the first genetic PC; B. Scatter plot of the ancestry corrected polygenic score of T2D and the first genetic PC.

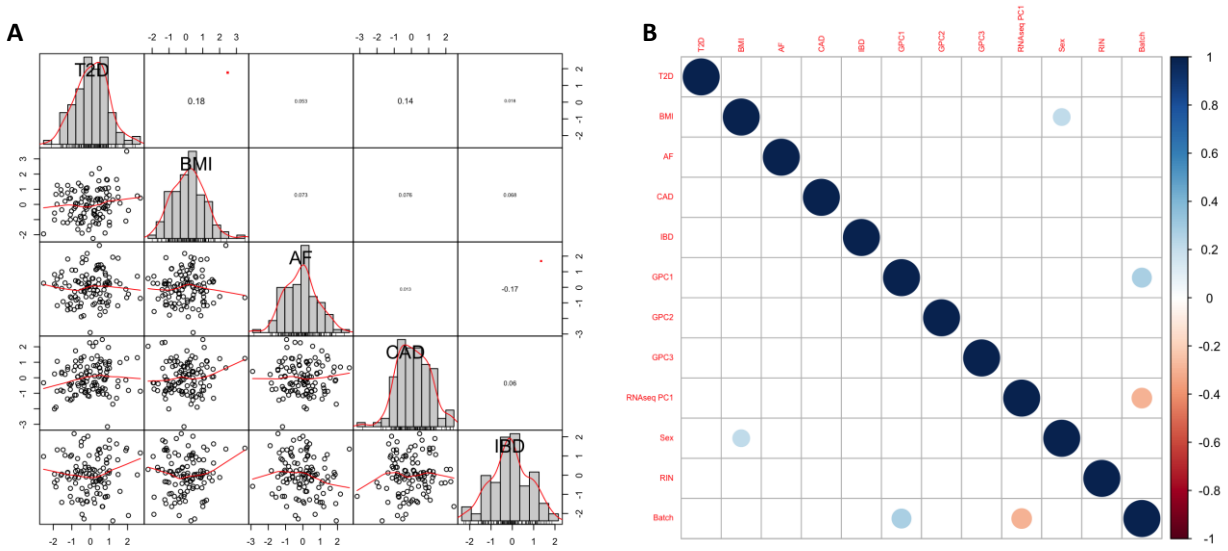


Figure 1.16 Correlation matrix of the ancestry corrected polygenic scores

A. The distribution and pair-wise correlation of the ancestry corrected polygenic scores of 5 metabolic diseases; B. Correlation matrix between the corrected polygenic scores of 5 metabolic diseases and covariates. Scale bar represents the Pearson correlation coefficient, and only correlation with $p < 0.05$ was plotted. T2D: type 2 diabetes, BMI: body mass index, AF: Atrial fibrillation, CAD: Coronary artery disease, IBD: Inflammatory bowel disease.

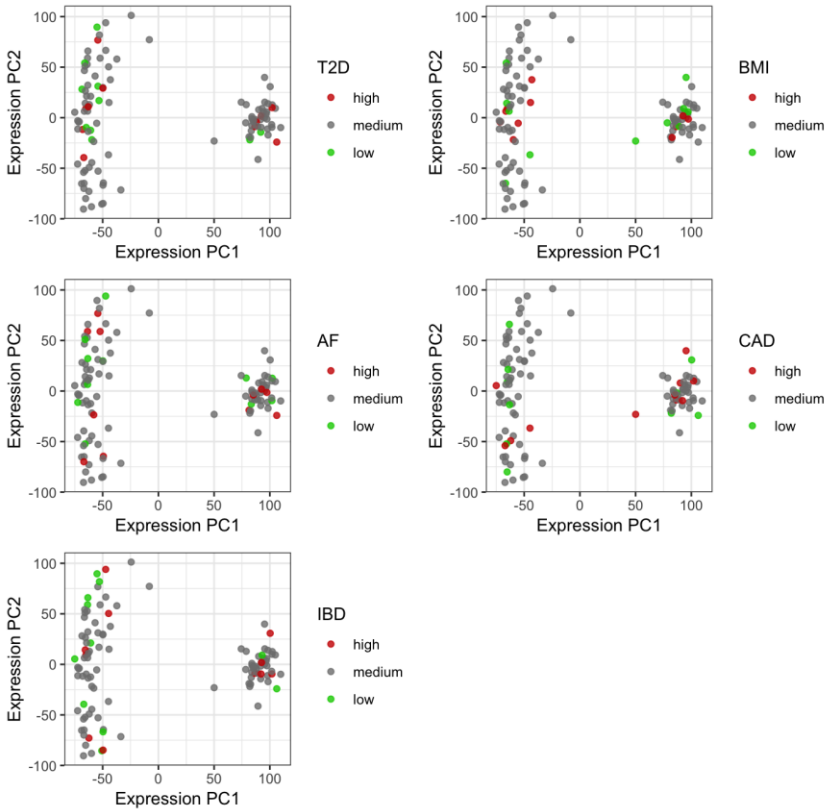


Figure 1.17 Disease risk stratification

Risk stratification for the 5 metabolic disease. People with the top 10% of the ancestry corrected polygenic scores were considered as high risk, the tail 10% as low risk, and the rest as medium. T2D: type 2 diabetes, BMI: body mass index, AF: Atrial fibrillation, CAD: Coronary artery disease, IBD: Inflammatory bowel disease.

A total of 32 genes have been identified to be differentially expressed (FDR adjusted $p < 0.05$ and fold change > 2) between the high and low risk groups. Interestingly, some of the identified genes have been reported to be involved in the pathogenesis of the corresponding diseases. For example, RSC1A1 is down-regulated in the prenatal samples characterized with high-risk for obesity (Figure 1.18 B, which is consistent with the study showing that RSC1A1 knockout mice displayed increased glucose transport and developed obesity (Osswald et al., 2005). Similarly, we found CD69, which is a critical regulator and potential therapeutic target of IBD (Radulovic & Niess, 2015), was up-regulated in the IBD high-risk group (Figure 1.18 E).

To capture the overall signature of the differentially expressed genes, we performed the KEGG pathway analysis with a relatively liberal threshold (nominal $p < 0.05$ and fold change > 2). We found the top enriched pathways were highly functional relevant to the corresponding diseases. For example, the PI3K/Akt signaling pathway, which is a critical pathway underlying the development of T2D (X. Huang, Liu, Guo, & Su, 2018), was found to be enriched in the identified T2D risk genes (Figure 1.19). These findings suggest that hepatic molecular profiles have already been significantly altered at the early stage of life among the individuals with high genetic risk for chronic disease.

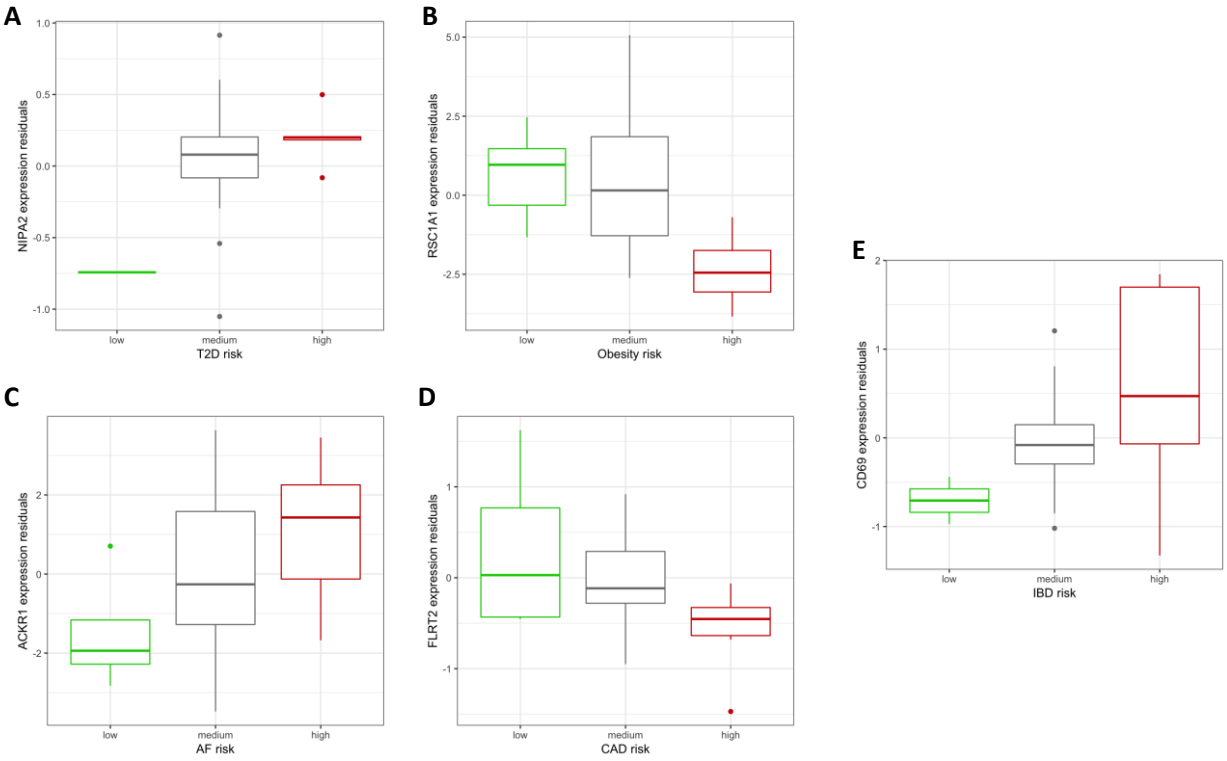


Figure 1.18 Boxplot of the representative differentially expressed genes in the prenatal livers

A. Boxplot of NIPA2 expression in prenatal livers; B. Boxplot of RSC1A1 expression in prenatal livers; C. Boxplot of ACKR1 expression in prenatal livers; D. Boxplot of FLRT2 expression in prenatal livers; E. Boxplot of CD60 expression in prenatal livers.

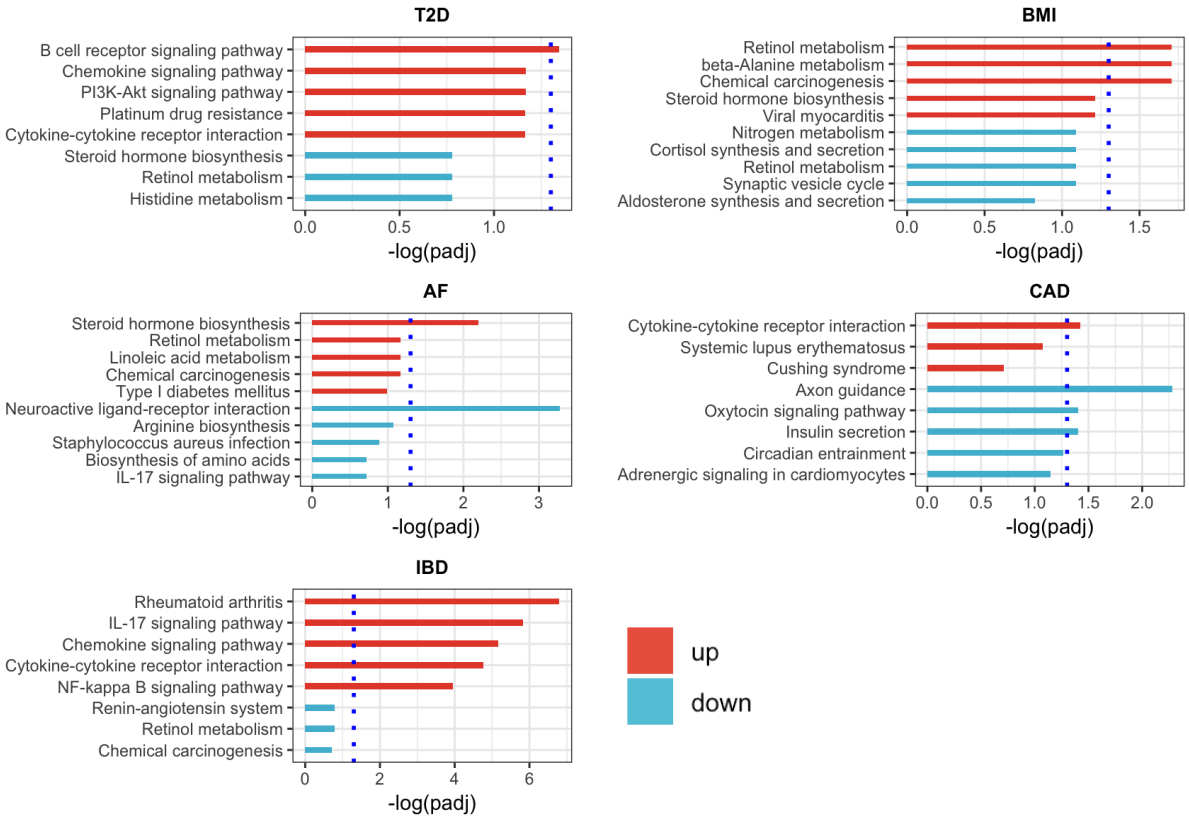


Figure 1.19 KEGG pathway analysis of the differentially expressed genes in the prenatal livers.

T2D: type 2 diabetes, BMI: body mass index, AF: Atrial fibrillation, CAD: Coronary artery disease, IBD: Inflammatory bowel disease.

1.4 Discussions

The transcriptome profiles of liver tissues in different developmental stages were analyzed in the present study to provide insights into the human liver development and the pathogenesis of metabolic diseases at both gene and isoform level. Despite ongoing endeavors (Cardoso-Moreira et al., 2019; Huse, Gruppuso, Boekelheide, & Sanders, 2015), this is the first large-scale study to depict the multiple facets of human liver development. In specific, this study is significant in the following aspects.

Firstly, we characterized a list of temporally dynamic genes using the largest-to-date cohort of human pediatric liver samples. Consistent with previous studies (Cardoso-Moreira et al., 2019;

Huse et al., 2015), we observed dramatic changes in human liver transcriptome during the course of development. In specific, the developing livers tend to display a proliferative profile, in which cell cycle, DNA replication and repair related genes were highly expressed. In the contrast, the developed livers tend to have differentiated liver function (e.g. drug metabolism and fatty acid degradation) related genes highly expressed. Our findings supported a transcriptome-wide transition from a stem-cell like to a liver-specific gene expression signature. This is consistent with the notion that development is a dynamic balancing process between proliferation and differentiation (Ruijtenberg & van den Heuvel, 2016), but was the first time to show in human livers at the transcriptome-wide level. Interestingly, we found the temporal regulation is more dynamic at the isoform level compared to the gene level. This suggested the developmental regulation might be controlled with more subtle and complex mechanisms in an isoform-specific way, which is consistent with the observations in other tissues (Gandal et al., 2018; M. Li et al., 2018).

Secondly, the emerging single cell data provide us an opportunity to explore the human liver development at the cell type specific resolution. We successfully decomposed the bulk RNA expression into cellar composition and estimated the trajectory of major hepatic cell types over development. Consistent with the knowledge that hepatocyte is the major liver functional cell type, we observed a surge in hepatocyte proportion from prenatal to postnatal stages. The proportion of hepatocyte correlated closely with the expression levels of the hepatocyte differentiation controlling transcription factor HNF4A, suggesting our estimation accurately reflected the true trajectory of hepatocytes proportion. Interestingly, we observed a decreased cell type heterogeneity from the embryonic stage till the end of infancy. The increasing

uniformness is consistent with the switch to a more functional profile. Despite the significance of the cellular composition analysis, our study is limited in the selection of the cell type signatures. Due to the limited availability of the single cell data of prenatal livers, we performed our analysis using only the mature cell type signatures collected from adult livers. Therefore, we don't have the power to observe the transition from the progenitor cells to the mature cells in our system. Future single-cell or single-nucleus RNA sequencing analysis are in need to investigate the cell type heterogeneity change in the prenatal liver samples.

Moreover, our polygenetic score analysis stratified the disease risk and identified both known and novel disease risk genes. Our analysis framework is conceptually similar to the transcriptome-wide association study (TWAS) (Gamazon et al., 2015; Gusev et al., 2016) or summary-data-based mendelian randomization (SMR) approaches (Zhu et al., 2016), in which genes associated with the predisposed genetic risk are more likely to be the causal genes due to fewer biases from confounding effects and reverse causality. What's more, we chose to focus on the prenatal samples due to less unknown environmental exposures in the prenatal stage. Several identified risk genes are highly relevant to the development of the corresponding disease. For example, CD69, a determining regulator of inflammation and potential therapeutic target for IBD (Radulovic & Niess, 2015), was found to be up-regulated in the prenatal liver tissues of the IBD high-risk individuals. This finding not only suggested the involvement of liver in IBD given the similar expression pattern of liver and small intestine (Consortium et al., 2017), more importantly, revealed that the transcriptomic signatures of high-risk people might be dysregulated at the very beginning of the development process due to the predisposed genetic risk. This is significant for the disease risk prediction and development of therapeutic treatment

for the target genes. Despite the significant findings revealed by the polygenic risk analysis, it is of note that the GPS score calculation (even after ancestry correction) is still prone to error, especially in the non-Caucasian ethnic groups (De La Vega & Bustamante, 2018). The GPS analysis is also limited in understanding liver disorders such as NAFLD, this is because construction of robust GPS model needs a large volume of independent genomics data which is not available for liver disorders yet. Future studies on the construction of GPS for liver diseases are needed to identify corresponding risk genes.

Taken together, our study significantly improved the current understanding of human liver development through a systematic analysis of the transcriptomic signature of human pediatric liver samples. Future studies at the single cell and epigenome level are warranted to further decipher the detailed mechanisms of human liver development and the pathogenesis of liver disorders.

CHAPTER 2. ALLELE SPECIFIC EXPRESSION ANALYSIS TO INVESTIGATE THE GENETICS CONTROL OF GENE EXPRESSION IN HUMAN PEDIATRIC LIVERS

2.1 Disclaimer

The area under the curve (AUC) data in Figure 2.11 were collected by ChienWei (Jack) Chiang and Lang Li from the Department of Biomedical Informatics of the Ohio State University.

2.2 Introduction

Numerous studies have revealed the significant associations between genetic variations and human gene expression in adult tissues (Consortium et al., 2017). However, recent studies have shown that the genetic control of gene expression is not only tissue-specific, but also highly dependent on the stages of the developmental process (Flutre, Wen, Pritchard, & Stephens, 2013; D. Wang et al., 2018). Consistent with the drastic transcriptome change (see details in **Chapter I**) between prenatal and postnatal liver, genetic mutations would affect multiple aspects of liver development and lead to serious liver disorders (Rao, Asch, & Yamada, 2017; Saha et al., 2014). Due to the limited availability of human pediatric liver samples, the associations and interactions between genetic variants and hepatic gene expression in the process of development remain largely unknown.

Expression quantitative trait loci (eQTL) analysis is the most common approach to investigate the relationships between genetic variations and gene expression. However, due to the multiple testing burden (Q. Q. Huang, Ritchie, Brozynska, & Inouye, 2018), large sample size is needed to identify significant associations, which is unavailable for the pediatric liver samples yet. Also, despite the advancement of fine mapping approaches (Roytman, Kichaev, Gusev, & Pasaniuc,

2018; Tehranchi et al., 2019), the identified significant eQTLs might not be the causal regulatory variants due to the existence of LD structure (Q. Q. Huang et al., 2018).

Allele specific expression (ASE) analysis, as an alternative approach, aims to identify significant allelic imbalance at the heterozygous sites. Therefore, ASE can be used to interrogate the cis-regulating effects at the individual level (Bell & Beck, 2009). As such, ASE is less likely to be affected by the confounding issues due to the existence of perfect internal control (Castel, Levy-Moonshine, Mohammadi, Banks, & Lappalainen, 2015). To systematically investigate the genetics control of hepatic gene expression during development, we tested the ASE at all the heterozygous sites of each of our pediatric liver samples. Then we analyzed if prenatal and postnatal samples display different ASE pattern. To understand the interactive effects between genetic variants and age on gene expression, we analyzed the associations between ASE and developmental stage using a newly developed powerful approach EAGLE (Knowles et al., 2017a). In the end, we intergraded both genetics and non-genetics information e.g. age and sex to build up predictive models for important pharmacogenomics (PGx) genes, with the aim of improving pharmacokinetic /pharmacodynamic (PK/PD) modeling for pediatric patients. A diagram summarizing the key question and major analyses in this chapter was shown in Figure 2.1.

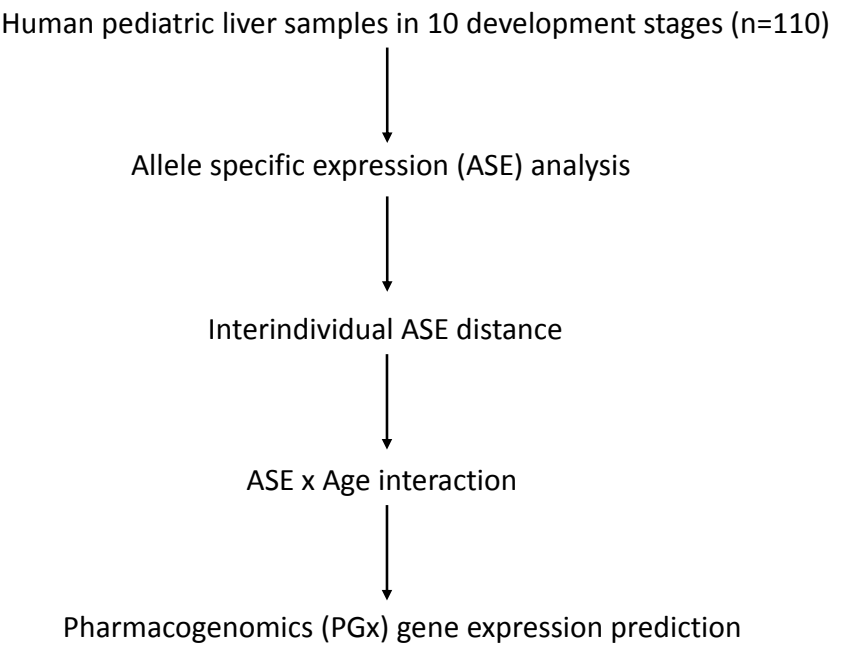


Figure 2.1 Analysis workflow

2.3 Materials and methods

2.3.1 Raw data generating

A total of 190 DNA samples and 6,233,264 high-quality variants were genotyped or imputed. The RNA samples (n =110) of the same set of livers were sequenced and mapped to the GRCh38.90 reference genome. More details on the raw data generating were introduced in the sections 1.2.3, 1.2.4, and 1.2.5.

2.3.2 Sequencing filtering by WASP

Allele specific expression (ASE) analysis is known to be sensitive to mapping biases, in which the reference allele is more likely to be mapped than the alternative allele (Castel, Levy-Moonshine, Mohammadii, Banks, & Lappalainenii, 2015). To reduce the potential mapping

biases, we filtered out reads that mapped differently to the reference and alternative genome using the WASP method (van de Geijn, McVicker, Gila, & Pritchard, 2015). In brief, for the reads overlapping genetic variants of interest (see section 2.3.3.1 for details), the allele in the read was flipped and tested if the read was able to be mapped to the original position. The original WASP method was re-implemented in the mapping tool STAR v.2.7.0c (Dobin & Gingeras, 2015) by adding the vW tag to the output bam files. We only kept reads with vM:i:1 tag (passed the WASP filtering) for the subsequent analyses. Finally, we de-duplicated the reads using the rmdup.py function of WASP tools. Importantly, one copy of the duplicated reads was randomly retained instead of the copy with the highest score to reduce the potential mapping biases.

2.3.3 ASE analysis

2.3.3.1 Genetic variants of interest

A total of 6,233,264 high-quality variants (see section 1.2.3 for details) existed in our liver samples were used as the core set of SNPs to be analyzed. We removed SNPs located in the ENCODE blacklist regions (Dunham et al., 2012) and areas with mappability < 1 from the 75-mer UCSC genome browser mappability track (http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeMapability/wgEncodeCrgMapabilityAlign75mer.bigWig). Following the suggestion of Panousis NI et al. (Panousis, Gutierrez-Arcelus, Dermitzakis, & Lappalainen, 2014), we then removed SNPs displaying > 5% mapping bias based on 50bp simulated reads (ftp://jungle.unige.ch/Allelic_map_bias//paired_end_genome_based_BWA/EUR01_50bp_result_stats_05bias.txt). The left SNPs were subject to be analyzed for allele specific expression.

2.3.3.2 Allele specific read count

We processed our data following the QuASAR pipeline (Harvey et al., 2015). Briefly, we used the mpileup function of SAMtools (H. Li et al., 2009) to compile counts of read covering both reference and non-reference alleles at each of the SNPs of interest with GRCh38.90 reference genome. Then SNPs that were not covered by a read or within a splice junction were removed to avoid potential biases.

2.3.3.3 Heterozygous SNPs confirmation

Genotyping or imputation error would lead to serious false-positive issues for ASE inference. For example, if it a homozygous site was mistakenly considered as heterozygous during DNA genotyping or imputation steps, then the reference ratio (ref allele count / total count) of this SNP would be extreme and prone to be inferred as a significant ASE site. To reduce the biases introduced by the genotyping/imputation error, we used the QuASAR method (Harvey et al., 2015) to orthogonally call the genotypes from the RNA-Seq reads. Then the final list of heterozygous SNPs to be analyzed for ASE were selected: a. heterozygous by DNA genotyping/imputation; b. heterozygous by mRNA-based genotyping (posterior probability of being heterozygous > 0.99); c. the total coverage is larger than 20. Overall, we observed a high concordance between DNA and RNA based genotypes. In average, 95.8% of the heterozygous SNPs were confirmed by both DNA and RNA based genotyping, while 2.5% and 1.7% were characterized as DNA and RNA only, respectively (Figure 2.2 A). The inconsistency between DNA and RNA genotypes could due to technical issues (e.g. genotyping, imputation, or sequencing error) or biological factors such as imprinting or RNA editing.

As a quality check of the mapping bias, we calculated the genome-wide reference ratio (defined as the mean reference ratio of all the Het SNPs in a sample) of each sample. The average genome-wide reference ratio of our sample is 0.498 (Figure 2.2 B), which is very close to the theoretical threshold of 0.5 (Castel, Levy-Moonshine, Mohammadii, et al., 2015), suggesting the mapping bias is unlikely to confound the ASE inference.

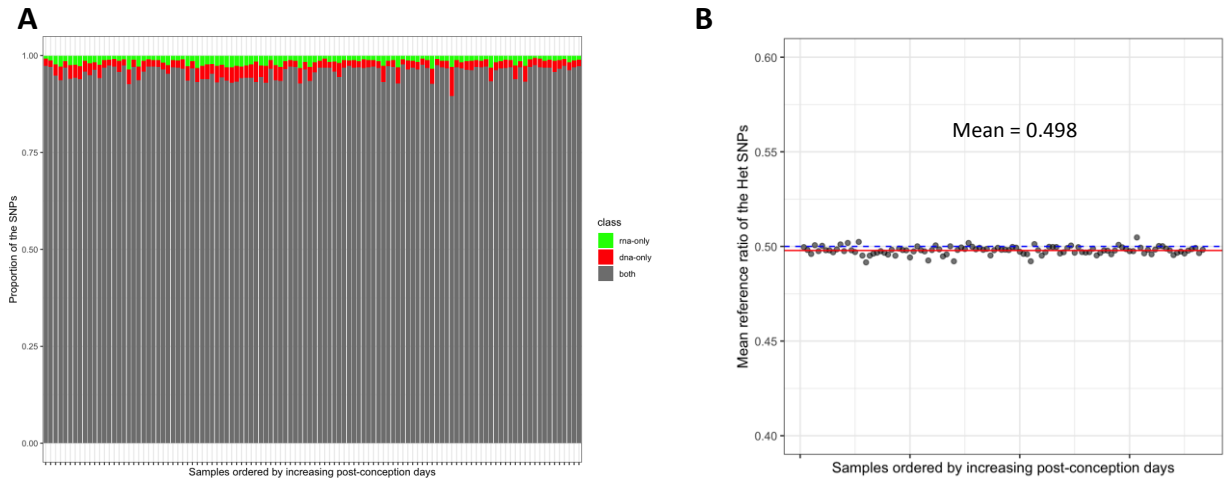


Figure 2.2 Mapping biases quality control

A. The proportion of DNA-confirmed, RNA-confirmed, and both RNA and DNA confirmed heterozygous SNPs in each liver sample; B. The genome-wise reference ratio (ref count/ total count) of each liver sample, blue dash line indicates the theoretical threshold of 0.5, the red line indicates the mean reference ratio of all the samples.

2.3.3.4 ASE inference

Inspection of the distribution of the effect size (defined as reference ratio – 0.5) of the tested heterozygous SNPs didn't observe obvious overdispersion issue (Figure 2.3 A), which is the major concern of using binomial test in ASE inference. Therefore, we chose to use the binomial test for the better power over the beta-binomial model (Figure 2.3 D).

The two-sided binomial test was used to determine whether the reference ratio was significantly deviated from 0.5, following by the Benjamini-Hochberg (BH) multiple testing adjustment. As ASE was inferred at a per-SNP base, different coverages at different sites would affect the statistical power. Therefore, we required the converge > 20 to ensure adequate power while filtered out Het SNPs with absolute effect size less than 0.1 to account for the potential false-positive at the high coverage sites. Moreover, we only considered the recurrent ASE events (FDR < 0.05 and |effect size| > 0.1 in at least two individuals) for further analysis. The significant ASE variants were annotated by an online SNP annotation tool SNPnexus (https://snp-nexus.org/) (Ullah et al., 2018).

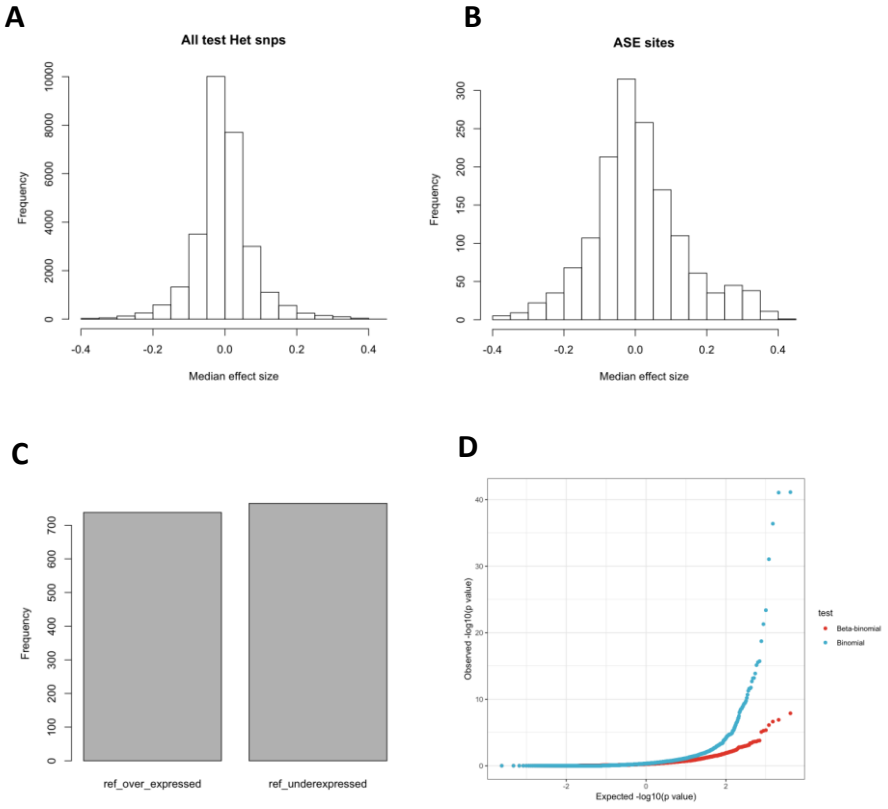


Figure 2.3 Distribution of the median effect size

A. Distribution of the median effect size (ref count / total count – 0.5) across all heterozygous individuals for all tested Het SNPs; B. Distribution of the median effect size (ref count / total count – 0.5) across all heterozygous individuals for the significant ASE sites; C. The number of over-expressed and under-expressed allele; D. QQ-plot of ASE inference using binomial test and beta-binomial test in one example sample.

2.3.4 Interindividual ASE distance

To examine the similarity within individuals with regard to the ASE pattern, we defined the ASE distance between two individuals as $1 - \frac{\mathrm{N_{sharedASE}}}{N_{sharedHet}}$, where $\mathrm{N_{sharedASE}}$ represents the number of shared significant ASE sites (FDR < 0.05 and |effect size| > 0.1), $N_{sharedHet}$ represents the number of overlapped heterozygous SNPs between two individuals. We then visualized the ASE distance matrix using multidimensional scaling (MDS) plot with two dimensions.

2.3.5 Comparison with ASE in human adult livers

To make a comparison with the published adult ASE data using GTEx tissues (V6, October 2015), we download the ASE track from UCSC genome browser for liver tissues (http://hgdownload.soe.ucsc.edu/hubs/gtexAnalysis/hg38/gtexAwgAseLiver.bb).

The effect size of an ASE site was defined as the median allelic imbalance (|0.5 – ref count / total count|) across all individuals. More details on the ASE inference in adult tissues can be found in the previous publication (Castel, Levy-Moonshine, Mohammadii, et al., 2015). Kolmogorov–Smirnov test was used to determine if the two densities were equal.

2.3.6 ASE x Age interaction

To systematically evaluate the interactions between genetic variation and age, we used EAGLE (Knowles et al., 2017b), a recently developed method to identify significant GxE interactions using ASE data. EAGLE is conceptually similar to the spearman association between allelic imbalance and environmental factor of interest such as age but has been shown to improve power and reduce false-positive through modeling the read counts directly and accounting for the overdispersion (Knowles et al., 2017b). For the input of Het SNPs to be tested, we used the same

threshold as in the ASE identification i.e. both DNA and RNA confirmed Het sites and converge > 20 counts. Following the author's suggestion, we required at least 20 heterozygous individuals to test for the significance of the interaction effect between ASE and age, which is quantified as the log10 transformed post-conception days. FDR adjusted p value less than 0.1 was considered significant. Two-sided Fisher's exact test was used to determine the enrichment of the significant ASE-age interaction in the significant ASE events. KEGG pathway analysis was used to identify significantly enriched pathways in the genes harboring GxE variants.

2.3.7 Pharmacogenomics (PGx) genes expression prediction

To help improve the pharmacokinetic/pharmacodynamic (PK/PD) modeling for pediatric patients, we constructed the predictive model for several important PGx genes using both genetic variation and non-genetics factor including sex and age as predictors. In specific, we used the bayesian variable selection regression (BVSR) method to model the hepatic gene expression using varbvs (http://github.com/pcarbo/varbvs) due to its ability to select significant variants at the genome-wide level. The predicted PGx genes include CYP3A4, CYP3A5, ABCC10, and SLC29A3. The SNPs within 1Mb of each gene, log10 PCD, sex, and the first 3 genetic principle components were used to train the predictive model using our pediatric liver set. Then the model was applied to an independent cohort of pediatric patients to calculate their virtual hepatic gene expression levels. To validate the predictive performance, we correlated the predicted gene expression levels with the area under the curve (AUC) data of drug metabolism collected by our collaborators.

2.4 Results

2.4.1 Identification and characterization of ASE in human pediatric liver samples

Allele specific expression analysis is a sensitive approach to explore the cis-regulating effects through identifying significant allelic imbalance (Figure 2.4 A). After stringent quality control processing steps, we characterized a total of 1,503 significant ASE variants out of 28,810 tested heterozygous SNPs. Consistent with the findings in eQTL and GWAS studies (Park et al., 2011), variants with lower minor allele frequency tend to have larger effect size (Figure 2.5 B) and be more likely to be characterized as significant ASE (Figure 2.5 C).

The ASE sites were annotated to be within the transcripts of 777 genes. These ASE-overlapping genes tend to be enriched in the pathways related to liver functions such as drug metabolism and fatty acid degradation (Figure 2.4 B), suggesting the involvement of cis-regulating effects in liver functions.

The majority of the ASE sites located in the 3 prime downstream or the 3 prime UTR regions of the corresponding transcripts, while 5 prime UTR has the least ASE sites. This is consistent with the finding that 3' UTR (1028bp on average) is around 5 times longer than 5' UTR (210bp on average) for human beings (Mignone, Gissi, Liuni, & Pesole, 2002). Moreover, the 3' overrepresentation of RNA-Seq coverage (Figure 1.3 C) might contribute to the most ASE identification in the 3' UTR as well.

Most of the ASE variants are synonymous mutations or benign non-synonymous substitutions (Figure 2.4 D), however, we identified several ASE sites that would lead to potentially damaging

effects such as stop codon gain/loss mutations. For example, the A allele of rs1790218 introduced a stop codon gain mutation in SLC22A10, which is a liver-specific organic anion transporter (OAT) gene involved in transporting anionic drugs such as antibiotics (Klein et al., 2010; Sweet, 2005). The stop gain mutation leads to significant expression reduction at both gene ($p = 0.001$, Figure 2.6 A) and isoform level ($p = 0.003$, 0.0002, and 0.001, respectively, Figure 2.6 B, C, and D).

We also observed that the ASE variants tend to within the regulatory regions of the genome (Figure 2.4 E). For example, we identified the most ASE sites within the promoter regions, indicating the cis-regulating effects might be mediated by affecting the promoter binding affinity.

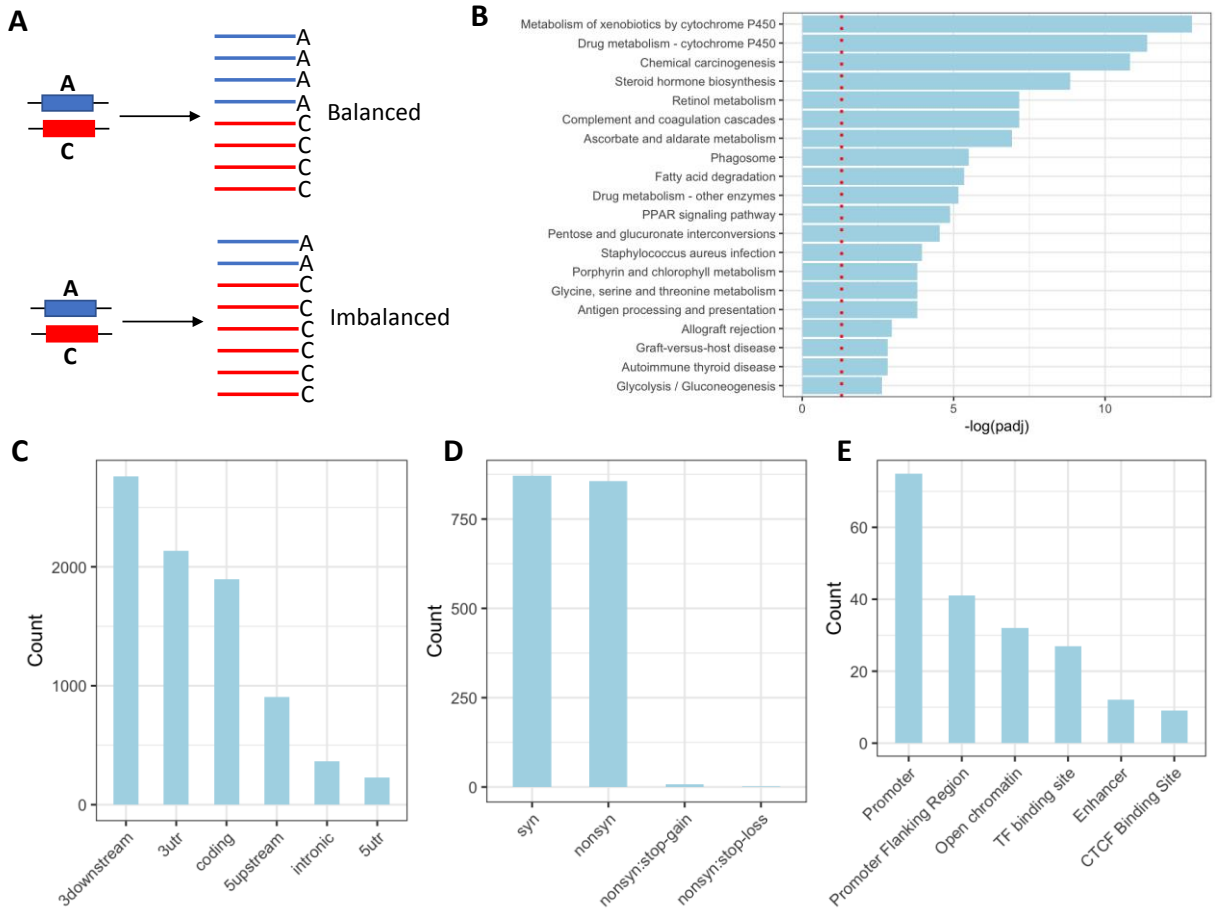


Figure 2.4 ASE identification and characterization

A. The diagram of ASE; B. KEGG pathway analysis of the genes containing ASE sites; C. The distribution of the location of ASE in the transcripts; D. The distribution of the type of mutation for ASE in the coding region; E. The distribution of the ASE sites that located in the regulatory regions of the human liver HepG2 cells.

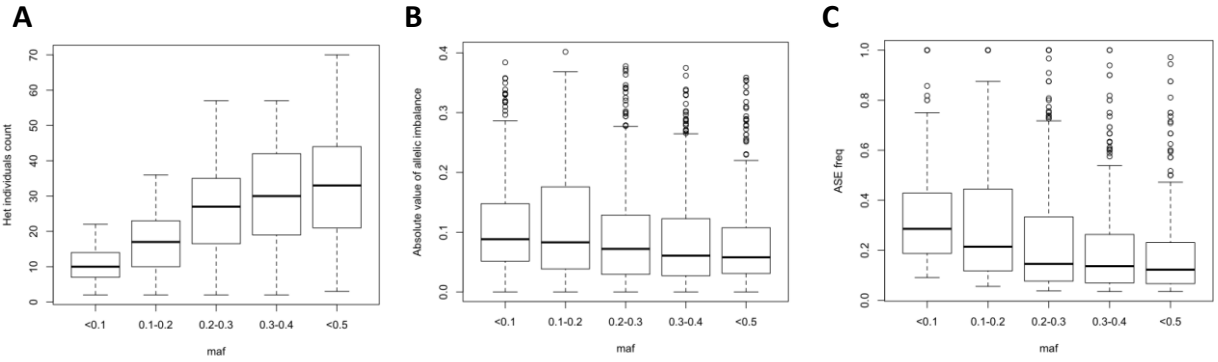


Figure 2.5 ASE and allele frequency

A. The distribution of the number of tested heterozygous SNPs in different allele frequency groups; B. The distribution of the absolute allelic imbalance (|ref count/ total count – 0.5|) of the ASE sites in different allele frequency groups; C. The distribution of the ASE frequency (number of ASE sites / number of the tested Het SNPs) in different allele frequency groups.

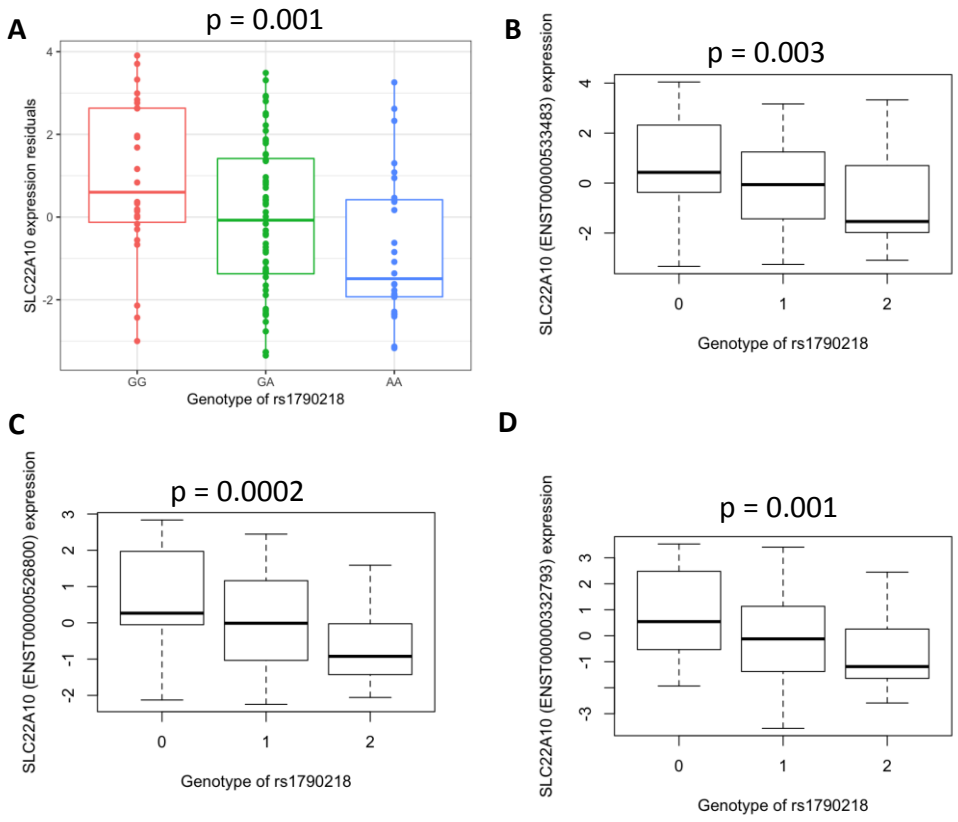


Figure 2.6 An example stop-gain ASE in gene SLC22A10

A. Boxplot of the SLC22A10 expression; B-D: boxplots of the expression levels of three SLC22A10 isoforms. Linear regression was used to determine the significance.

2.4.2 Prenatal and postnatal livers exhibited different ASE signatures

To address the question of whether development affects the ASE pattern in human livers, we explored the ASE frequency (proportion of ASE sites in Het SNPs) of each sample. Interestingly, we found that the postnatal samples tend to have higher ASE frequency compared to the prenatal samples (1.8% and 3.2% in prenatal and postnatal samples, respectively, $p < 2.2e-16$) (Figure 2.7 A). Moreover, we calculated the pair-wise interindividual ASE distance and found that samples were clustered by their developmental stages (Figure 2.7 B). Due to the different genetic background of prenatal and postnatal samples (most of the prenatal samples were from black people, while most of the postnatal samples were from white people), we explored whether

ancestry would confound the observed ASE signatures. The correlation matrix indicated that the ASE frequency is positively correlated with age, but not correlated with the first genetic PC (Figure 2.8 A). To account for the potential confounding effects, we regressed sex, genetic PC1, RIN, library batch, and read depth on the ASE frequency and found the residuals were still positively correlated with age (r = 0.4, p = 1.6e-5, Figure 2.8 B). The MDS plot of ASE distance also demonstrated that ancestry didn't separate the samples into obvious clusters (Figure 2.8 C). Therefore, we concluded age/developmental stage was the major factor leading to the different ASE signatures.

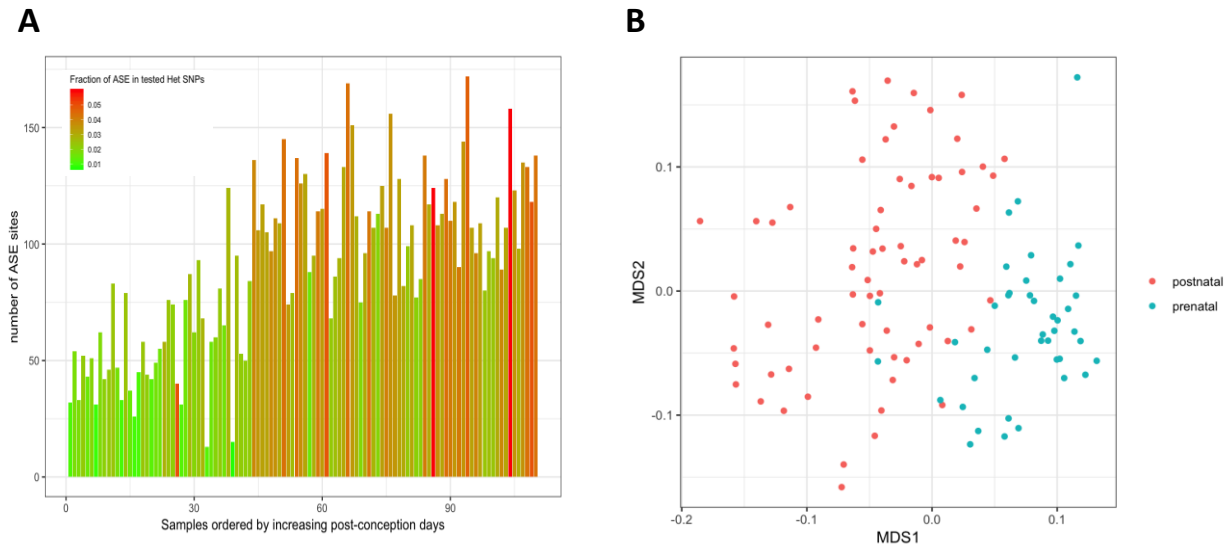


Figure 2.7 ASE patterns in prenatal and postnatal samples

A. The number of ASE sites and ASE frequency of the liver samples. The color represent the frequency (number of ASE sites / number of the tested Het SNPs) of the ASE sites in each individual. B. The MDS plot of the interindividual ASE distance.

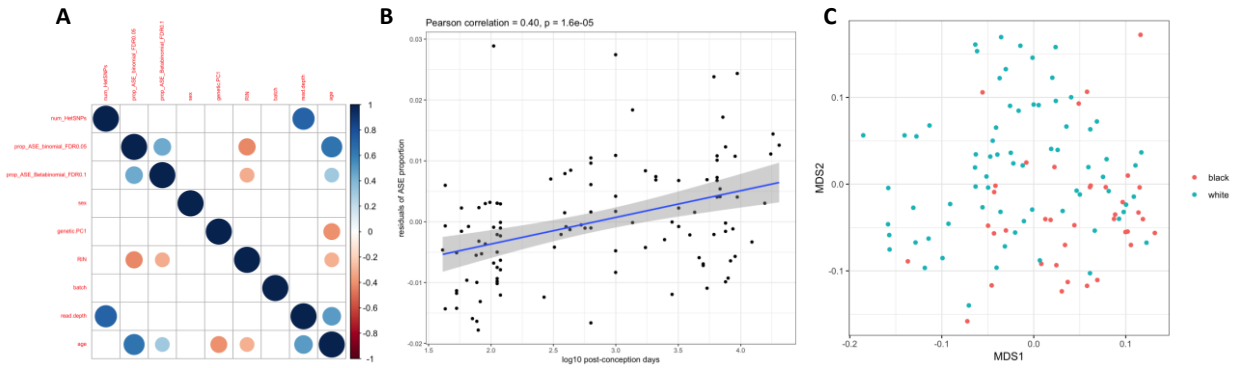


Figure 2.8 Ancestry and ASE frequency

A. Correlation matrix between ASE frequency and covariates; B. Scatter plot of the age (log10 post-conception days) and ASE frequency adjusted for covariates; C. The MDS plot of the interindividual ASE distance.

2.4.3 Comparison between ASE in pediatric livers and adult livers

Next, we compared the allelic imbalance in our pediatric set with the published human adult liver tissues (GTEx V6). Overall, we observed good concordance between pediatric liver and adult livers (Figure 2.9 A, B), suggesting ASE as a general mechanism regulating hepatic gene expression. In specific, postnatal samples tend to have more similar allelic imbalance (rho = 0.48 [0.43, 0.52], p = 6.8e-66, Figure 2.9 B) than prenatal samples (rho = 0.34 [0.26, 0.41], p = 4.3e-16, Figure 2.9 A). While both prenatal and postnatal livers have different densities as adult livers (both $p < 2.2e-16$ by K-S test), the density of postnatal samples is relatively similar to the adult counterpart (p = 0.001 by K-S test, Figure 2.9 C), which is consistent with the notion that age plays a role in shaping ASE signatures.

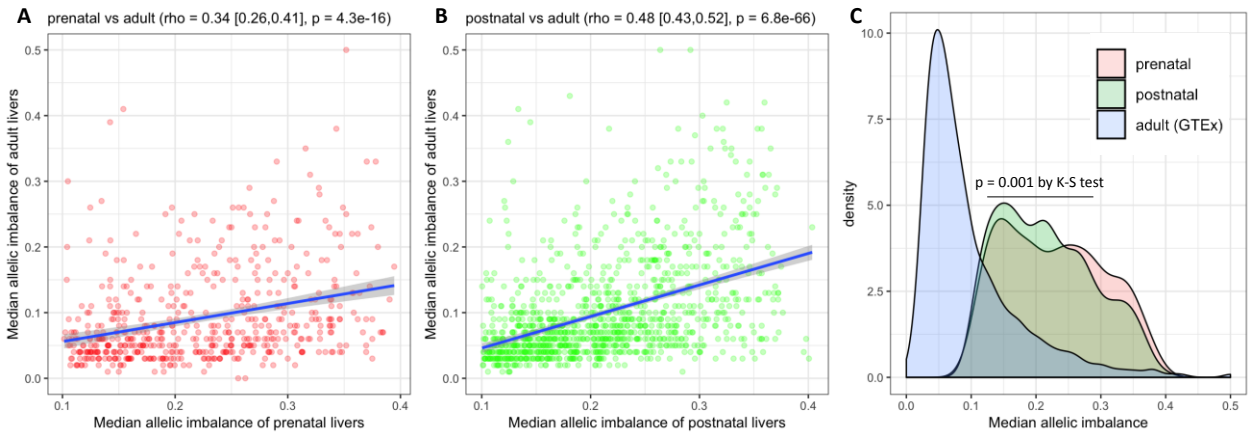


Figure 2.9 Comparison between ASE in pediatric and adult livers

A. The correlation between median effect size of ASE sites in prenatal and adult livers; B. The correlation between median effect size of ASE sites in postnatal and adult livers; C. The densities of the distribution of the effect sizes. K-S test was used to determine the significance.

2.4.4 ASE x age interaction

Given the different ASE patterns in the prenatal and postnatal samples (section 2.4.3), we aimed to identify the specific age-dependent ASE sites at the population level (Figure 2.10 A). In total, we identified 35 significant (FDR < 0.1) ASE-age interactions. These age-dependent ASE sites were significantly enriched in the ASE sites identified at the individual level ($p < 2.2e-16$ by two-sided Fisher's exact test). As with individual-level ASE sites, age-dependent ASE variants tend to be located in the genes related to major liver functions such as fatty acid and glucose metabolism (Figure 2.10 B). For example, the allelic imbalance of the rs174546 in FADS2, the critical enzyme for long-chain polyunsaturated fatty acid (LC-PUFA) metabolism in the liver (L. Wang et al., 2015), was found to be increased over development (FDR adjusted p = 1.4e-6, Figure 2.10 C), suggesting an age-dependent difference in PUFA metabolism for people with different genetic background. Similarly, we also found the allelic imbalance of KLF10, the determining transcription factor regulating hepatic glucose metabolism (Yang et al., 2017), and

PON1, a liver fibrosis protect gene (Marsillach et al., 2009), were significantly changed over the course of development (FDR adjusted p = 0.02 and 0.09, respectively, Figure 2.10 D, E).

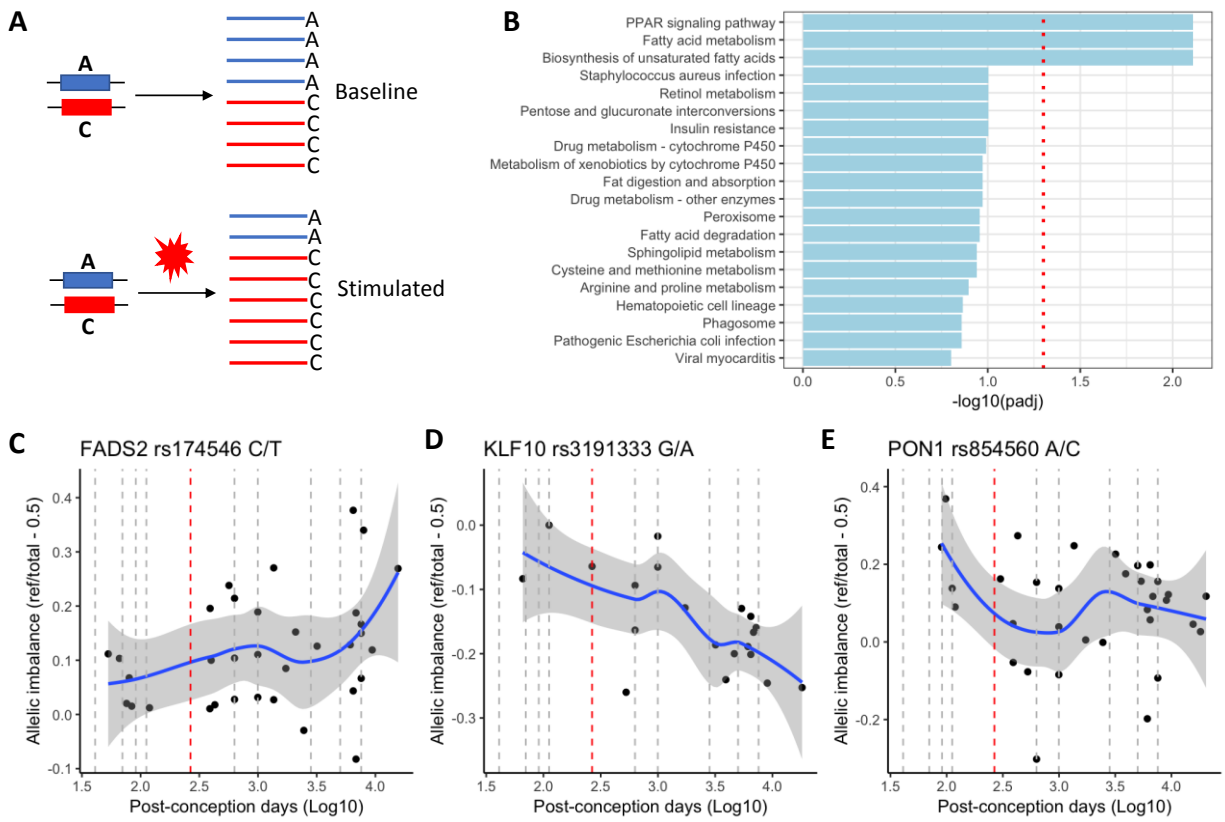


Figure 2.10 ASE x Age interaction

A. The diagram of the age-dependent ASE; B. KEGG pathway analysis of the genes containing age-dependent ASE; C-E. Allelic imbalance of three example age-dependent ASE sites. Loess regression was used to fit the trend line. Shades represent 95% CI. The red vertical line indicates the birth day, grey vertical dash lines represent the different developmental stages.

2.4.5 Hepatic gene expression imputation through the integration of both genetics and non-genetics factors

Given the pivotal role of developmental stage in shaping liver gene expression, the current adult tissue-based gene expression imputation model would not be applied well to the pediatric people (Gamazon et al., 2015). Therefore, we constructed gene expression predictive models using our pediatric liver set to best fit the need of pediatric people. As a proof of concept, we first applied

our predictive model to several important PGx genes, with the aim to improve the pharmacokinetic/pharmacodynamic (PK/PD) modeling for pediatric patients. Interestingly, we found the predicted expression levels are significantly correlated with the area under the curve (AUC) data of vincristine in an independent pediatric cohort of patients. For example, the predicted expression levels of CYP3A4, which is the major enzyme responsible for the metabolism of around half of the prescribed drugs (Zanger & Schwab, 2013), was negatively correlated with the vincristine AUC data (beta = -0.03, p = 0.04, Figure 2.11 A), which is consistent with the notion that a higher level of CYP3A4 leads to metabolism lower drug exposure. Another significant example is the transporter gene ABCC10, which has been shown to play a critical role in the resistance of multiple drugs (J. J. Chen et al., 2013; Oguri et al., 2008; Zhao et al., 2018) . We found the predicted ABCC10 expression levels were positively correlated with the vincristine exposure time (beta = 0.04, p = 0.02, Figure 2.11 C), suggesting the possibility to model drug resistance using genetic variation and non-genetics factors such as age and sex.

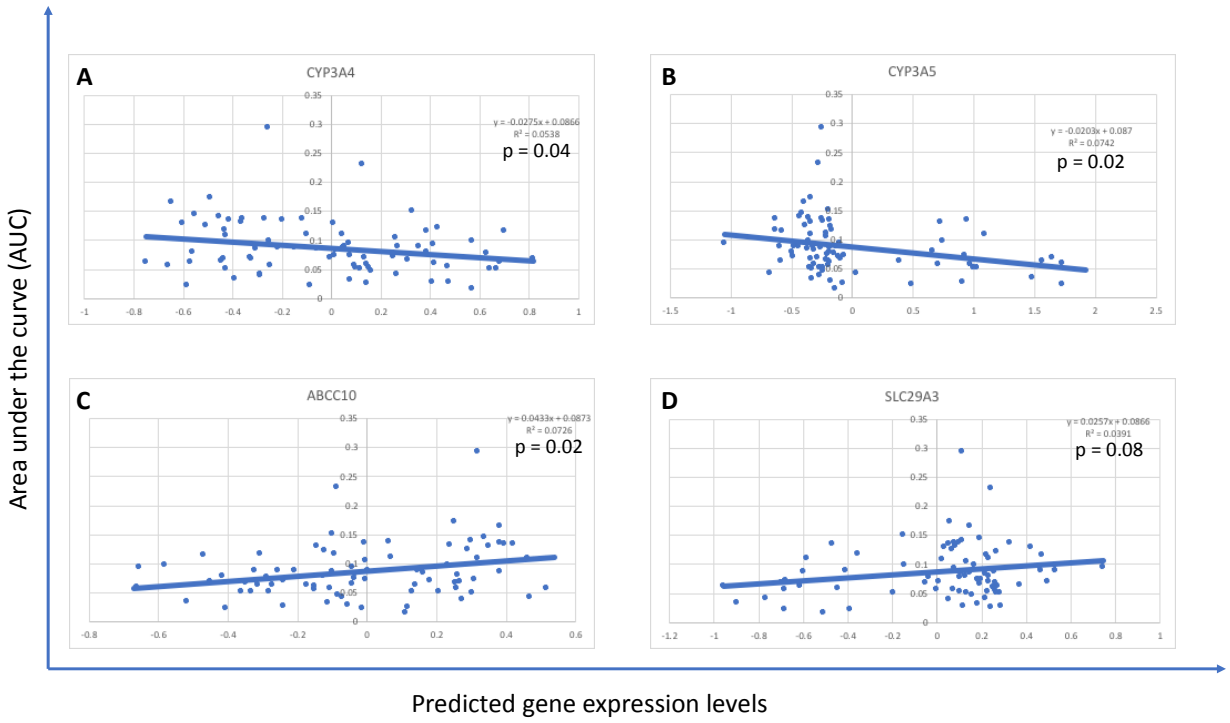


Figure 2.11 PGx gene expression prediction

Correction between the predicted levels of PGx genes and vincristine drug exposure (AUC) in an independent cohort of pediatric patients.

2.5 Discussions

We explored the effects of genetic variations on the interindividual variability of hepatic gene expression in the context of development. To our best knowledge, this is the first study to investigate the genetic control of gene expression in human pediatric livers. Our study is significant in the following aspects.

Firstly, we observed genes in the postnatal samples were more likely to display allele specific expression pattern. In other words, development/ageing is an induce factor of ASE in at least human liver tissues. As the liver is one of the major metabolic organs interacted with the environment, the more environmental exposures accumulated after birth might lead to the higher ASE rate in the postnatal stage. This is consistent with the recent studies showing that ASE is

under selection (Chen et al., 2016; Shao et al., 2019; Tian et al., 2018), in which ASE sites were less conservatory than the non-ASE sites. Given the transition from a proliferative to a differentiated expression profile from prenatal to postnatal (Chapter I), the less ASE rate in the prenatal stage might indicate the existence of a protective mechanism to tolerate genetic variations in the critical developing periods. As we also observed the predisposed genetic risk plays a role in the early stage gene expression (section 1.3.7), future studies are warranted to study the balance between the predisposed genetic load and the potential tolerate system.

Secondly, motivated by the findings that both genetics variations and development play pivotal roles in regulating hepatic gene expression. As a proof-of-concept, we constructed the predictive model for important PGx genes and found the predicted levels were significantly correlated with the drug metabolism data in an independent cohort. The virtual hepatic expression prediction model is of translational interest. For the purpose of precision medication, our study is promising in improving the modeling of PK/PD for pediatric patients due to the age-dependent PGx gene expression patterns. As for the diagnosis of serious liver disorders such as NASH and fibrosis, no biomarker was existed to screen people for their disease risks. The virtual hepatic expression indicates the baseline levels/activities of important disease-related genes, thus could serve as a potential biomarker to screen people for the need to perform further invasive diagnostic procedures such as biopsy.

In summary, through the ASE analysis in the different developmental stages of human livers, we, for the first time, identified an age-dependent allele specific expression pattern. This motivates

the construction of the pediatric-specific hepatic gene expression system, which is of important translational significance in drug and disease diagnosis.

CHAPTER 3. INTEGRATIVE OMICS ANALYSIS IDENTIFIES MACROPHAGE MIGRATION INHIBITORY FACTOR SIGNALING PATHWAYS UNDERLYING HUMAN HEPATIC FIBROGENESIS AND FIBROSIS

3.1 Disclaimer

The materials in this chapter were adapted from the author's published paper (Z. Liu, Chalasani, et al., 2019). The copyright has been granted by Wolters Kluwer Health, Inc.

3.2 Introduction

Hepatic fibrosis is a highly conserved and protective response characterized by the excessive accumulation of extracellular matrix proteins following either acute or chronic liver injury (Bataller & Brenner, 2005). Chronic viral hepatitis, alcohol abuse, and non-alcoholic fatty liver disease are the main causes of liver fibrosis (Hernandez-Gea & Friedman, 2011). Although normal liver composition can be restored following acute or transient insult, sustained chronic liver injury will lead to a progressive formation of fibrous scar tissue, which will destroy the liver architecture and eventually produce hepatocellular dysfunction (Benyon & Iredale, 2000) .The pathogenesis of liver fibrosis is not fully understood, and it remains largely unclear why the severity of the disease displays substantial variability in patients with the same set of known risk factors (Pellicoro, Ramachandran, Iredale, & Fallowfield, 2014). Therefore, there is a pressing need to understand the genetic basis underlying the liver fibrosis which may subsequently allow for identifying critical drug targets can be identified and early intervention strategies for subjects with high risk of liver disease can be developed.

The expression of alpha-smooth muscle actin (α-SMA) reflects the activation of hepatic stellate cells to myofibroblast-like cells and is closely related to human liver fibrogenesis (Mann & Smart, 2002). Sirius red staining can accurately quantify the total hepatic collagen content (Lattouf et al., 2014), which is among the predominant hepatic extracellular matrix proteins (Schuppan, 1990). Therefore, quantitative α-SMA expression and Sirius red staining are accurate and reliable markers indicating liver fibrogenesis and fibrosis, respectively. In this study, we explored the genetic basis underlying the variability of these two quantitative molecular phenotypes using a step-wise genome-wide analysis, using donor liver samples which reflects a general American population. We performed genome-wide association study on the quantitative level of the two markers. We further investigated the potential function of candidate single nucleotide polymorphisms (SNPs) in regulating mRNA expression for their nearby genes. A pathway enrichment analysis was further conducted to identify critical genes and pathways that are potentially underlying the genetic susceptibility to liver fibrosis.

3.3 Materials and methods

3.3.1 Datasets

The tissue procurement procedure and related information of the liver samples (n = 121) used in this cross-sectional study have been described in our previous studies (Gamazon et al., 2013; Innocenti et al., 2011; L. Wang et al., 2015). In brief, these liver samples were obtained from unrelated liver transplantation donors of self-reported European and African descent. Subjects with heavy alcohol consumption (> 20 g/day), hepatitis B and C virus infection, or drug-induced liver injury, were excluded from this study. The genotype and gene expression profiling of these samples have been previously analyzed and deposited to the Gene Expression Omnibus database

(accession number: GSE26106, https://www.ncbi.nlm.nih.gov/geo/). The demographical and histological characteristics of the donor liver samples used in this study have been summarized in Table 3.1. Analyses in this study were performed on anonymous individuals, thus this study is not considered to involve "human subjects". The study was reviewed and has been approved by the Institutional Review Board (IRB) of Wayne State University (approval No. 201842) on May 17th of 2018.

Table 3.1 Demographical and histological characteristics of the donor liver tissues

Item	Data
Male	82 (67.8)
Age (year)	40 (17-57)
body mass index (kg/m^2)	25.8 (21.8-29.8)
Race	
White	102 (84.3)
Black	19 (15.7)
Percentage of alpha-smooth muscle actin expression	3.6 (1.8-7.1)
Percentage of total collagen content	9.4 (5.5-14.1)
Fibrosis stage*	
Focal perisinusoidal	1 (0.9)
Perisinusoidal	6 (5.2)
No fibrosis	109 (93.9)

N=121. *missing information for 5 participants. Data are expressed as number (percent) in male, race, and fibrosis stage, and median (interquartile range) in others.

3.3.2 Liver histology characterization and α-SMA and sirius red staining and quantification

Formalin-fixed, paraffin-embedded liver sections were stained with hematoxylin and eosin and Masson trichrome stains for histological evaluation. The biopsies were scored by an experienced hepatopathologist (JL) in a blinded fashion according to the non-alcoholic steatohepatitis clinical research network liver histology criteria published by Kleiner et al. (Kleiner et al., 2005) and it showed normal in 59% and non-alcoholic fatty liver disease in 41% [fatty liver 5%, borderline non-alcoholic steatohepatitis 23%, and definite non-alcoholic steatohepatitis in 13%. Formalin-

fixed, paraffin-embedded sections were stained for α-SMA (marker for stellate cell activation) and Sirius red (total collagen content) and were digitally quantitated and expressed as a percent of total liver biopsy area using SPSS Sigma Scan Pro 5.0 software (SPSS Inc., Chicago, IL, USA).

3.3.3 Genome-wide association study (GWAS) analyses

Genotyping was performed using Illumina Human610-Quad v1.0 BeadChip array (Illumina, San Diego, CA, USA)(Innocenti et al., 2011). The overall genotyping rate was 95.16%. After excluding rare (minor allele frequencies <5%) and low quality (call rate < 90% and deviation from Hardy-Weinberg equilibrium P <1e-3) variants, there are 533,687 remaining SNPs for the linear regression analysis. Using an additive genetic model, each SNP was tested for association to α-SMA expression and hepatic collagen content, respectively. The phenotypes were normalized with log base 10 transformation. Age, gender, body mass index, and the first two genetic principal components were adjusted as covariates for the association. SNPs with P value less than 1e-4 were considered as candidate loci for the following analysis. The quality control and association test was performed using the package PLINK 1.07 (http://pngu.mgh.harvard.edu/purcell/plink/)(Purcell et al., 2007). The regional plots were generated using the package LocusZoom (http://csg.sph.umich.edu/locuszoom/)(Pruim et al., 2010).

3.3.4 Expression quantitative trait loci (eQTL) analyses

Gene expression profiling was measured using Agilent-014850 Whole Human Genome Microarray 4x44K (Agilent, Santa Clara, CA, USA) for the liver tissues of the same set of subjects (Innocenti et al., 2011). Linear regression model was used to detect transcripts

significantly associated with the lead GWAS loci within ±1 Mbp region. Age, gender, body mass index, and the first two genetic principal components were used as covariates. Associations with P value less than 0.05 were considered as significant for the following analysis. The eQTL analyses were performed using R package Matrix eQTL (Shabalin, 2012).

3.3.5 Pathway enrichment analyses

We used QIAGEN's Ingenuity® Pathway Analysis (IPA®, QIAGEN Redwood City, CA, USA; www.qiagen.com/ingenuity) to identify overrepresented signaling pathways in eQTL-controlling genes. Right-tailed Fisher's exact test was used to determine the significance level of signaling pathways, and $P < 0.05$ was considered significant.

3.3.6 Gene interactions from curated databases and text-mining

The Gene Interactions tool in University of California, Santa Cruz Genome Browser (https://genome.ucsc.edu) was used to search the gene-gene interactions. Two genes were considered to be interacted if the interaction has been supported by either curated databases or text-mining. The curated databases consist of 23 pathway or protein interactions databases. A full list of databases that have been included can be found in the user guide of Gene Interactions tool (https://genome.ucsc.edu/goldenPath/help/hgGeneGraph.html). The text-mining supported gene interactions were generated by the Literome machine-reading program, which read and extracted the gene interactions from 20 million PubMed abstracts by the end of 2014 (Poon, Quirk, DeZiel, & Heckerman, 2014). The gene-gene interactions among a given list of genes were visualized through igraph 1.0.0 (http://igraph.org).

3.3.7 Statistical analysis

Linear regression was used to assess the association between genetic variants and fibrosis markers or gene expression levels. Age, gender, body mass index, and the first two genetic principal components were used as covariates in the linear regression model. For GWAS analysis, P value less than 1e-4 were considered as candidate loci for the following analysis. As for the eQTL analysis, P value less than 0.05 were considered as significant for the following analysis. The demographical and histological features of the samples were expressed as number (percentage) for categorical variables and median (interquartile range) for continuous variables. The correlations between MIF gene expression and α-SMA expression and total collagen content were evaluated by Pearson correlation test, and P <0.05 was considered statistically significant. The statistical tests were performed using R 3.4 (https://www.r-project.org).

3.4 Results

3.4.1 GWAS analysis identifies multiple loci affecting ***α***-SMA expression and total collagen content

The workflow of the study is shown in Figure 3.1. The two phenotypes explored by GWAS are positively correlated with each other (r = 0.31, P =6.6e-4). After GWAS, no SNP was identified with a typical genome-wide significance (5e-8) in correlation with each phenotype. At a suggestive level of P <1e-4, there were 73 and 71 candidate genetic variants associated with α-SMA expression and total collagen content, respectively. Although there is a moderate correlation between the two markers, only 3 SNPs (rs1274369, rs1274351, rs1274323) were commonly associated with both markers. The Manhattan plots and Q-Q plots of the GWAS are shown in Figure 3.2, and the characteristics of the top 10 association loci are summarized in Table 3.2.

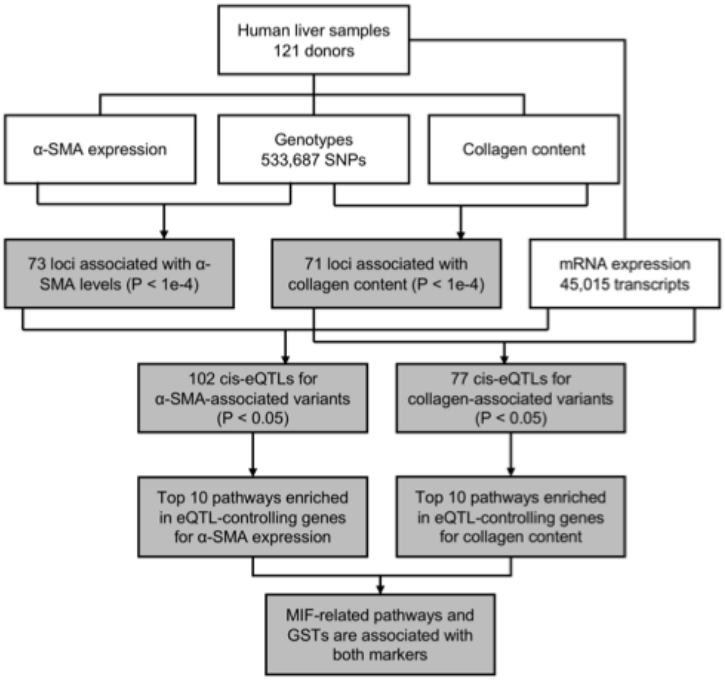


Figure 3.1 Flowchart showing the workflow of the analysis.

White boxes represent the input data of human liver samples. Grey boxes indicate the significant output of analysis. α-SMA=alpha-smooth muscle actin, eQTL=expression quantitative trait loci, GST=glutathione S-transferase, MIF=macrophage migration inhibitory factor, SNP=single nucleotide polymorphism.

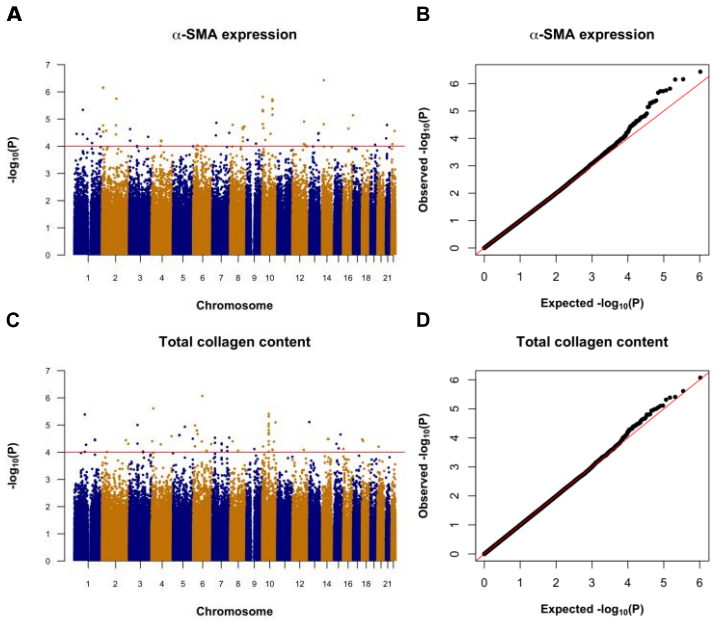


Figure 3.2 Genome-wide association studies of **α**-SMA expression and total collagen content.

(A, B) Manhattan plot and quantile-quantile plot for GWAS of α-SMA expression. (C, D) Manhattan plot and quantile-quantile plot for GWAS of total collagen content. *P* values were calculated by using multiple linear regression model adjusted for age, gender, body mass index, and first two principle components. The horizontal red lines indicate the suggestive significance threshold (P <1e-4) used for further analysis. α-SMA=alpha-smooth muscle actin, GWAS=genome-wide association study.

Top GWAS hits for **α**-SMA expression include an intronic SNP rs8015303 in *BAZ1A* (bromodomain adjacent to zinc finger domain 1A) which encodes a protein subunit of the ATP-dependent chromatin assembly factor that involved in chromatin remodeling. An intronic SNP rs1012580 in the *NOL10* (nucleolar protein 10) was also significantly associated with **α**-SMA expression. A few other top hits indicated that several genes involved in inflammation and immune response especially *IL2RA* (interleukin 2 receptor subunit alpha), *HTR7* (5-hydroxytryptamine receptor 7), *ST6GALNAC3* (ST6 N-acetylgalactosaminide alpha-2,6-sialyltransferase 3) and *AOAH* (acyloxyacyl hydrolase).

Top GWAS hits for total hepatic collagen content are mainly genes that are significantly related to cell skeleton structure, extracellular matrix and cell adhesion. These include *ZWINT* (ZW10 Interacting kinetochore protein) (spindle assembly)(Woo Seo et al., 2015), *BCAR3* (breast cancer anti-estrogen resistance 3) (cytoskeletal remodeling and adhesion)(Wilson, Schrecengost, Guerrero, Thomas, & Bouton, 2013), *EFNA5* (ephrin-A5) (cell adhesion and morphology)(Buensuceso & Deroo, 2013) and *COL11A2* (collagen type XI alpha 2 chain).

Table 3.2 Top 10 GWAS Loci associated with α-SMA expression and total collagen content

SNP	Gene	Locus	Position	A1	A2	MAF	Beta	SE	P
	α-SMA expression								
rs8015303	*BAZ1A*	14q13.2	34316223	G	A	0.06967	-0.7229	0.1313	3.70e-7
rs1012580	*NOL10*	2p25.1	10725199	A	G	0.2869	-0.4247	0.07935	6.96e-7
rs12722561	*IL2RA*	10p15.1	6109899	A	G	0.1557	-0.5144	0.09968	1.54e-6
rs12479413	*AK311291*	2q14.3	129343258	G	A	0.05328	-0.9579	0.1869	1.78e-6
rs1274351	*BC037970*	10q23.31	92237989	A	G	0.123	-0.5068	0.09921	1.90e-6
rs1274323	*HTR7*	10q23.31	92264462	G	A	0.127	-0.5038	0.09925	2.17e-6
rs724999	*ST6GALNAC3*	1p31.1	76961614	A	G	0.07438	-0.8518	0.1741	4.58e-6
rs12207	*TUBB3*	16q24.3	88529861	A	G	0.05785	-0.9667	0.2023	7.23e-6
rs1345546	*NUAK1*	12q23.3	104826063	A	G	0.09426	-0.6404	0.1381	1.23e-5
rs10281171	*AOAH*	7p14.2	36773906	G	A	0.3852	-0.3295	0.07149	1.37e-5
	Total collagen content								
rs6908031	*FAM46A*	6q14.1	81900082	G	A	0.2992	-0.2554	0.04814	8.46e-7
rs2108935	*LDB2*	4p15.32	16453008	G	A	0.2541	-0.2545	0.05043	2.44e-6
rs7089692	*ZWINT*	10q21.1	58603555	G	A	0.2459	-0.2431	0.0493	3.89e-6
rs3950119	*BCAR3*	1p22.1	93821825	G	A	0.2336	0.2674	0.05438	4.10e-6
rs9512950	*TPTE2*	13q12.11	18870233	C	A	0.2992	-0.2572	0.05406	7.74e-6
rs181662	*EMX2*	10q26.11	119372445	G	A	0.3279	0.2253	0.04743	7.97e-6
rs13092046	*CNTN3*	3p12.3	74617642	A	G	0.4262	0.2266	0.04828	9.97e-6
rs2237172	*ATXN1*	6p22.3	16709566	G	A	0.3512	-0.2444	0.05214	1.03e-5
rs6894788	*EFNA5*	5q21.3	106875173	G	A	0.2645	0.2476	0.05317	1.16e-5
rs9368758	*COL11A2*	6p21.32	33245999	A	G	0.06557	-0.4443	0.09714	1.58e-5

The single nucleotide polymorphism is mapped to its nearest gene; A1 is the minor allele, and A2 is the major allele; All positions refer to hg18. α-SMA=alpha-smooth muscle actin, MAF=minor allele frequencies, SE=standard error.

3.4.2 Expression analysis identifies significant eQTLs for lead variants

The eQTL analysis can help establish the potential causality for the GWAS findings. To further explore the effects of GWAS loci on gene expression, we performed a *cis*-eQTL analyses in the same set of liver tissues as the GWAS, by focusing on genes that are within the distance of ±1 Mbp to GWAS identified candidate loci (P <1e-4). By accepting a liberal significance level of P =0.05 for eQTLs, we identified 102 significant eQTLs for 44 α-SMA-associated variants and 77 eQTLs for 44 candidate GWAS loci associated with collagen content.

Our results show that several variants are associated with the mRNA expression of their nearest genes. For example, rs1012580 in nucleolar protein 10 (*NOL10*) for **α**-SMA expression and rs13092046 in contactin 3 (*CNTN3*) for total collagen content are found to be correlated with the mRNA expression levels of their most nearby genes. However, other significant eQTLs may exert their effects on gene expression in a relatively broad range. For instance, fibrogenesis candidate variant rs12207 at *TUBB3* locus is associated with the expression levels of several nearby genes including charged multivesicular body protein 1A (*CHMP1A*), fanconi anemia complementation group A (*FANCA*), VPS9 domain containing 1 (*VPS9D1*), and paraplegin matrix AAA peptidase subunit (*SPG7*) instead of *TUBB3* itself.

Top eQTLs for α-SMA expression include several variants in 22q11.23 locus that are significantly associated with the transcription levels of macrophage migration inhibitory factor (*MIF*) and glutathione S-transferase theta 2 (*GSTT2*). Top eQTL-controlling genes for total collagen content include two glycosyltransferases encoding genes, namely solute carrier family 35 member B4 (*SLC35B4*) and beta-1,3-galactosyltransferase 4 (*B3GALT4*). The transcription of prostaglandin-endoperoxide synthase 2 (*PTGS2*), which encodes a cyclooxygenase involved in

the MIF-related apoptosis repression, is significantly associated with collagen-related variants in 1q31.1 locus.

3.4.3 Pathway enrichment analysis of the eQTL-controlling genes

Given the polygenic nature of liver fibrosis, a single genetic variant or gene may only possess limited effect on the development and progression of the disease. We aim to identify pathways underlying the liver fibrosis that individual SNP-based GWAS analysis might miss. We conducted enrichment analyses on the candidate eQTL-controlling genes using the ingenuity pathway analysis package. In order to avoid missing key pathways, we assumed that those genes whose transcriptions are associated with the candidate SNPs at a nominal $P < 0.05$ level are the candidate genes for an enrichment analysis. As shown in Table 3.3, eumelanin biosynthesis, ataxia telangiectasia mutated signaling, glutathione redox reactions I, vascular endothelial growth factor signaling, glutathione-mediated detoxification, MIF-mediated glucocorticoid regulation, and MIF regulation of innate immunity pathways are top enriched pathways for the α-SMA activation. On the other hand, MIF-mediated glucocorticoid regulation, MIF regulation of innate immunity, branched-chain α-keto acid dehydrogenase complex, and glutathione ascorbate recycling pathways are enriched in candidate genes for total collage content.

Table 3.3 Results from ingenuity pathway analysis analyses

Item	Ingenuity canonical pathways	*P*	Molecules
α-SMA expression		7.079e-5	
	Eumelanin biosynthesis		MIF, DDT
	ATM signaling	1.318e-3	MDM4, RNF8, ZNF420
	Glutathione redox reactions I	1.585e-3	GSTT2/GSTT2B, GSTT1
	Vascular endothelial growth factor signaling	2.692e-3	ROCK2, VCL, NOS3
	Glutathione-mediated detoxification	2.754e-3	GSTT2/GSTT2B, GSTT1
	MIF-mediated glucocorticoid regulation	3.548e-3	MIF, PLA2G2D
	MIF regulation of innate immunity	5.495e-3	MIF, PLA2G2D
	Actin nucleation by ARP-WASP complex	1.000e-2	ROCK2, ARPC4
	Citrulline-nitric oxide cycle	1.349e-2	NOS3
	Remodeling of epithelial adherens junctions	1.445e-2	VCL, ARPC4
Total collagen accumulation		2.884e-3	
	MIF-mediated glucocorticoid regulation		PLA2G4A, PTGS2
	MIF regulation of innate immunity	4.365e-3	PLA2G4A, PTGS2
	Branched-chain α-keto acid dehydrogenase complex	9.550e-3	BCKDHB
	Ascorbate recycling (cytosolic)	9.550e-3	GSTO2
	Endothelin-1 signaling	1.047e-2	PLA2G4A, ADCY5, PTGS2
	Eicosanoid signaling	1.148e-2	PLA2G4A, PTGS2
	Role of MAPK signaling in the pathogenesis of influenza	1.318e-2	PLA2G4A, PTGS2
	Estrogen-dependent breast cancer signaling	1.479e-2	IGF1, HSD17B8
	Adenine and adenosine salvage III	1.660e-2	ADAL
	Purine ribonucleosides degradation to ribose-1-phosphate	1.905e-2	ADAL

α-SMA=alpha-smooth muscle actin, ATM=ataxia telangiectasia mutated, MAPK=mitogen-activated protein kinase, MIF=macrophage migration inhibitory factor.

Interestingly, although there is only a minimal overlap between the candidate genes associated with the two phenotypes, the MIF related pathways as well as glutathione-S-transferases (GSTs) are significantly enriched for both markers. A detailed inspection of the *MIF* gene SNPs identifies several variants near the *MIF* locus to be negatively associated with α-SMA levels (β = -0.43; 95%CI: -0.64, -0.23; P =8.38e-5) (Figure 3.3 B), but positively associated with *MIF* gene expression (β = 0.37; 95%CI: 0.30, 0.44; P = 6.34e-7) (Figure 3.3 C). We found that *MIF* gene expression is significantly correlated with decreased levels of α-SMA (Figure 3.4 A, r = -0.21, P =0.026), suggesting that variations in *MIF* locus might affect the susceptibility of fibrogenesis through controlling *MIF* gene expression. There is no significant correlation between *MIF* gene expression and total collagen content (Figure 3.4 B, P =0.29).

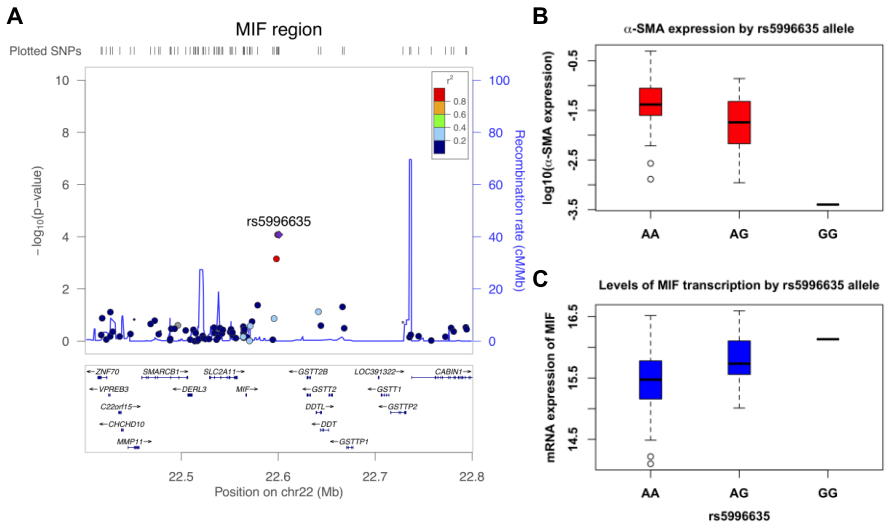


Figure 3.3 Association between variants near MIF locus and MIF gene expression and **α**-SMA levels.

(A) Regional plot for MIF locus. Each SNP is plotted with its P value (shown as $-\log_{10}$ P value) as a function of its genomic coordinate (Hg 18). The local LD structure was estimated based on 1000 Genomes 2010 CEU; B: Negative association between *MIF* rs5996635 and **α**-SMA expression levels (β = -0.43; 95%CI: -0.64, -0.23; P =8.38e-5). P values were calculated by using multiple linear regression model adjusted for age, gender, BMI, and first two principle components. (C) Positive association between *MIF* rs5996635 and MIF transcription (β = 0.37; 95%CI: 0.30, 0.44; P =6.34e-7). A multiple linear regression model was used to test the association between MIF expression levels and nearby GWAS loci within ±1 Mbp region. Age, gender, body mass index, and the first two genetic principal components were used as covariates. Numbers of individuals with AA, AG, and GG genotypes are 88, 32, and 1, respectively. α-SMA=alpha-smooth muscle actin, MIF=macrophage migration inhibitory factor, SNP=single nucleotide polymorphism.

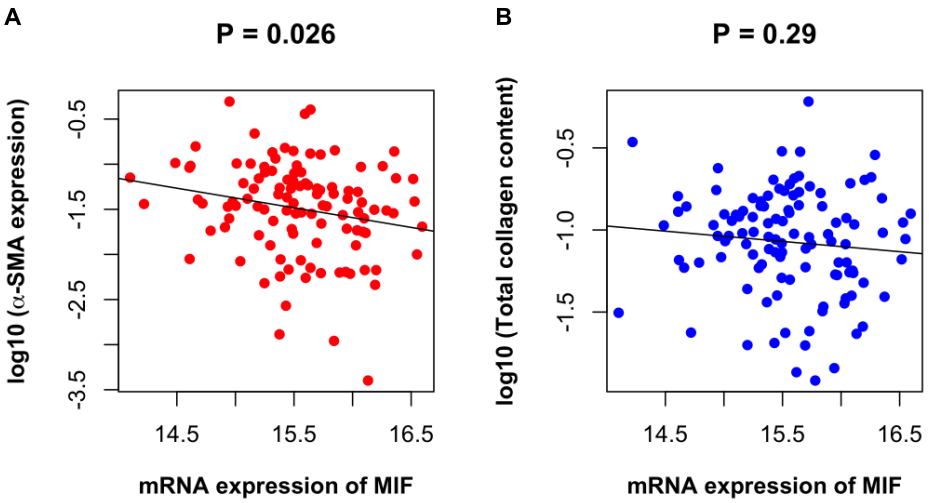


Figure 3.4 Correlation between MIF gene expression and **α**-SMA expression and total collagen content.

(A) Negative correlation (r = -0.21, P = 0.026) between *MIF* gene expression and α-SMA expression. (B) Correlation (r = -0.10, P = 0.29) between *MIF* gene expression and total collagen content. P values were calculated by Pearson correlation test. α-SMA=alpha-smooth muscle actin, MIF=macrophage migration inhibitory factor.

To further confirm this pathway enrichment, and to explore the possible mechanisms through which eQTL-controlling genes are involved in the development of fibrosis, we searched for the gene-gene interactions among candidate genes for α-SMA expression and total collagen content based on a text-mining based gene interaction database. Two genes were considered to be interacted as long as this relationship has been supported by either curated gene interactions databases or text-mining evidence. As shown in Figure 3.5 A, nitric oxide synthase 3 (*NOS3*) and *MIF* are the two genes that have the largest number of connections with other α-SMA expression-related genes. *PTGS2*, which mediates the MIF induced apoptosis suppression, is one of the most highly connected genes for the total collagen content-related gene network (Figure 3.5 B).

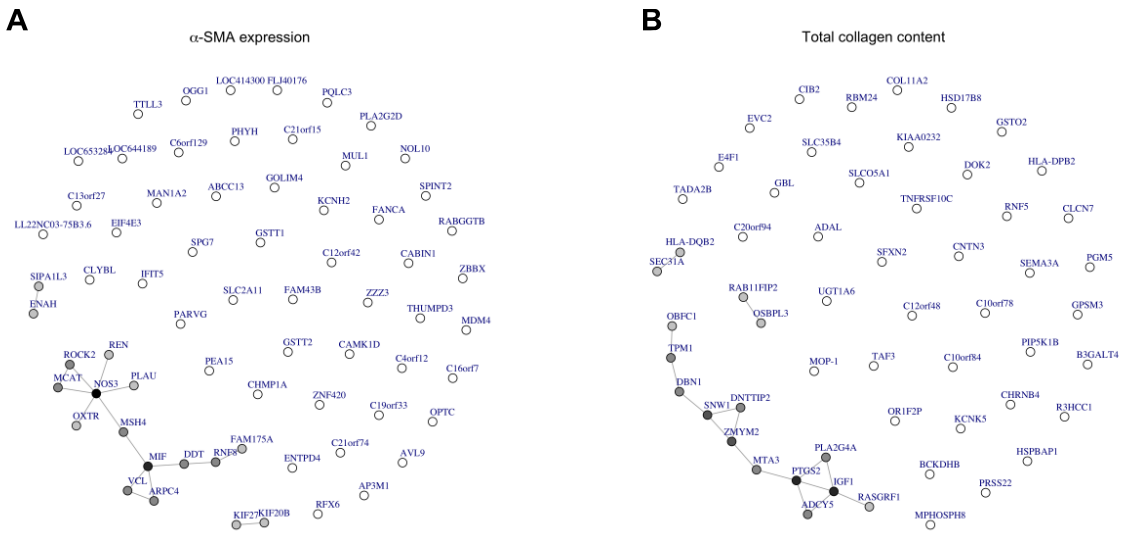


Figure 3.5 Gene interaction networks for eQTL-controlling genes.

(A) Gene interactions for α-SMA expression-related genes. (B) Gene interactions for total collagen content-related genes. The grayscale intensity of the node represents the number of its edges. The darker the node is, the more connections it has. α-SMA=alpha-smooth muscle actin, eQTL=expression quantitative trait loci.

3.5 Discussions

Hepatic collagens accumulation and **α**-SMA activation are reliable and quantitative markers indicating the risk of developing liver fibrosis and cirrhosis (Y. Huang et al., 2013; Yamaoka, Nouchi, Marumo, & Sato, 1993), thus are important intermediate phenotypes for the disease. Interrogating the genetic susceptibility loci for these intermediate phenotypes will gain insight into the molecular mechanisms involved in the pathogenesis of hepatic fibrosis and provide potential targets for early diagnosis and treatment. Our study for the first time identifies candidate genetic variants, genes, and pathways contributing to human liver fibrogenesis and fibrosis in a general population.

GWAS have been widely used to investigate the genetic basis of human complex diseases, and thousands of susceptibility loci have been identified thus far (Welter et al., 2014). However, to

date most GWAS identified alleles only account for a modest proportion of total variance in traits. This is mainly because of the polygenic nature of complex traits and at least in part due to the multiple testing burden in test statistics (Visscher, Brown, McCarthy, & Yang, 2012). A growing body of knowledge acknowledges that there are causal variants remain undetected owe to the adoption of stringent genome-wide significance threshold level (P <5e-8)(Amin, van Duijn, & Janssens, 2009; Sveinbjornsson et al., 2016). Therefore, we set a relatively moderate threshold at P <1e-4 for GWAS to systematically evaluate the overall effects of candidate genetic variants on the two phenotypes. In addition, incorporating the information of eQTLs and evaluating them in a network way is beneficial to interpret the biological mechanisms underlying the discovered loci, thus potentially identifying causal alleles underlying the genetic association. As such, although our study did not identify any SNP that reaches the typical GWAS significance level, which could be largely attributed to the small sample size, our combined analyses indeed narrowed down a few interesting genes and pathways that are broadly supported by many previous studies.

Our analyses firstly demonstrated that a few genes involved in inflammation and immune response as top GWAS hits for **α**-SMA activation. This is consistent with the known relationship between **α**-SMA expression and hepatic stellate cell activation as a key step for liver fibrogenesis. One notable loci is *IL2RA* (CD25) across which multiple SNPs were previously associated with various inflammatory disorders and traits including allergy (Hinds et al., 2013), Epstein-Barr virus nuclear antigen-1 (EBNA-1) IgG level (Y. Zhou et al., 2016), diisocyanate-induced asthma (Yucesoy et al., 2015), inflammatory bowel disease (J. Z. Liu et al., 2015), autoimmune diseases (Jin et al., 2010; Y. R. Li et al., 2015), rheumatoid arthritis (Orozco et al.,

2014), multiple sclerosis (International Multiple Sclerosis Genetics Consortium et al., 2011; Patsopoulos et al., 2011), and levels of autoantibodies in type 1 diabetes (Plagnol et al., 2011). The top hit rs12722561 identified in our study is in complete linkage disequilibrium with rs12722489, a polymorphism previously identified in a genome-wide meta-analyses as a susceptibility locus to multiple sclerosis (Patsopoulos et al., 2011) and Crohn's disease (Franke et al., 2010). A recent study has shown that variant in *IL2RA* was also significantly associated with the concentration of circulating IL2RA (Ahola-Olli et al., 2017). Importantly, the IL-2 signaling pathway was also enriched as one of the top pathways associated advanced fibrosis and cirrhosis in a genome-wide pathway analysis for non-alcoholic fatty liver disease (Q. R. Chen et al., 2013). Mechanistically, the IL-2/IL2RA signaling has been indicated to be involved in the liver fibrosis as well. The stellate cell-lymphocyte interaction in the liver plays a pivotal role in stellate cell activation and liver fibrogenesis, and the IL2RA (CD25) and CD4 positive regulatory T cells ($CD4^+/CD25^+$) has been demonstrated to be anti-fibrotic by suppression the pro-fibrotic effect of CD8 cells on stellate cells (Horani et al., 2007).

It is also not surprising that top GWAS hits for total collagen content are associated with multiple genes involved in cytoskeletal structure, extracellular matrix, cell migration and adhesion. In addition, two highly linked intronic SNPs at the *COL11A2* (collagen type XI alpha 2 chain) locus are strongly associated with the total collagen content as well. Interestingly, these two SNPs are also significantly associated with mRNA expression of multiple genes within a ~250 kb region in our liver tissue set including *COL11A2* and multiple *HLA* genes. We further searched the genotype-tissue expression portal for these two SNPs, which turns out that they are also strong eQTLs for the similar set of the genes within the region among multiple tissues.

Notably, this *HLA* locus has been previously linked with hepatitis B virus/hepatitis C virus-related liver cirrhosis (Mangia et al., 1999; Zhang et al., 2015) as well as primary biliary cirrhosis (Almasio et al., 2016), suggesting an essential role of these genes in increasing the susceptibility to liver fibrosis.

It is known that the typical GWAS approach may miss important genetic loci as only the very top significant SNPs are selected. We therefore further performed a pathway enrichment analysis by focusing on the genes whose mRNA expression is likely to be affected by the candidate SNPs that are associated with the two markers using a liberal cut-off of P <1e-4. Interestingly, the glutathione-related genes and MIF-related pathways are the two common major pathways enriched for both markers. It should be noted that there is only a minimal overlap between the candidate genes associated with each of the two markers, suggesting that despite different genes, the underlying pathways still stand out. While it is known that GSTs are protective for oxidative stress-mediated liver damage (S. Li et al., 2015). MIF related signaling pathway has been strongly linked to liver fibrosis as well. A recent study demonstrated that MIF-deleted mice ($Mif^{-/-}$) tend to show exaggerated fibrogenic phenotypes in two chronic liver injury models (Heinrichs et al., 2011). This study is consistent with our finding that MIF may exert anti-fibrotic effects in human livers. Moreover, other eQTL-controlling genes in MIF pathway, including *PLA2* and *PTGS2*, are key mediators of MIF induced apoptosis suppression signaling, indicating apoptosis repression might be one of the mechanisms for the antifibrotic effects of MIF (Elsharkawy, Oakley, & Mann, 2005).

In addition, angiogenesis has been significantly involved in liver fibrosis (Srivastava et al., 2018; Thomas, 2018). Indeed, our pathway enrichment analysis identified that NOS3 and vascular endothelial growth factor signaling pathways are significantly associated with α-SMA expression, while the endothelin-1 signaling pathway is significantly pathway is associated with total collagen level. A considerable body of evidence demonstrates that vascular endothelial growth factor signaling promotes liver fibrogenesis by stimulating activated stellate cells growth, migration, and collagen production (Novo et al., 2007; Yoshiji et al., 2003). It has been also shown that *NOS3* expression and activation plays a critical role in the development of FLD and liver fibrosis (Leung et al., 2008; Persico et al., 2017; Sheldon, Laughlin, & Rector, 2014).

We further explore the potential interaction network between the key genes and pathways. Again, the *NOS3* and *MIF* are the most highly connected genes for the α-SMA expression gene network. Our results suggested that the two hub genes, *NOS3* and MIF, and their related genes were connected through MutS homolog 4 (*MSH4*), suggesting that the crosstalk between NOS3 and MIF signaling pathways might be critical for the liver fibrogenesis. However, this hypothesis needs to be validated through further investigations. As for the gene interaction network of the total collagen content related genes, *PTGS2* has the largest number of connections. Again, it is notable that PTGS2 is also involved in both the MIF and endothelin-1 signaling pathway, which may further indicate that the MIF and angiogenesis signaling are both involved in liver fibrosis.

It should be noted again that the interaction networks for both phenotypes highlighted the MIF signaling pathway although there is limited overlap between eQTL-controlling genes for the two phenotypes. Therefore, the role of MIF signaling, especially its potential interaction with the angiogenesis pathway needs to be further investigated.

There are also several limitations of our study. Due to the moderate sample size, our analyses are limited in its power for identifying genetic risk loci. In addition, we adopted a generous threshold to incorporate more potential causal variants into analysis. This will inevitably increase the false positive rate of our test. Therefore, the susceptibility loci identified in our study should be considered as suggestive and need to be validated independently and within more diverse cohorts.

In conclusion, our study identified candidate genetic variants and pathways significantly associated with intermediate markers for liver fibrogenesis and fibrosis. Our findings would be helpful to elucidate the genetic basis underlying the inter-individual differences in the development of liver fibrosis and provide candidate targets for developing therapeutic strategies.

CHAPTER 4. MENDELIAN RANDOMIZATION ANALYSIS DISSECTS THE RELATIONSHIP BETWEEN NAFLD, T2D, AND OBESITY AND PROVIDES IMPLICATIONS TO PRECISION MEDICINE

4.1 Disclaimer

The materials of this chapter were adapted from the author's preprint at the bioRxiv database (https://www.biorxiv.org/content/10.1101/657734v1) (Z. Liu, Zhang, et al., 2019). The copyright holds by the author. The GWAS analysis in the UK biobank samples (section 4.3.2.2) were performed by Sarah Graham from the University of Michigan, the transgenic animal experiments (section 4.4.5) were finished by Yang Zhang from the Wayne State University.

4.2 Introduction

Non–alcoholic Fatty Liver Disease (NAFLD) is characterized by the presence of excess hepatic fat accumulation ($\geq 5\%$) without significant alcohol use, hepatitis virus infection, or other secondary causes of hepatic fat accumulation (Chalasani et al., 2018). The spectrum of NAFLD ranges from simple non–alcoholic fatty liver (NAFL) to non–alcoholic steatohepatitis (NASH), which over time can lead to cirrhosis, hepatocellular carcinoma, and organ failure(Hardy, Oakley, Anstee, & Day, 2016). Compelling observational epidemiological studies have shown that NAFLD is highly correlated with metabolic disorders such as type 2 diabetes (T2D)(M. Hu, Phan, Bourron, Ferre, & Foufelle, 2017; Lonardo et al., 2019; Mantovani, Byrne, Bonora, & Targher, 2018) and obesity (Chalasani et al., 2018; Fabbrini, Sullivan, & Klein, 2010; Loomis et al., 2016). All three diseases together affect over 50% of the U.S. population (Hruby & Hu, 2015; Xu et al., 2018; Younossi et al., 2018).

Dissecting the causal relationship between the three diseases is crucial for both understanding the disease etiology and developing effective therapeutic or preventive strategies. However, observational associations are limited in elucidating the causality due to various confounding factors (e.g. lifestyle, socioeconomic status) or reverse causation bias (Lawlor, Harbord, Sterne, Timpson, & Smith, 2008). As a result, the three diseases are often treated as comorbidities for each other in various biomedical research settings. On the other hand, current prevention and treatment strategies focus on each of the three diseases individually. Without clearly knowing the causality among the three diseases, treatments of preventive interventions may often lead to conflicting research findings and inconsistent responses to disease prevention and treatment among patients.

Mendelian randomization (MR) analysis, which uses genetic variants as proxies for the risk factors of interest, has been widely applied in understanding the causal relationship between various risks factors and human diseases, e.g. estimation of the causal effect of plasma HDL cholesterol on myocardial infarction risk (Voight et al., 2012). Since during the process of meiosis the alleles of the parents are randomly segregated to the offspring, the MR method is considered to be analogous to a randomized controlled trial (RCT) but less likely to be influenced by confounding factors and reverse causation (Paternoster, Tilling, & Smith, 2017). Bidirectional MR is an extension of traditional MR in which the exposure–outcome causal relationship was explored from both sides. The bidirectional framework provides an efficient way to ascertain the direction of a causal relationship, which helps alleviate the potential bias from reverse causation (Welsh et al., 2010).

Recent MR studies have partially explored the causal relationships among the three diseases. Dongiovanni et al. showed genetically instrumented hepatic steatosis was associated with insulin resistance and a small increase in T2D risk (Dongiovanni et al., 2018). A study by De Silva et al. indicated that genetically raised circulating ALT and AST increased the risk for T2D (De Silva et al., 2019). Stender et al. found that the genetic predictors of BMI were associated with increased hepatic triglyceride content (Stender et al., 2017). However, a systematic bidirectional MR study leveraging the latest GWAS data is particularly needed to elucidate the causal relationships among the three metabolic diseases under a uniformed setting. In addition, experimental analysis e.g. animal models with a characterized natural history under controlled conditions would also help further establish the causality.

In this study, we first aimed to explore whether NAFLD casually increases risks for T2D, obesity and their related intermediate traits. We then investigated the reverse relationships, i.e. whether T2D and obesity causally affect NAFLD risk. Further, we constructed a transgenic mice model expressing human PNPLA3 isoforms, a known genetic NAFLD model to test the causal effects of hepatic perturbations on T2D and obesity.

4.3 Methods and materials

4.3.1 Ethics statement

The summary–level GWAS data used for mendelian randomization analyses are publicly available (Figure 4.1 and Table 4.1) (Dupuis et al., 2010; Mahajan et al., 2018; Manning et al., 2012; Prokopenko et al., 2014; Pulit et al., 2019; Saxena et al., 2010; Speliotes et al., 2011; Strawbridge et al., 2011; Wheeler et al., 2017; Willer et al., 2013). Therefore, no specific ethical

approval is required. The study of the transgenic mice experiments has been reviewed and approved by the IACUC of the Indiana University School of Medicine. This research has been conducted using the UK Biobank Resource under application number 24460.

4.3.2 MR analyses

4.3.2.1 GWAS summary data

The summary statistics of association with computerized tomography (CT) measured hepatic steatosis were taken from a meta–analysis of 7,176 individuals by the Genetics of Obesity–related Liver Disease (GOLD) consortium (Speliotes et al., 2011). The results of the association with histologic NAFLD were taken from the same study, in which 592 biopsy proven NAFLD patients from NASH Clinical Research Network (NASH CRN) and 1,405 controls from Myocardial Infarction Genetics Consortium (MIGen) were involved. The full GWAS summary data of this study are not publicly available, therefore only the results of the top GWAS loci associated with steatosis and histologic NAFLD were extracted. The full GWAS summary statistics of NAFLD were generated in UK Biobank (UKBB) samples consisting of 1,122 cases and 399,900 controls. The details of the GWAS of NAFLD in UKBB are described in the section below.

We downloaded full GWAS summary data of 22 glycemic and obesity traits from the largest published studies as of March 2019. These traits include T2D, glycated hemoglobin A1c (HbA1c), fasting glucose, fasting insulin, fasting proinsulin, 2–h glucose, homeostatic model assessment of insulin resistance (HOMA–IR), β–cell function (HOMA–B) and seven insulin secretion and action indices during oral glucose tolerance test (OGTT) including area under the curve of insulin levels (AUCins), ratio of AUC insulin and AUC glucose (AUCins/AUCgluc),

incremental insulin at 30 min (Incre30), insulin response to glucose during the first 30 min adjusted for BMI (Ins30adjBMI), insulin sensitivity index (ISI), corrected insulin response (CIRadjISI), disposition index (DI), body mass index (BMI), waist–hip ratio (WHR), WHR adjusted for BMI (WHRadjBMI), and four plasma lipid levels. Only association results from participants of European descent were used in the present study. The details on the phenotype information, sample size, and PubMed ID of the original study are summarized in Table 4.1.

Table 4.1 Characteristics of the GWAS summary data

Phenotype	Participants	Data source	Phenotype transformation	Units of effect size	Phenotype description	PubMed ID
NAFLD						
Steatosis	7,176	GOLD	inverse normally transformation	SD	Computerized tomography measured hepatic steatosis	21423719
Histologic NAFLD	592 cases and 1,405 controls	NASH CRN/MIGen	none	OR	Biopsy–proven NAFLD	21423719
NAFLD*	1,122 cases 399,900 controls	UKBB	none	OR	NAFLD from UK Biobank database	Not published
T2D and glycemic traits						
T2D	74,124 cases and 824,006 controls	DIAGRAM	none	OR	Type 2 diabetes adjusted for body mass index (BMI)	30297969
HbA1c	123,665	MAGIC	none	%	Glycated hemoglobin	28898252
Fasting glucose	58,074	MAGIC	none	mmol/L	Fasting glucose adjusted for BMI	22581228
Fasting insulin	51,750	MAGIC	log transformation	pmol/L	Fasting insulin adjusted for BMI	22581228
Fasting proinsulin	10,701	MAGIC	log transformation	pmol/L	Fasting proinsulin adjusted for fasting insulin	21873549
2h glucose	15,234	MAGIC	none	mmol/L	2 h glucose levels after an oral glucose challenge adjusted for BMI	20081857

Table 4.1 continued

HOMA–IR	37,037	MAGIC	log transformation	original	The homeostatic model assessment (HOMA) insulin resistance	20081858
HOMA–B	36,466	MAGIC	log transformation	original	HOMA beta cell function	20081858
AUCins	4,324	MAGIC	log transformation	mU*min/L	Area under the curve (AUC) of insulin levels during OGTT	24699409
AUCins/AUCgluc	4,213	MAGIC	log transformation	mU/mmol	Ratio of AUC insulin and AUC glucose	24699409
Incre30	4,447	MAGIC	log transformation	mU/L	Incremental insulin at 30 min	24699409
Ins30adjBMI	4,409	MAGIC	log transformation	original	Insulin response to glucose during the first 30 min adjusted for BMI =insulin at 30 min/ (glucose at 30 min×BMI)	24699409
ISI	4,769	MAGIC	log transformation	mg/dL	Insulin sensitivity index =10,000/square root (fasting plasma glucose (mg/dl) ×fasting insulin×mean glucose during OGTT (mg/dl) ×mean insulin during OGTT)	24699409
CIRadjISI	4,789	MAGIC	log transformation	original	Corrected Insulin Response =(100 × insulin at 30 min)/ (glucose at 30 min × (glucose at 30 min– 3.89)), adjusted for ISI	24699409
DI	5,130	MAGIC	log transformation	original	Disposition index =CIR×ISI	24699409
Obesity traits						
BMI	806,834	GIANT	inverse normally transformation	SD	Body mass index	30239722
WHR	697,734	GIANT	inverse normally transformation	SD	Waist–hip ratio	30239722
WHRadjBMI	694,649	GIANT	inverse normally transformation	SD	Waist–hip ratio adjusted for BMI	30239722
HDL	188,577	GLGC	quantile normalization	SD	Plasma high–density lipoprotein cholesterol	24097068
LDL	188,577	GLGC	quantile normalization	SD	Plasma low–density lipoprotein cholesterol	24097068

Table 4.1 continued

TC	188,577	GLGC	quantile normalization	SD	Plasma total cholesterol	24097068
TG	188,577	GLGC	quantile normalization	SD	Plasma triglycerides	24097068

*NAFLD was defined based on ICD–9 571.8 "Other chronic nonalcoholic liver disease") and ICD– 10 K76.0 ["Fatty (change of) liver, not elsewhere classified"] from inpatient hospital diagnosis within the UK Biobank dataset. Individuals with Hepatitis B or C or with other known liver diseases (e.g. liver transplant, hepatomegaly, jaundice, or abnormal liver function study results) were excluded from the analysis.
GOLD: Genetics of Obesity–related Liver Disease; NASH CRN: NASH Clinical Research Network; MIGen: Myocardial Infarction Genetics consortium; UKBB: UK Biobank; DIAGRAM: DIAbetes Genetics Replication And Meta–analysis; MAGIC: Meta–Analyses of Glucose and Insulin–related traits Consortium; GIANT: The Genetic Investigation of ANthropometric Traits consortium; GLGC: The Global Lipids Genetics Consortium; OR: odds ratio; SD: standard deviation.

4.3.2.2 GWAS of NAFLD and T2D in UK Biobank

NAFLD was defined based on ICD–9 571.8 "Other chronic nonalcoholic liver disease" and ICD–10 K76.0 ["Fatty (change of) liver, not elsewhere classified"] from inpatient hospital diagnosis within the UK Biobank dataset. Individuals with Hepatitis B or C or with other known liver diseases (e.g. liver transplant, hepatomegaly, jaundice, or abnormal liver function study results) were excluded from the analysis. The white British subset of UK Biobank was used for analysis in SAIGE (W. Zhou et al., 2018) with sex, birth year, and 4 principle components as covariates.

4.3.2.3 Genetic predictors selection

We used the two strongest genetic predictors of NAFLD, Patatin–like phospholipase domain–containing protein 3 (*PNPLA3*) rs738409 and Transmembrane 6 superfamily member 2 (*TM6SF2*) rs58542926, as the proxy for hepatic steatosis and histologic NAFLD. These two variants have been consistently shown to be associated with the whole spectrum of the NAFLD (Eslam, Valenti, & Romeo, 2018; Kozlitina et al., 2014; Y. L. Liu et al., 2014; Romeo et al., 2008; Speliotes et al., 2011; X. Wang et al., 2016). Multiple MR analyses have previously been

performed using these two variants to test the causal relationship between NAFLD and diseases such as Vitamin D deficiency(N. Wang et al., 2018), ischaemic heart disease (Lauridsen et al., 2018), and liver damage (Dongiovanni et al., 2018). As *TM6SF2* rs58542926 is not genotyped in most of the GWAS summary data used in this study, rs2228603 at the *NCAN* gene locus, which is in strong linkage disequilibrium with *TM6SF2* rs58542926 (pairwise $R^2 = 0.76$ based on the Phase 3 data of the 1000 Genomes Project in European individuals) and significantly associated with both hepatic fat content and NAFLD histology (Gorden et al., 2013), was used as the proxy for *TM6SF2* rs58542926.

For the 22 glycemic and obesity traits, we selected the significant and independent genetic predictors in two steps: we first obtained all the variants that passed the genome–wide association significance level of $p<5e–8$. Then the independent genetic predictors were identified by clumping the top GWAS loci through PLINK 1.9 (https://www.cog–genomics.org/plink2) with the threshold of R^2 < 0.1 in a 500–kb window. The linkage disequilibrium was estimated based on the European samples in phase 3 of the 1000 Genomes Project (Genomes Project et al., 2015). We omitted 10 traits including 2h glucose, HOMA–IR, HOMA–B, and seven OGTT traits due to lack of enough significant and independent genetic variants (number of valid variants < 3). Therefore, 12 traits were analyzed for the causal effects on NAFLD.

The instrument strength of each genetic predictor was assessed by F statistics = $(\frac{n-k-1}{k})(\frac{R^2}{1-R^2})$, where n represents sample size, k represents the number of genetic variants, and R^2 is the proportion of phenotypic variance explained by the genetic variants. The F statistics of all the

genetic predictors in the present study were larger than the empirical threshold of 10 (Table 4.2)(Staiger & Stock, 1997).

Table 4.2 Characteristics of the associations of PNPLA3 rs738409, NCAN rs2228603 with phenotype

Phenotype	**Participants**	***PNPLA3* rs738409 G allele**			***NCAN* rs2228603 T allele**		
		Effect	**SE**	***p***	**Effect**	**SE**	***p***
NAFLD							
Steatosis (SD)	7,176	0.26	0.021	4.3E–34	0.24	0.035	1.22E–11
Histologic NAFLD (logOR)	592 cases and 1,405 controls	1.18	0.31	3.6E–43	0.5	0.23	5.29E–05
T2D and glycemic traits							
T2D (OR)	74,124 cases and 824,006 controls	0.062	0.0089	3.5E–12	0.086	0.014	1.30E–09
HbA1c (%)	123,665	–0.0037	0.003	2.2E–01	0.0007	0.0033	8.33E–01
Fasting glucose (mmol/L)	58,074	0.0029	0.0038	4.4E–01	0.02	0.0063	1.77E–03
Fasting insulin (pmol/L)	51,750	0.0077	0.0032	1.6E–02	0.0024	0.0053	6.49E–01
Fasting proinsulin (pmol/L)	10,701	0.004	0.0087	6.4E–01	0.02	0.014	1.44E–01
2h glucose (mmol/L)	15,234	–0.016	0.022	4.7E–01	–0.019	0.036	6.09E–01
HOMA–IR ((mU/L)*(mmol/L))	37,037	0.007	0.0047	1.4E–01	0.0097	0.0081	2.33E–01
HOMA–B ((mU/L)/(mmol/L))	36,466	0.0007	0.0039	8.5E–01	0.0053	0.0066	4.26E–01
AUCins (mU*min/L)	4,324	–0.0066	0.026	8.00E–01	0.0079	0.044	8.56E–01
AUCins/AUCgluc (mU/mmol)	4,213	–0.0057	0.026	8.30E–01	0.0073	0.044	8.69E–01
Incre30 (mU/L)	4,447	–0.00015	0.025	9.95E–01	–0.0018	0.043	9.66E–01
Ins30adjBMI	4,409	–0.0027	0.026	9.16E–01	0.026	0.043	5.50E–01
ISI (mg/dL)	4,769	0.03	0.029	2.94E–01	–0.066	0.043	1.27E–01
CIRadjISI	4,789	–0.0062	0.026	8.09E–01	0.013	0.043	7.60E–01
DI	5,130	0.03	0.024	2.15E–01	–0.012	0.041	7.62E–01
Obesity traits							
BMI (SD)	806,834	–0.0082	0.0021	7.2E–05	–0.004	0.0031	1.98E–01
WHR (SD)	697,734	0.0029	0.0022	1.8E–01	0.0134	0.0033	3.60E–05

Table 4.2 continued

WHRadjBMI (SD)	694,649	0.0078	0.0022	3.5E–04	0.0178	0.0033	5.95E–08
HDL (SD)	188,577	–0.016	0.0058	9.8E–03	0.0054	0.0067	4.51E–01
LDL (SD)	188,577	–0.0151	0.0063	2.0E–02	–0.104	0.0072	4.43E–44
TC (SD)	188,577	–0.0219	0.0062	6.0E–04	–0.12	0.0069	1.05E–62
TG (SD)	188,577	0.0051	0.0056	2.5E–01	–0.11	0.0065	1.74E–57

T2D: type 2 diabetes; HOMA–IR: The homeostatic model assessment (HOMA) insulin resistance; HOMA–B: HOMA beta cell function; AUCins: area under the curve (AUC) of insulin levels during oral glucose tolerance test (OGTT); AUCins/AUCgluc: ratio of AUC insulin and AUC glucose; Incre30: incremental insulin at 30 min; Ins30adjBMI: Insulin response to glucose during the first 30 min adjusted for BMI; ISI: insulin sensitivity index; CIRadjBMI: Corrected Insulin Response adjusted for ISI; DI: disposition index; BMI: body mass index; WHR: waist–hip ratio; WHRadjBMI: WHR adjusted for BMI; HDL: high–density lipoprotein cholesterol; LDL: low–density lipoprotein cholesterol; TC: total cholesterol; TG: triglycerides; OR: odds ratio; SD: standard deviation; SE: standard error.

4.3.2.4 MR estimation

We set the exposure increasing allele as the effect allele. If the effect size was reported on the alternative allele in the original study, we multiplied the reported effect size by –1 for harmonization of the effect allele. For palindromic SNPs, we checked the reported allele frequency to avoid potential strand flipping issues (Hartwig, Davies, Hemani, & Davey Smith, 2016). Palindromic SNPs with minor allele frequency larger than 0.42 or lack of allele frequency information were considered as ambiguous and thus removed for MR estimation. We only kept SNPs that had association results in both exposure and outcome GWAS studies.

For MR estimation with NAFLD as the exposure, we used the inverse variance weighted (IVW) method to estimate the combined causal effect of *PNPLA3* and *TM6SF2 (NCAN)* variants by assuming a fixed–effect model (Burgess, Butterworth, & Thompson, 2013). As a sensitivity analysis, Wald's method (Wald, 1940) was used to estimate the causal effect with each of the genetic variant respectively. We considered a significant causal relationship if the directions of

the estimates by *PNPLA3*, *TM6SF2 (NCAN)*, and the two variants combined were consistent, and the combined estimate passed the Bonferroni–adjusted significance of $p<0.05$.

For the MR estimation with NAFLD as the outcome, besides the IVW method, we estimated the causal effects using additional methods including weighted median estimator (Bowden, Davey Smith, Haycock, & Burgess, 2016) and MR–Egger (Bowden, Davey Smith, & Burgess, 2015) as a sensitivity analysis. IVW method provides greater precision in MR estimates in the absence of directional pleiotropy (Bowden, Del Greco, et al., 2016). Weighted median estimator relaxes the assumption by requiring at least 50% of the genetic variants to be valid instruments. MR–Egger provides unbiased MR estimates even if the genetic variants exhibit pleiotropic effects given the impendence between instrument strength and pleiotropic effects.

To assess the heterogeneity and identify horizontal pleiotropic outliers, we used the Q' statistics (Bowden, Del Greco, et al., 2018) with modified second order weights and MR–PRESSO global test (Verbanck, Chen, Neale, & Do, 2018). If the horizontal pleiotropy is significant, MR–PRESSO was used to identify outliers at $p<0.05$. We then removed the outliers and retested if the causal relationship and pleiotropic effects were significant.

We considered as significant if the directions of the estimates by three methods (IVW, weighted median, and MR–Egger) were consistent, IVW method passed the Bonferroni–adjusted significance of $p<0.05$, and no significant pleiotropy tested by MR–PRESSO global test and modified Q' statistics. MR analyses were performed with "MendelianRandomization" (Yavorska

& Burgess, 2017), “MRPRESSO” (Verbanck et al., 2018), and “RadialMR” (Bowden, Spiller, et al., 2018) packages in R version 3.5.0 (http://www.r–project.org/).

4.3.2.5 Sample overlap

Participant overlap in two–sample MR might lead to inflated Type 1 error rates due to the weak instrument bias (Burgess, Davies, & Thompson, 2016). We examined the samples used to estimate the genetic correlations with exposure and outcome, respectively. The maximum sample overlapping rate was then calculated as the percentage of overlapped samples in the larger dataset (Burgess et al., 2016). For example, if sample size of study 1 is n1, sample size of study 2 is n2, sample size of the overlapped cohorts in study 1 is m1, sample size of the overlapped cohorts in study 2 is m2, then the maximum overlapping rate = min(m1,m2) / max(n1, n2). As shown in Table 4.3, the degree of overlap in most MR analyses is very low (<1%). For the analyses with substantial overlap (>1%), we have used robust instruments (F statistics ranges from 53 to 110) for MR estimate. Based on a simulation study (Burgess et al., 2016), overlapping bias is unlikely to affect the results given instruments of this degree of strength.

Table 4.3 Sample overlap

Phenotype	Steatosis (GOLD)		Histologic NAFLD (NASH CRN/MIGen)		NAFLD (UKBB)	
	Overlapping cohort	Overlapping percentage	Overlapping cohort	Overlapping percentage	Overlapping cohort	Overlapping percentage
T2D (DIAGRAM)	FHS	0.13%	None	None	UKBB	0.12%
HbA1c (MAGIC)	FHS	1.61%	None	None	None	None
Fasting glucose (MAGIC)	FHS	5.04%	None	None	None	None
Fasting insulin (MAGIC)	FHS	5.67%	None	None	None	None
Fasting proinsulin (MAGIC)	FHS	27.38%	None	None	None	None
2h glucose (MAGIC)	FHS	17.87%	None	None	None	None
HOMA–IR (MAGIC)	FHS	7.92%	None	None	None	None
HOMA–B (MAGIC)	FHS	8.04%	None	None	None	None
Seven oGTT traits (MAGIC)	None	None	None	None	None	None
BMI (GIANT)	None	None	None	None	UKBB	49.70%
WHR (GIANT)	None	None	None	None	UKBB	57.47%
WHRadjBMI (GIANT)	None	None	None	None	UKBB	57.73%
Four plasma lipids (GLGC)	Amish	0.57%	None	None	None	None

The sample overlapping rate was calculated as the percentage of overlapped samples in the larger dataset. For example, if sample size of study 1 is n1, sample size of study 2 is n2, sample size of the overlapped cohorts in study 1 is m1, sample size of the overlapped cohorts in study 2 is m2, then the maximum overlapping rate = min(m1,m2) / max(n1, n2). GOLD: Genetics of Obesity–related Liver Disease; NASH CRN: NASH Clinical Research Network; MIGen: Myocardial Infarction Genetics consortium; UKBB: UK Biobank; DIAGRAM: DIAbetes Genetics Replication And Meta–analysis; MAGIC: Meta–Analyses of Glucose and Insulin–related traits Consortium; GIANT: The Genetic Investigation of ANthropometric Traits consortium; GLGC: The Global Lipids Genetics Consortium; FHS: Framingham Heart Study

4.3.3 Animal experiments

4.3.3.1 Construction of transgenic mice and induction of the NAFLD phenotypes

The human PNPLA3–I148I or PNPLA3–I148M transgenic mice (TghPNPLA3–I148I or TghPNPLA3–I148M) were generated by using the PiggyBac transgenic technology at Cyagen (USA). The human BAC clone CTD–2316P10 were modified to generate the PNPLA3–I148I and PNPLA3148M isoforms and transduced into the embryo stem cell of the C57BL/6 mouse, respectively, without knocking out the mPNPLA3. All animal experiments were carried out at the animal facility with the approval from the Institutional Animal Care and Use Committee of Indiana University School of Medicine in accordance with National Institutes of Health guidelines for the care and use of laboratory animals. Six– to eight–week–old male mice were allowed for free access to water and fed regular chow (Teklad Diets 2018SX: 24% calories from protein, 18% calories from fat, and 58% calories from carbohydrate) in a 12–h/12–h light/dark cycle. For the dietary challenge studies, transgenic (TghPNPLA3–I148I and TghPNPLA3–I148M) and nontransgenic wild type (WT) mice were fed either a high–sucrose (Research Diets, D17070603: 73·5 kcal%, NJ, USA) diet which has been demonstrated to strongly induce hepatic steatosis (but not NASH) as widely reported in the literature including a report on the Pnpla3 I148M knock in mice(Smagris et al., 2015) for 4 weeks, or a high–fat, high–fructose, high–cholesterol diet (HFFC) (Research Diets D18021203: 20% calories from protein, 40% calories from fat, 40% calories from carbohydrate and 1% cholesterol, NJ, USA) for 20 weeks. The HFFC diet mimics the modern Western diet and has been widely demonstrated inductive to induce NASH (Clapper et al., 2013; Denk, Abuja, & Zatloukal, 2018; Panasevich et al., 2018; Van Herck, Vonghia, & Francque, 2017).

4.3.3.2 Glucose and insulin tolerance test

After 16 weeks of HFFC diet feeding, glucose tolerance test (GTT) and insulin tolerance test (ITT) for TghPNPLA3–I148I, TghPNPLA3–I148M, and WT were performed by intraperitoneal (i.p.) injection of 2 g/kg glucose ((Millipore–Sigma, MO, USA) or 0·75 unit/kg insulin (Lily, IN, USA). All of the mice were fasted for 5 hours for GTT or 4 hours for ITT prior to intraperitoneal injection. Plasma glucose levels were determined before the injection of glucose or insulin and at 15, 30, 60 and 120 min after the injection by a Contour glucometer (Bayer HealthCare, IN, USA).

4.3.3.3 Glycemic and lipids profile and fat composition measurement

Fasting blood samples were collected biweekly after fasting for 4 h. A 16–hour fasting was performed for blood glucose measurement at the 16th week for the mice fed with the HFFC diet. Plasma glucose levels were measured by a Contour glucometer. The fasting plasma insulin concentration were determined by a mouse ultrasensitive insulin ELISA kit (ALPCO, NH, USA) according to the manufacturer's protocol.

Hepatic lipid was extracted as previously described (Tao, Xiong, DePinho, Deng, & Dong, 2013). Total cholesterol and triglycerides concentrations in hepatic or plasma were measured according to manufacturer instruction from commercial assay kits (Wako Chemicals, VA, USA), respectively. Mouse body weights were measured weekly. Total body fat mass and lean mass were evaluated using magnetic resonance imaging (MRI) with an EchoMRI 500–Analyzer (EchoMRI, TX, USA) at the Islet and Physiology Core of Indiana University Center for Diabetes and Metabolic Diseases. The tests were performed at week 20 prior to mice sacrifice. At the end

of treatment, mice were sacrificed, and blood, liver, and fat tissue were collected for various analyses.

4.3.3.4 Histological analysis

The liver specimens were fixed in 10% formalin solution and routinely processed for paraffin–embedding, sectioning (5 μm thickness) and hematoxylin and eosin (H&E) staining were performed at the Histology Core of Indiana University School of Medicine following the standard protocol. For Sirius red staining, sections were incubated with a 0·1% direct red 80 (Sigma–Aldrich, MO, USA) plus 0·1% fast green FCF (Fisher Scientific, MA, USA) solution dissolved in saturated aqueous picric acid (1.2% picric acid in water) for 1 hour at room temperature, dehydrated, and mounted with a Permount™ mounting medium (Fisher Scientific, MA, USA). The images of H&E and Sirius red staining were captured using a Leica DM750 microscope equipped with an EC3 digital camera (100× magnifications). The lipid droplets area and Sirius red positive area were quantified using ImageJ software.

4.3.3.5 Immunofluorescence (IF)

Sections were deparaffinized, rehydrated by washes in graded alcohols (ethanol 100%, 95%, 70%, 50%) and distilled water, and retrieved in 0·01 M sodium citrate buffer (10 mM Sodium Citrate, 0.05% Tween 20, pH 6.0) for 20 minutes at 98°C by microwave heating. Sections were blocked with blocking buffer (10% normal serum, 1% bovine serum albumin (BSA) and 0·1% Triton X–100 in PBS) for 2 h at room temperature and incubated overnight at 4°C with rabbit polyclonal anti–myeloperoxidase (MPO, 1:100, Invitrogen, CA, USA) or rabbit monoclonal anti–F4/80 (1:100, Invitrogen, CA, USA) antibody, respectively. After washing with 0·025% Triton X–100 in PBS, sections were incubation with Alexa–Fluor–488 secondary antibodies (Molecular Probes, OR, USA) for 1 h at room temperature, and then mounted with

VECTASHIELD® antifade mounting media with DAPI (Vector Laboratories, CA, USA). Immunofluorescence images were captured by a ZEISS fluorescence microscope using an AxionVison Rel 4·8 software (200× magnification). The positive cells were counted in randomly selected fields (five fields per section).

4.3.3.6 Statistical analyses

Two–way repeated measures ANOVA was used to assess the effects of genotype, time, and genotype–time interaction on body weight, glucose, insulin, cholesterol, and triglycerides levels. Tukey's multiple comparisons test was used to determine the significance of pair–wise comparisons at each time point. Tukey corrected p value<0.05 was considered significant. All the statistical analyses for animal experiments were performed using GraphPad Prism Version 6·00 (GraphPad Software, CA, USA).

4.4 Results

4.4.1 Study overview

The design of this study consists of three steps (Figure 4.1). We first aimed to explore whether NAFLD (both CT measured hepatic steatosis and biopsy–proven histologic NASH) casually affects T2D, obesity and their related intermediate traits. To this end, we used the summary GWAS data for CT scan–measured steatosis (GOLD) as well as the histologic NASH progression (NASH CRN/MIGen) presented in Speliotes et al.(Speliotes et al., 2011), which is thus far the largest GWAS comprehensively covering multiple phenotypes of NAFLD. We also use the summary GWAS meta–analysis data for T2D (DIAGRAM),(Mahajan et al., 2018) obesity (GIANT)(Pulit et al., 2019), glycemic (MAGIC)(Dupuis et al., 2010; Manning et al., 2012; Prokopenko et al., 2014; Saxena et al., 2010; Strawbridge et al., 2011; Wheeler et al.,

2017), and lipid (GLGC)(Willer et al., 2013) traits as outcomes (Step 1), which are the largest–to–date GWAS data on these phenotypes. The causal role of NAFLD on T2D and obesity was conducted by using two well–established NAFLD–associated polymorphisms in PNPLA3 and TM6SF2/NCAN loci as a proxy for hepatic steatosis and NASH progression. We further performed a GWAS on "fatty liver disease" in UK biobank. Using the aforementioned summary level data of DIAGRAM and GIANT, we then investigated the reverse relationships, i.e. whether T2D or obesity causally affect NAFLD risk in the UK Biobank samples (Step 2). Finally, we constructed a transgenic mouse model expressing human PNPLA3–I148I or PNPLA3–I148M isoform to test the causal effects of hepatic steatosis on T2D and obesity (Step 3). To test the effect of steatosis and NASH disease progression on the susceptibility to T2D and obesity, we used a high sucrose diet (HSD) known to induce steatosis in PNPLA3148M mice (Smagris et al., 2015), as well as a high fat, high fructose and high cholesterol diet (HFFC) known to induce NASH in mice (Clapper et al., 2013; Denk et al., 2018; Panasevich et al., 2018; Van Herck et al., 2017). A schematic representation of the three assumptions for an MR analysis was shown in Figure 4.2 A, and the MR methods and heterogeneity tests used in the study were listed in Figure 4.2 B.

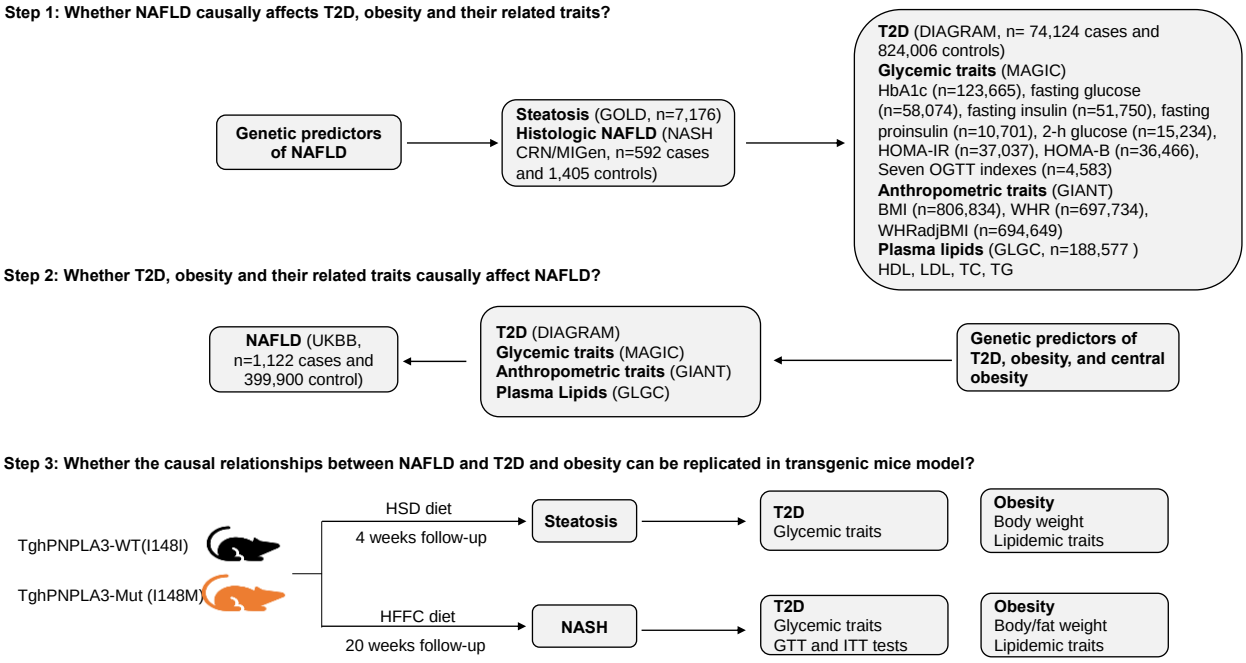
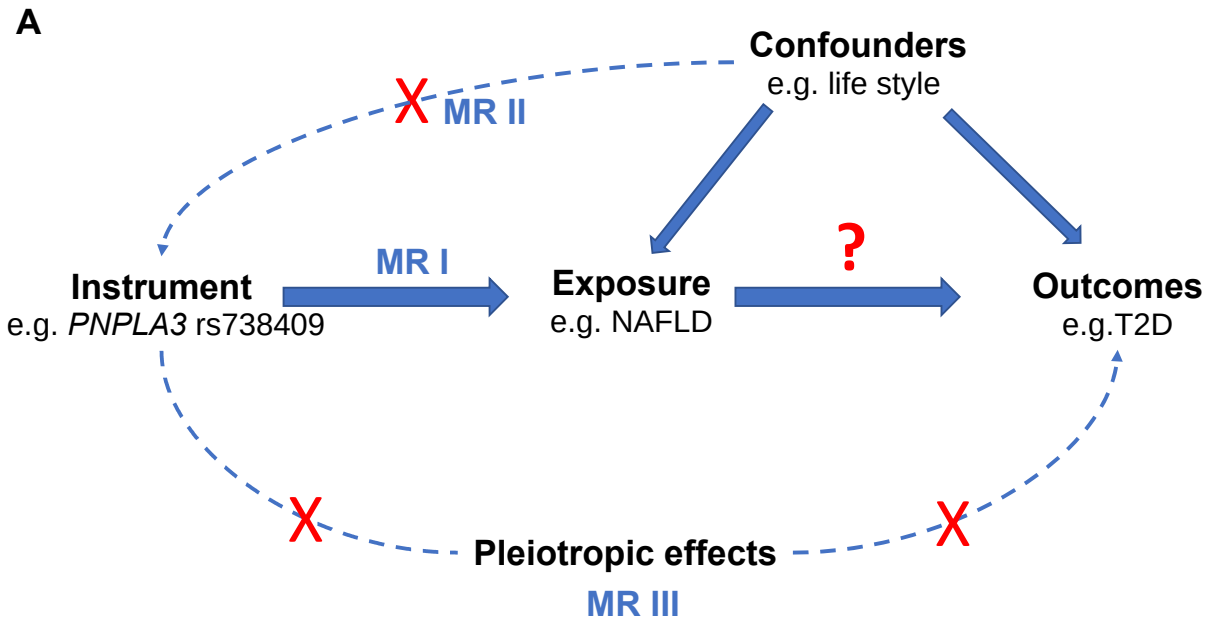


Figure 4.1 Flowchart of the study design

The summary–level associations were taken from the following genomics consortium: GOLD (Genetics of Obesity–related Liver Disease) for computerized tomography (CT) measured hepatic steatosis (Speliotes et al., 2011); NASH Clinical Research Network (NASH CRN) and Myocardial Infarction Genetics Consortium (MIGen) for biopsy–proven NAFLD (Speliotes et al., 2011); DIAbetes Genetics Replication And Meta–analysis (DIAGRAM) for T2D (Mahajan et al., 2018); Meta–Analyses of Glucose and Insulin–related traits (MAGIC) consortium for glycemic traits including HbA1c (Wheeler et al., 2017), fasting glucose (Manning et al., 2012), fasting insulin (Manning et al., 2012), fasting proinsulin (Strawbridge et al., 2011), 2–h glucose (Saxena et al., 2010), homeostatic model assessment of insulin resistance (HOMA–IR)(Dupuis et al., 2010) and β–cell function (HOMA–B) (Dupuis et al., 2010), and seven insulin secretion and action indices during oral glucose tolerance test (OGTT)(Prokopenko et al., 2014); The Genetic Investigation of ANthropometric Traits (GIANT) consortium for body mass index (BMI), waist–hip ratio (WHR), and WHR adjusted for BMI (WHRadjBMI)(Pulit et al., 2019); The Global Lipids Genetics Consortium (GLGC) for plasma high–density lipoprotein cholesterol (HDL), low–density lipoprotein cholesterol (LDL), total cholesterol (TC), and triglycerides (TG) levels (Willer et al., 2013).



B

	MR tests	Heterogeneity
NAFLD as exposure	IVW	Wald test for each of the variants
NAFLD as outcome	IVW, Median, MR-Egger	MR-PRESSO, Q'

Figure 4.2 Diagram of the two–sample MR design, three assumptions, and methods used.

(A) MR analysis was used to explore the causal relationships among NAFLD, T2D and obesity. Three assumptions of the genetic instrumental variable are as follows: 1) The genetic variant (e.g. PNPLA3 rs738409) is robustly associated with the exposure of interest (e.g. NAFLD), 2) the genetic variant is not associated with confounding factors (e.g. life style), and 3) there is no alternative pathway through which the genetic variant affects the outcomes (e.g. T2D) other than via the exposure. (B) MR methods and heterogeneity tests used in the study. IVW: inverse variance weighted; Median: weighted median estimator; MR–PRESSO: MR pleiotropy residual sum and outlier; Q': Q' statistics with modified second order weights.

4.4.2 The causal effect of NAFLD on T2D risk and glycemic traits

Using two well–established NAFLD–associated variants in PNPLA3 (rs738409) and TM6SF2/NCAN (rs2228603) gene loci as the genetic predictors of steatosis and histologic NASH progression, we tested the causal effect of NAFLD on T2D and glycemic traits in the latest publicly available GWAS data. As listed in Table 4.4, we observed a significant

association between genetically instrumented hepatic steatosis and T2D risk, in which a one–standard deviation (SD) increase in CT measured hepatic steatosis caused a 30% increased risk of T2D (OR: 1.3, 95% CI: [1.2, 1.4], *p*=8.3e–14). As for glycemic traits, we detected nominal associations of steatosis with increased fasting glucose (β: 0.026 mmol/L, 95% CI: [8.5e–5, 0.051], *p*=0·049), fasting insulin (β: 0.025 pmol/L, 95% CI: [0.0035, 0.046], *p*=0.022), and insulin resistance (HOMA–IR) (β: 0·03 (mU/L)*(mmol/L), 95% CI: [–0.0016, 0·061], *p*=0.063) levels. However, these results were not significant after adjusted with Bonferroni correction (p<2·3e–3, correction for 22 traits). Other tested glycemic traits including HbA1c, fasting proinsulin, 2–h glucose, HOMA–IR, HOMA–B, and seven insulin secretion and action indices during OGTT did not show any significance.

We then tested if genetically increased risk for the disease progression of NASH also has causal effect on T2D susceptibility and glycemic traits. Consistent with the result of hepatic steatosis, there was evidence of a significant effect of histologically characterized NAFLD severity on an increased risk of T2D (OR: 1·06, 95% CI: [1·03, 1·09], *p*=2·8e–4). No significant causal relationship was found between genetically driven NAFLD and glycemic traits except a nominal association with increased fasting insulin levels (β: 0·0064 pmol/L, 95% CI: [0·00034, 0·012], *p*=0·038) (Table 4.4).

Table 4.4 MR estimate with NAFLD as exposure

	Hepatic steatosis (per SD)		Histologic NAFLD (per logOR)	
T2D and glycemic traits	**Effect (95% CI)**	***p***	**Effect (95% CI)**	***p***
T2D (OR)	1.30 (1.21, 1.39)	**8.30E–14**	1.06 (1.03, 1.09)	**2.80E–04**
HbA1c (%)	–0.0072 (–0.025, 0.01)	4.20E–01	–0.0025 (–0.0074, 0.0024)	3.10E–01
Fasting glucose (mmol/L)	0.026 (8.5E–05, 0.051)	4.90E–02	0.0032 (–0.0031, 0.0096)	3.20E–01
Fasting insulin (pmol/L)	0.025 (0.0035, 0.046)	2.20E–02	0.0064 (0.00034, 0.012)	3.80E–02
Fasting proinsulin (pmol/L)	0.032 (–0.026, 0.089)	2.80E–01	0.0051 (–0.0091, 0.019)	4.80E–01
2h glucose (mmol/L)	–0.066 (–0.21, 0.079)	3.70E–01	–0.015 (–0.051, 0.021)	4.10E–01
HOMA–IR ((mU/L)*(mmol/L))	0.03 (–0.0016, 0.061)	6.30E–02	0.0066 (–0.0016, 0.015)	1.10E–01
HOMA–B ((mU/L)/(mmol/L))	0.007 (–0.019, 0.033)	5.90E–01	0.0011 (–0.0052, 0.0074)	7.30E–01
AUCins (mU*min/L)	–0.012 (–0.18, 0.16)	8.90E–01	–0.0043 (–0.046, 0.038)	8.40E–01
AUCins/AUCgluc (mU/mmol)	–0.01 (–0.18, 0.16)	9.10E–01	–0.0037 (–0.046, 0.038)	8.60E–01
Incre30 (mU/L)	–0.0021 (–0.17, 0.16)	9.80E–01	–0.00033 (–0.041, 0.04)	9.90E–01
Ins30adjBMI	0.017 (–0.15, 0.19)	8.40E–01	0.00083 (–0.041, 0.043)	9.70E–01
ISI (mg/dL)	0.011 (–0.18, 0.2)	9.10E–01	0.017 (–0.032, 0.065)	5.00E–01
CIRadjBMI	–0.0055 (–0.18, 0.17)	9.50E–01	–0.0034 (–0.045, 0.039)	8.80E–01
DI	0.078 (–0.082, 0.24)	3.40E–01	0.022 (–0.018, 0.063)	2.80E–01
Obesity traits				
BMI (SD)	–0.027 (–0.041, –0.013)	**1.30E–04**	–0.0071 (–0.012, –0.0023)	3.40E–03
WHR (SD)	0.021 (0.0062, 0.036)	5.40E–03	0.0029 (–0.00091, 0.0067)	1.30E–01
WHRadjBMI (SD)	0.039 (0.023, 0.054)	**8.20E–07**	0.0072 (0.0022, 0.012)	4.50E–03
HDL* (SD)	–0.06 (–0.1, –0.016)	8.20E–03	–0.013 (–0.025, –0.0014)	2.80E–02
LDL* (SD)	–0.058 (–0.11, –0.0097)	1.90E–02	–0.013 (–0.025, –4E–04)	4.30E–02
TC* (SD)	–0.084 (–0.13, –0.036)	**6.80E–04**	–0.019 (–0.033, –0.0044)	9.90E–03
TG* (SD)	0.02 (–0.023, 0.062)	3.60E–01	0.0043 (–0.0052, 0.014)	3.80E–01

Effect was estimated by a combined genetic vector of PNPLA3 and TM6SF2(NCAN) variants through inverse variance weighted (IVW) method. The F statistics of the genetic predictors of hepatic steatosis and histologic NAFLD are 110 and 10, respectively. Significant results at the Bonferroni–adjusted level of significance ($p<0.05/22=2.3e–3$) were highlighted in bold. *estimate was made using PNPLA3 variant only due to the pleiotropic effect of TM6SF2(NCAN). T2D: type 2 diabetes; HOMA–IR: The homeostatic model assessment (HOMA) insulin resistance; HOMA–B: HOMA beta cell function; AUCins: area under the curve (AUC) of insulin levels during oral glucose tolerance test (OGTT); AUCins/AUCgluc: ratio of AUC insulin and AUC glucose; Incre30: incremental insulin at 30 min; Ins30adjBMI: Insulin response to glucose during the first 30 min adjusted for BMI; ISI: insulin sensitivity index; CIRadjBMI: Corrected Insulin Response adjusted for ISI; DI: disposition index; BMI: body mass index; WHR: waist–hip ratio; WHRadjBMI: WHR adjusted for BMI; HDL: high–density lipoprotein cholesterol; LDL: low–density lipoprotein cholesterol; TC: total cholesterol; TG: triglycerides; OR: odds ratio; CI: confidence interval; SD: standard deviation.

In order to avoid potential biases on the selection of genetic instrument, a post–hoc analysis by separately examining the causal role of PNPLA3 and TM6SF2/NCAN polymorphisms was also performed and generated similar results (Table 4.5).

Table 4.5 Full results of MR estimate with NAFLD as exposure

Phenotype	**Hepatic steatosis (per SD)**						**Histologic NAFLD (per logOR)**					
	***PNPLA3* rs738409**		***NCAN* rs2228603**		**PNPLA3 rs738409 + NCAN rs2228603**		***PNPLA3* rs738409**		***NCAN* rs2228603**		**PNPLA3 rs738409 + NCAN rs2228603**	
T2D and glycemic traits	**Effect (95% CI)**	***p***	**Effect (95% CI)**	***p***	**Effect (95% CI)**	***p***	**Effect (95% CI)**	***p***	**Effect (95% CI)**	***p***	**Effect (95% CI)**	***p***
T2D (OR)	1.27 (1.17, 1.36)	1.20E–09	1.43 (1.23, 1.68)	5.20E–06	1.30 (1.21, 1.39)	8.30E–14	1.05 (1.02, 1.09)	9.10E–04	1.19 (1.02, 1.39)	4.30E–02	1.06 (1.03, 1.09)	2.80E–04
HbA1c (%)	–0.014 (–0.037, 0.0085)	2.20E–01	0.0029 (–0.024, 0.03)	8.30E–01	–0.0072 (–0.025, 0.01)	4.20E–01	–0.0031 (–0.0084, 0.0021)	2.40E–01	0.0014 (–0.026, 0.029)	8.30E–01	–0.0025 (–0.0074, 0.0024)	3.10E–01
Fasting glucose (mmol/L)	0.011 (–0.017, 0.04)	4.50E–01	0.084 (0.027, 0.14)	4.00E–03	0.026 (8.5e–05, 0.051)	4.90E–02	0.0025 (–0.004, 0.0089)	4.50E–01	0.04 (–0.017, 0.097)	7.50E–02	0.0032 (–0.0031, 0.0096)	3.20E–01
Fasting insulin (pmol/L)	0.03 (0.005, 0.054)	1.80E–02	0.01 (–0.034, 0.054)	6.50E–01	0.025 (0.0035, 0.046)	2.20E–02	0.0065 (0.00022, 0.013)	4.30E–02	0.0048 (–0.039, 0.049)	6.60E–01	0.0064 (0.00034, 0.012)	3.80E–02
Fasting proinsulin (pmol/L)	0.015 (–0.05, 0.081)	6.50E–01	0.084 (–0.034, 0.2)	1.60E–01	0.032 (–0.026, 0.089)	2.80E–01	0.0034 (–0.011, 0.018)	6.50E–01	0.04 (–0.078, 0.16)	2.30E–01	0.0051 (–0.0091, 0.019)	4.80E–01
2h glucose (mmol/L)	–0.061 (–0.23, 0.1)	4.70E–01	–0.08 (–0.38, 0.22)	6.00E–01	–0.066 (–0.21, 0.079)	3.70E–01	–0.014 (–0.051, 0.024)	4.80E–01	–0.038 (–0.34, 0.26)	6.10E–01	–0.015 (–0.051, 0.021)	4.10E–01
HOMA–IR ((mU/L)*(mmol/L))	0.027 (–0.0087, 0.062)	1.40E–01	0.041 (–0.027, 0.11)	2.40E–01	0.03 (–0.0016, 0.061)	6.30E–02	0.0059 (–0.0025, 0.014)	1.70E–01	0.019 (–0.048, 0.087)	3.00E–01	0.0066 (–0.0016, 0.015)	1.10E–01
HOMA–B ((mU/L)/(mmol/L))	0.0027 (–0.027, 0.032)	8.60E–01	0.022 (–0.032, 0.077)	4.30E–01	0.007 (–0.019, 0.033)	5.90E–01	0.00059 (–0.0059, 0.0071)	8.60E–01	0.011 (–0.044, 0.065)	4.50E–01	0.0011 (–0.0052, 0.0074)	7.30E–01
AUCins (mU*min/L)	–0.025 (–0.22, 0.17)	8.00E–01	0.033 (–0.33, 0.4)	8.60E–01	–0.012 (–0.18, 0.16)	8.90E–01	–0.0056 (–0.049, 0.038)	8.00E–01	0.016 (–0.35, 0.38)	8.60E–01	–0.0043 (–0.046, 0.038)	8.40E–01
AUCins/AUCgluc (mU/mmol)	–0.022 (–0.22, 0.17)	8.30E–01	0.031 (–0.33, 0.39)	8.70E–01	–0.01 (–0.18, 0.16)	9.10E–01	–0.0048 (–0.048, 0.038)	8.30E–01	0.015 (–0.35, 0.38)	8.70E–01	–0.0037 (–0.046, 0.038)	8.60E–01
Incre30 (mU/L)	–0.00057 (–0.19, 0.19)	1.00E+00	–0.0076 (–0.36, 0.35)	9.70E–01	–0.0021 (–0.17, 0.16)	9.80E–01	–0.00013 (–0.042, 0.041)	1.00E+00	–0.0036 (–0.36, 0.35)	9.70E–01	–0.00033 (–0.041, 0.04)	9.90E–01

Table 4.5 continued

Ins30adjBMI	–0.01 (–0.21, 0.18)	9.20E–01	0.11 (–0.25, 0.46)	5.50E–01	0.017 (–0.15, 0.19)	8.40E–01	–0.0023 (–0.045, 0.041)	9.20E–01	0.052 (–0.3, 0.41)	5.60E–01	0.00083 (–0.041, 0.043)	9.70E–01
ISI (mg/dL)	0.11 (–0.1, 0.33)	3.00E–01	–0.28 (–0.64, 0.086)	1.30E–01	0.011 (–0.18, 0.2)	9.10E–01	0.025 (–0.024, 0.075)	3.20E–01	–0.13 (–0.49, 0.23)	2.10E–01	0.017 (–0.032, 0.065)	5.00E–01
CIRadjBMI	–0.024 (–0.22, 0.17)	8.10E–01	0.055 (–0.3, 0.41)	7.60E–01	–0.0055 (–0.18, 0.17)	9.50E–01	–0.0052 (–0.048, 0.038)	8.10E–01	0.026 (–0.33, 0.38)	7.60E–01	–0.0034 (–0.045, 0.039)	8.80E–01
DI	0.11 (–0.066, 0.3)	2.10E–01	–0.05 (–0.39, 0.29)	7.70E–01	0.078 (–0.082, 0.24)	3.40E–01	0.025 (–0.017, 0.067)	2.40E–01	–0.024 (–0.36, 0.31)	7.70E–01	0.022 (–0.018, 0.063)	2.80E–01
Obesity traits												
BMI (SD)	–0.031 (–0.048, –0.015)	2.00E–04	–0.017 (–0.043, 0.0092)	2.00E–01	–0.027 (–0.041, –0.013)	1.30E–04	–0.0069 (–0.012, –0.0019)	6.70E–03	–0.008 (–0.034, 0.018)	2.70E–01	–0.0071 (–0.012, –0.0023)	3.40E–03
WHR (SD)	0.011 (–0.0055, 0.028)	1.90E–01	0.056 (0.025, 0.088)	4.90E–04	0.021 (0.0062, 0.036)	5.40E–03	0.0025 (–0.0014, 0.0063)	2.10E–01	0.027 (–0.0049, 0.058)	5.80E–02	0.0029 (–0.00091, 0.0067)	1.30E–01
WHRadjBMI (SD)	0.03 (0.013, 0.047)	6.50E–04	0.075 (0.04, 0.11)	2.40E–05	0.039 (0.023, 0.054)	8.20E–07	0.0066 (0.0016, 0.012)	9.80E–03	0.036 (0.00086, 0.07)	4.60E–02	0.0072 (0.0022, 0.012)	4.50E–03
HDL (SD)	–0.06 (–0.1, –0.016)	8.20E–03	0.023 (–0.033, 0.078)	4.20E–01	–0.028 (–0.062, 0.0071)	1.20E–01	–0.013 (–0.025, –0.0014)	2.80E–02	0.011 (–0.045, 0.066)	4.50E–01	–0.0096 (–0.021, 0.0013)	8.30E–02
LDL (SD)	–0.058 (–0.11, –0.0097)	1.90E–02	–0.44 (–0.58, –0.3)	7.60E–10	–0.098 (–0.14, –0.053)	2.30E–05	–0.013 (–0.025, –4e–04)	4.30E–02	–0.21 (–0.35, –0.068)	3.40E–02	–0.014 (–0.026, –0.0012)	3.10E–02
TC (SD)	–0.084 (–0.13, –0.036)	6.80E–04	–0.51 (–0.67, –0.35)	2.20E–10	–0.12 (–0.17, –0.074)	3.30E–07	–0.019 (–0.033, –0.0044)	9.90E–03	–0.24 (–0.4, –0.085)	3.30E–02	–0.019 (–0.033, –0.0054)	6.80E–03
TG (SD)	0.02 (–0.023, 0.062)	3.60E–01	–0.44 (–0.58, –0.31)	3.50E–10	–0.02 (–0.06, 0.021)	3.40E–01	0.0043 (–0.0052, 0.014)	3.80E–01	–0.21 (–0.35, –0.072)	3.30E–02	0.0038 (–0.0057, 0.013)	4.40E–01

Effect was estimated by Wald's method and inverse variance weighted (IVW) method for single variant, combined variants, respectively.

T2D: type 2 diabetes; HOMA–IR: The homeostatic model assessment (HOMA) insulin resistance; HOMA–B: HOMA beta cell function; AUCins: area under the curve (AUC) of insulin levels during oral glucose tolerance test (OGTT); AUCins/AUCgluc: ratio of AUC insulin and AUC glucose; Incre30: incremental insulin at 30 min;

Ins30adjBMI: Insulin response to glucose during the first 30 min adjusted for BMI; ISI: insulin sensitivity index; CIRadjBMI: Corrected Insulin Response adjusted for ISI; DI: disposition index; BMI: body mass index; WHR: waist–hip ratio; WHRadjBMI: WHR adjusted for BMI; HDL: high–density lipoprotein cholesterol; LDL: low–density lipoprotein cholesterol; TC: total cholesterol; TG: triglycerides; OR: odds ratio; CI: confidence interval; SD: standard deviation.

4.4.3 The causal effect of NAFLD on obesity

While obesity is a well–known risk factor for NAFLD, the reverse relationship, i.e. the causal effect of NAFLD on obesity, has not been explored before. Using the same genetic predictors of

steatosis and histologic NAFLD, we implemented MR and observed a significant causal association of a one–SD increase in hepatic fat with a 0.027–SD decrease in BMI (β:–0·027, 95%CI: [–0·043, –0·013], *p*=1·3e–4), but a 0·039–SD increase in WHRadjBMI (WHR adjusted for BMI) (β:0·039, 95%CI: [0·023, 0·054], *p*=8·2e–7), an established marker for abdominal or central obesity. Similar relationships were found between genetically raised histologic NAFLD with BMI (β:–0·0071, 95%CI: [–0·012, –0·0023], *p*=3·4e–3) and WHRadjBMI (β:0·0072, 95%CI: [0·0022, 0·012], *p*=4·5e–3). Taken together, our analyses suggested a consistent negative causal relationship between NAFLD and overall obesity (measured by BMI), but a positive correlation with central or visceral obesity (measured by WHRadjBMI).

We next investigated the causal effect of NAFLD on blood lipid levels including HDL, LDL, total cholesterol (TC), and triglycerides (TG). Since TM6SF2/NCAN plays a role in plasma lipid regulation and might lead to the invalidity of the MR assumptions, we used PNPLA3 rs738409 variant only with regard to estimating the causal effect on blood lipid profile. Our analyses demonstrated a significant negative correlation between genetically raised hepatic fat and TC levels (β:–0·084 SD, 95%CI: [–0·13, –0·036], *p*=6·8e–4). This negative relationship also exists for histologic NAFLD (β:–0·019 SD, 95%CI: [–0·033, –0·0044], *p*=9·9e–3). No significant causal relationship was found with other lipids (Table 4.4).

4.4.4 Reverse MR investigating the causal effects of T2D, obesity, and their related secondary traits on NAFLD

To understand the causal relationships among the three diseases, we implemented MR analyses to test the existence of the reverse or bidirectional causal relationships between T2D or obesity and NAFLD. We used the data from our newly performed GWAS on NAFLD as an outcome, while the DIAGRAM and GIANT summary data for T2D and obesity as predictor, respectively.

We found that genetic predictors of T2D exert positive effects on NAFLD (OR: 1·1, 95% CI: [1·0, 1·2], p=1·67e–3) without evidence of significant heterogeneity ($P_{MR–PRESSOGlobal}$=0·31, $P_{modified\ Q}$=0·33) after removing outlier variants identified by MR–PRESSO (Table 4.6). Initial MR estimates without removing outliers were shown in Table 4.7.

Consistent with the previous MR estimate (Stender et al., 2017), we found BMI causally increased the NAFLD risk (OR: 2·3, 95% CI: [2·0, 2·7], p=1·4e–25), but with remaining heterogeneity ($P_{MR–PRESSOGlobal}$<2·5e–3, $P_{modified\ Q}$=3·9e–3) after removing outliers. BMI adjusted WHR also significantly aggravated NAFLD risk (OR: 1·5 95% CI: [1·3, 1·8], p=1·1e–6) without evident heterogeneity ($P_{MR–PRESSOGlobal}$=0·46, $P_{modified\ Q}$=0·45).

Table 4.6 MR estimate with NAFLD as outcome

			IVW		Weighted median		MR–Egger		Pleiotropy test	
T2D and glycemic traits[#]	**#SNPs**	**F**	**OR (95% CI)**	***p***	**OR (95% CI)**	***p***	**OR (95% CI)**	***p***	**MR–PRESSO global test p**	**modified Q' *p***
T2D (logOR)*	315	69	1.1 (1, 1.2)	**1.67E–03**	1.1 (0.93, 1.2)	3.54E–01	0.99 (0.84, 1.2)	8.59E–01	3.11E–01	3.26E–01
HbA1c (%)	67	69	0.33 (0.17, 0.66)	**1.76E–03**	0.33 (0.11, 1.11)	7.44E–02	0.17 (0.04, 0.68)	1.28E–02	3.85E–01	3.69E–01
Fasting glucose (mmol/L)	33	68	0.42 (0.25, 0.73)	**1.60E–03**	0.33 (0.15, 0.80)	1.34E–02	0.44 (0.12, 1.52)	1.93E–01	8.78E–02	8.29E–02
Fasting insulin (pmol/L)*	9	36	7.7 (1.3, 46)	2.40E–02	4.6 (0.42, 51)	2.11E–01	4.2 (0.00035, 51000)	7.64E–01	3.18E–01	2.82E–01
Fasting proinsulin (pmol/L)	15	72	1.13 (0.79, 1.62)	4.94E–01	0.99 (0.58, 1.67)	9.55E–01	1.57 (0.49, 4.95)	4.51E–01	4.60E–02	4.74E–02
Obesity traits										
BMI (SD)*	1839	60	2.3 (2, 2.7)	1.41E–25	2.1 (1.6, 2.8)	1.02E–07	1.7 (1.1, 2.7)	1.58E–02	2.50E–03	3.89E–03
WHR (SD)	951	53	2.10 (1.72, 2.59)	**2.85E–12**	2.34 (1.67, 3.32)	6.91E–07	1.62 (0.94, 2.72)	8.14E–02	3.62E–01	3.64E–01
WHRadjBMI (SD)	1156	63	1.54 (1.30, 1.82)	**1.11E–06**	1.60 (1.20, 2.14)	1.53E–03	1.57 (1.06, 2.34)	2.37E–02	4.55E–01	4.54E–01
HDL (SD)*	226	123	0.88 (0.76, 1)	8.78E–02	0.83 (0.64, 1.1)	1.57E–01	0.93 (0.69, 1.3)	6.62E–01	2.00E–03	1.26E–03
LDL (SD)*	192	148	0.96 (0.84, 1.1)	5.47E–01	1 (0.79, 1.3)	9.69E–01	0.84 (0.64, 1.1)	2.11E–01	<5.00E–04	7.29E–06
TC (SD)*	239	110	0.94 (0.82, 1.1)	4.18E–01	0.95 (0.71, 1.3)	7.13E–01	0.88 (0.66, 1.2)	3.85E–01	1.00E–03	6.33E–04
TG (SD)*	149	119	1.6 (1.3, 1.9)	5.10E–07	1.7 (1.2, 2.4)	1.45E–03	1.7 (1.1, 2.4)	8.92E–03	7.00E–03	7.83E–03

Table 4.7 MR estimates following outlier removal

Trait	Initial Estimate						After outlier removal					
	# SNPs	F	IVW		Pleiotropy test		# SNPs	F	IVW		Pleiotropy test	
			Effect (95% CI)	*p*	MR–PRESSO global test p	modified Q' *p*			Effect (95% CI)	*p*	MR–PRESSO global test p	modified Q' *p*
T2D (logOR)	318	69	1.14 (1.05, 1.22)	9.60E–04	1.50E–03	7.34E–04	315	69	1.1 (1, 1.2)	1.67E–03	3.11E–01	3.26E–01
Fasting insulin (pmol/L)	10	37	2.36 (0.44, 12.18)	3.15E–01	2.00E–04	3.14E–04	9	36	7.7 (1.3, 46)	2.40E–02	3.18E–01	2.82E–01
BMI (SD)	1843	60	2.27 (1.95, 2.66)	4.31E–25	<2E–04	1.55E–04	1839	60	2.3 (2, 2.7)	1.41E–25	2.50E–03	3.89E–03
HDL (SD)	228	123	0.88 (0.76, 1.02)	7.78E–02	<2E–04	2.18E–05	226	123	0.88 (0.76, 1)	8.78E–02	2.00E–03	1.26E–03
LDL (SD)	193	149	0.93 (0.81 1.07)	2.95E–01	<2E–04	1.11E–07	192	148	0.96 (0.84, 1.1)	5.47E–01	<5.00E–04	7.29E–06
TC (SD)	242	112	0.92 (0.80, 1.06)	2.38E–01	<2E–04	4.23E–07	239	110	0.94 (0.82, 1.1)	4.18E–01	1.00E–03	6.33E–04
TG (SD)	151	120	1.48 (1.22 1.77)	3.74E–05	<2E–04	7.04E–06	149	119	1.6 (1.3, 1.9)	5.10E–07	7.00E–03	7.83E–03

IVW: inverse variance weighted; F: F statistics for the strength of correlation between instrument and exposure; MR–PRESSO: MR pleiotropy residual sum and outlier; Q': Q' statistics with modified second order weights; T2D: type 2 diabetes; BMI: body mass index; HDL: high–density lipoprotein cholesterol; LDL: low–density lipoprotein cholesterol; TC: total cholesterol; TG: triglycerides; OR: odds ratio; CI: confidence interval; SD: standard deviation.

4.4.5 Transgenic mice study on the relationship between NAFLD and susceptibility to T2D and obesity.

To further examine the causal effect of NAFLD on T2D and obesity, we set out to induce hepatosteatosis and NASH using transgenic animal models expressing human PNPLA3 isoforms, and during which, to observe the development of T2D and obesity phenotypes. To do so, we constructed mice models transduced with the bacterial artificial chromosome (BAC)–containing the PNPLA3–I148I isoform or that was engineered to the PNPLA3–I148M isoform. The mice were then fed with a previously established high sucrose diet (HSD) for 4 weeks to induce hepatosteatosis (Smagris et al., 2015). To examine the effect of NASH progression on T2D or obesity phenotypes, we also fed the mice with the "Western diet" characterized with high fat, high fructose and high cholesterol (HFFC) for 20 weeks, which has been an established NASH–inductive diet as demonstrated before(Clapper et al., 2013; Denk et al., 2018; Panasevich et al., 2018; Van Herck et al., 2017).

After 4 weeks of HSD diet, the TghPNPLA3–I148M mice developed severe hepatosteatosis as compared to their TghPNPLA3–I148I littermates or the non–transgenic controls, characterized with significantly increased lipid droplets formation in the liver (Figure 4.3 A) as well as hepatic triglycerides (TG) accumulation (Figure 4.3 B). Meanwhile, as compared to the TghPNPLA3–I148I controls, the TghPNPLA3–I148M mice also demonstrated a trend of increased circulating glucose (Figure 4.4 A), but the insulin level remained unchanged (Figure 4.4 B). After 4 weeks of HSD diet feeding, the TghPNPLA3–I148M mice demonstrated no significant change in total body weight (Figure 4.4 C), but a marginal trend to a reduced total circulating cholesterol

(Figure 3D, *p*=0·038), but not the circulating TG levels (Figure 4.4 E), as compared to their TghPNPLA3–I148I littermates.

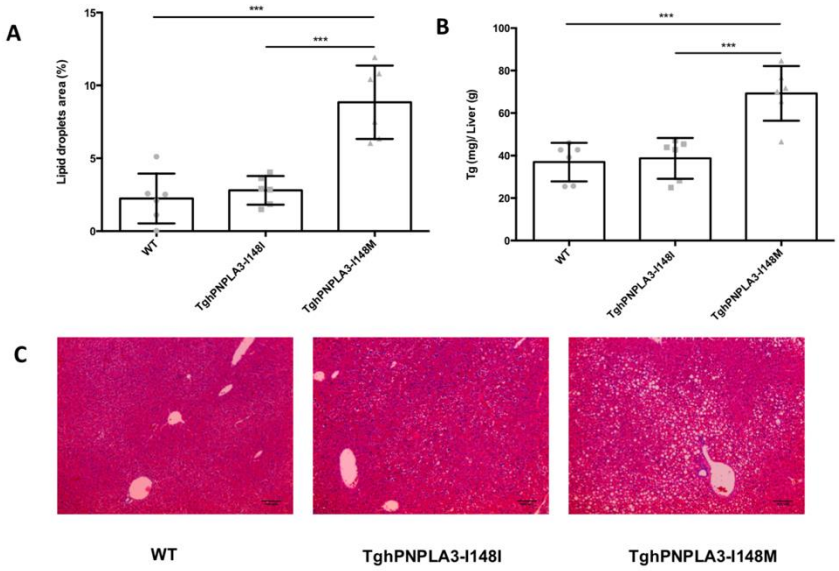


Figure 4.3 Lipid droplets and TG accumulation of mice fed with an HSD diet for 4 weeks

(A) Lipid droplets area of TghPNPLA3–I148I, TghPNPLA3–I148M, and non–transgenic wide type mice fed with an HSD diet for 4 weeks. (B) Liver triglycerides levels of the three groups after 4 weeks. (C) H&E staining of liver sections. Error bar represents standard deviation (SD). The significance level of the comparison between TghPNPLA3–I148I and TghPNPLA3–I148M was indicated as follows: *: Tukey adjusted p<0.05; **: Tukey adjusted p<0.01; ***: Tukey adjusted p<0.001.

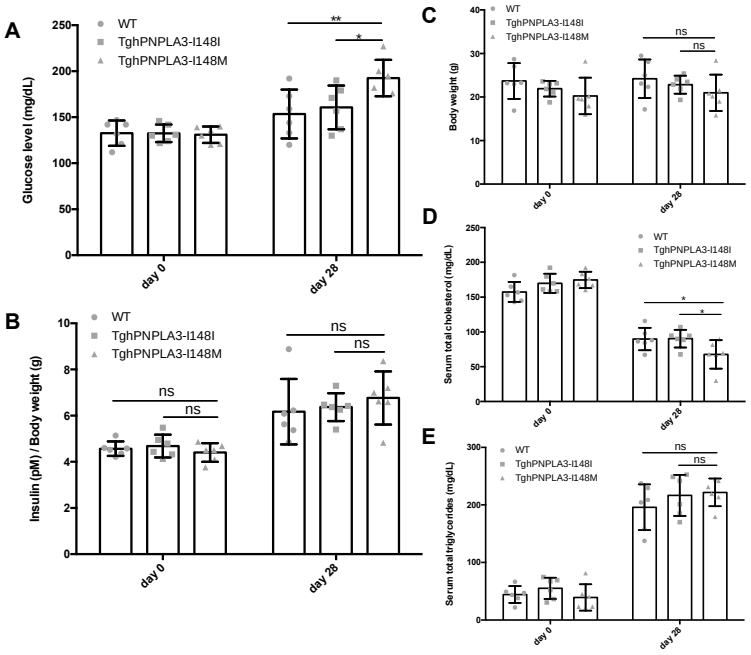


Figure 4.4 Effect of PNPLA3 I148M mutant on T2D and obesity with an HSD diet.

(A) Glucose levels of TghPNPLA3–I148I, TghPNPLA3–I148M, and non–transgenic wide type mice fed with a high sucrose diet (HSD) for 4 weeks. (B) Insulin levels. We normalized the insulin level to the body weight since the age of the TghPNPLA3–I148M group is younger than the TgHPNPLA3–I148I and non–transgenic control groups. We found that the insulin level, especially at the baseline, is highly correlated with body weight (**Figure 4.5 A**). The non–normalized data were also presented in Figure 4.5 B. (C) Body weight (D) Serum total cholesterol levels (E) Serum total triglycerides levels of the three groups fed with an HSD diet for 4 weeks. Error bar represents standard deviation (SD). *: Tukey adjusted $p<0.05$; **: Tukey adjusted $p<0.01$, ns: not significant.

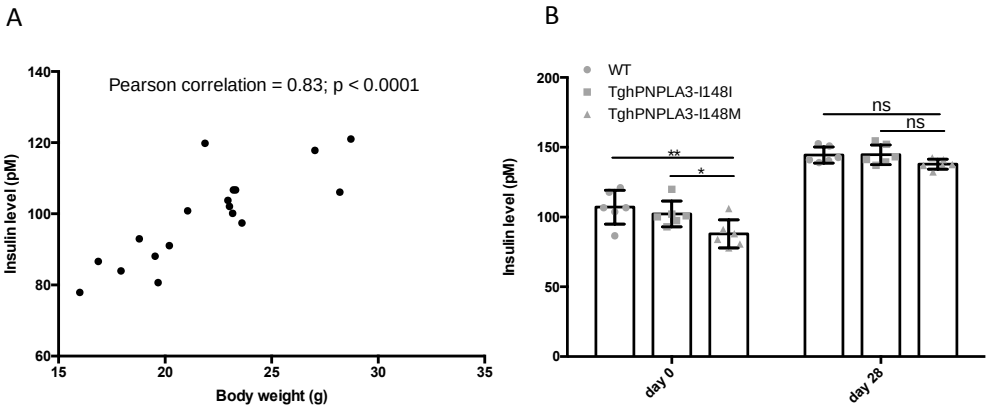


Figure 4.5 Insulin levels and body weight of the mice fed with an HSD diet

(A) Pearson correlation between insulin levels and body weight at the baseline of the HSD diet. (B) Insulin levels of TghPNPLA3–I148I, TghPNPLA3–I148M, and non–transgenic wide type mice without normalizing to the body weight. Error bar represents standard deviation (SD). *: Tukey adjusted $p<0.05$; **: Tukey adjusted $p<0.01$; ns: not significant.

To examine the effect of NASH progression on the susceptibility to T2D and obesity, we fed the mice with the NASH–inducing HFFC diet for 20 weeks. The TghPNPLA3–I148M mice developed significantly more severe NAFLD/NASH phenotypes as compared to the TghPNPLA3–I148I littermates, as characterized by increased inflammation and fibrosis (Figures 4.6–4.8), confirming that PNPLA3 I148M possesses a strong genetic predisposition to NAFLD and NASH. H&E staining results indicated that all the three groups have developed hepatosteatosis after 20 weeks of HFFC diet (Figure 4.9).

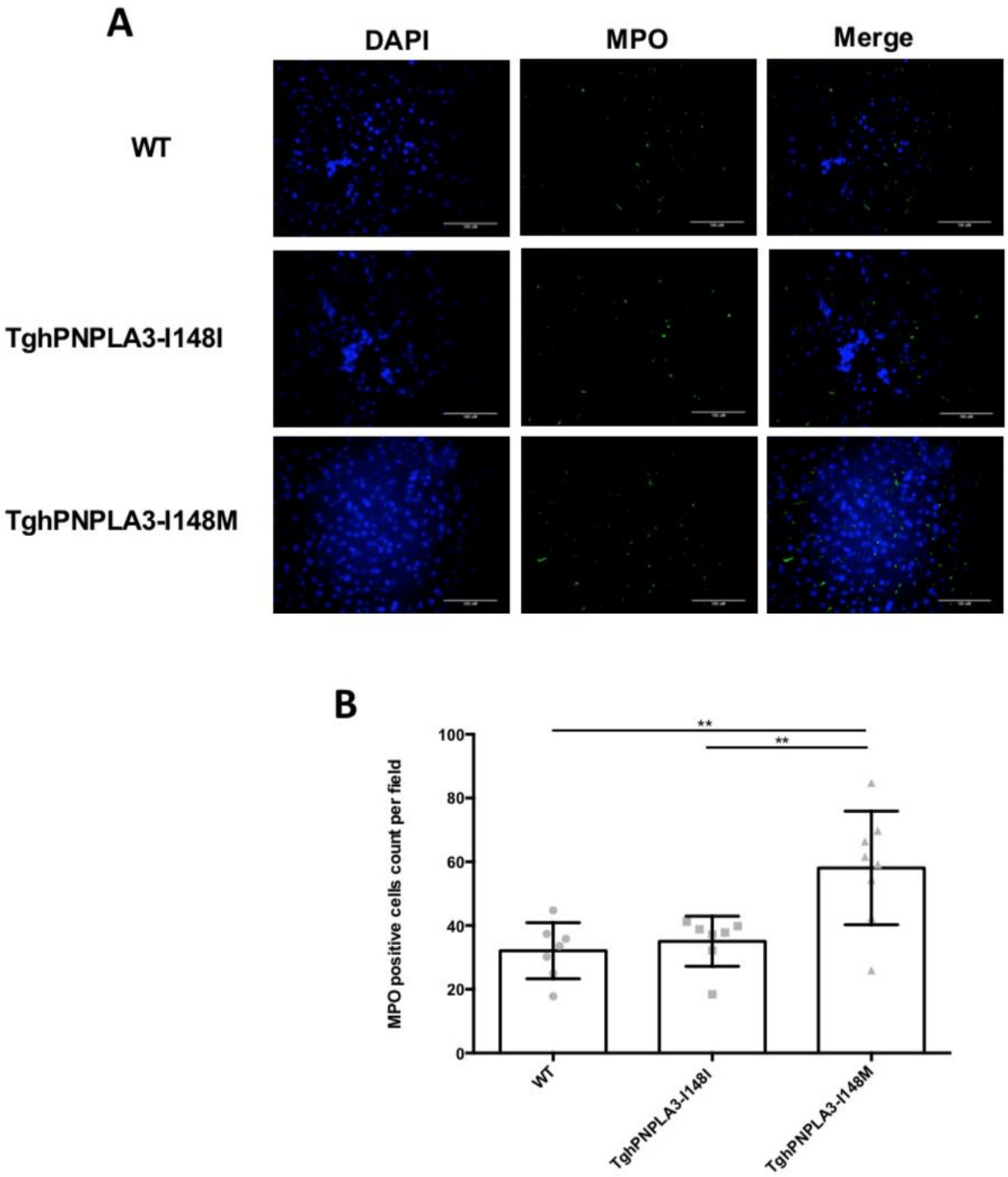


Figure 4.6 MPO staining of mice fed with an HFFC diet for 20 weeks

(A) Representative images of MPO immunofluorescence staining. (B) Counts of MPO positive cells in TghPNPLA3–I148I, TghPNPLA3–I148M, and non–transgenic wide type controls. The positive cells were counted in randomly selected fields (five fields per section). Error bar represents standard deviation (SD). *: Tukey adjusted $p<0.05$; ***: Tukey adjusted $p<0.001$.

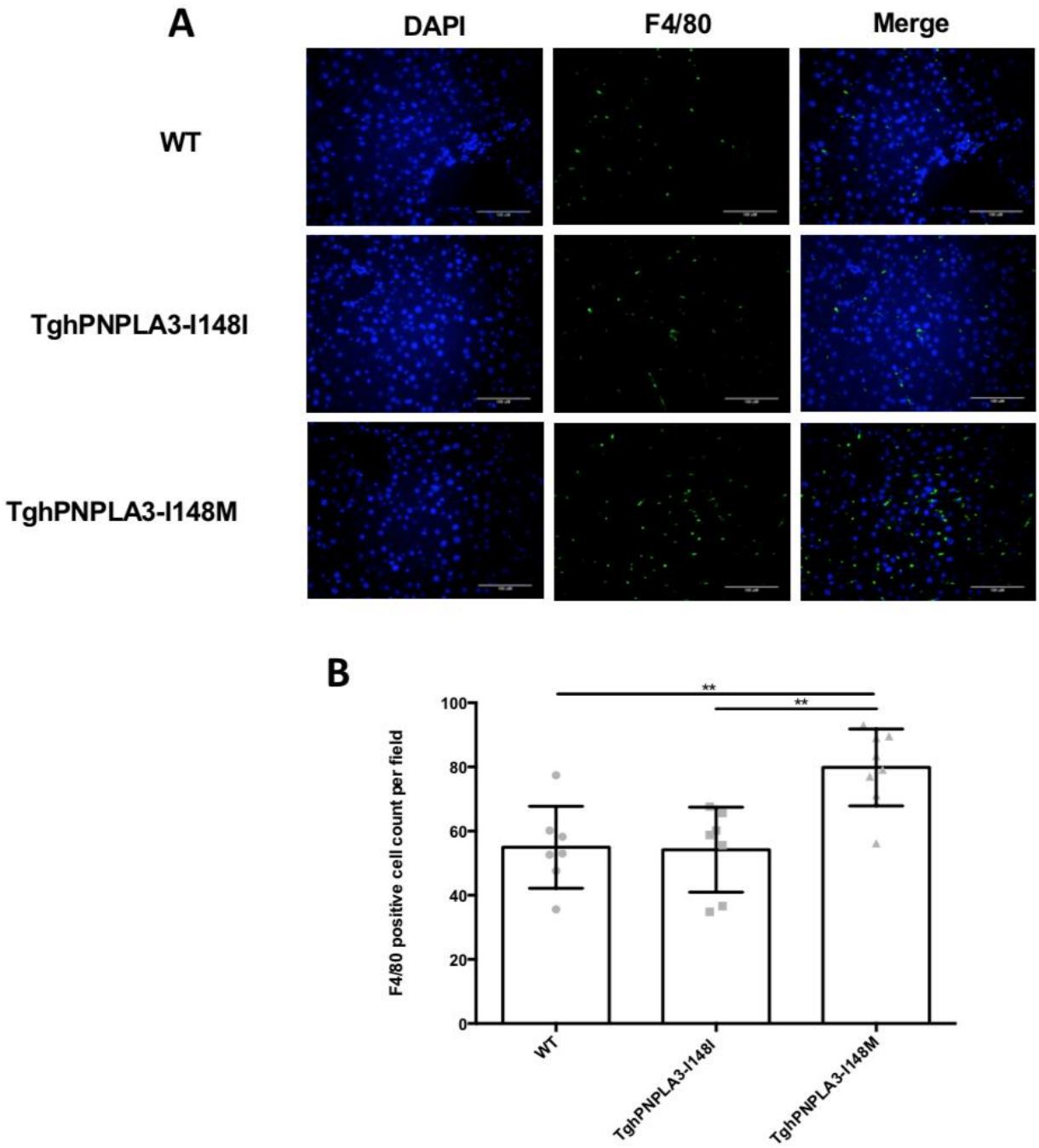


Figure 4.7 F4/80 staining of mice fed with an HFFC diet for 20 weeks

(A) Representative images of F4/80 immunofluorescence staining. (B) Counts of F4/80 positive cells in TghPNPLA3–I148I, TghPNPLA3–I148M, and non–transgenic wide type controls. The positive cells were counted in randomly selected fields (five fields per section). Error bar represents standard deviation (SD). *: Tukey adjusted $p<0.05$; ***: Tukey adjusted $p<0.001$.

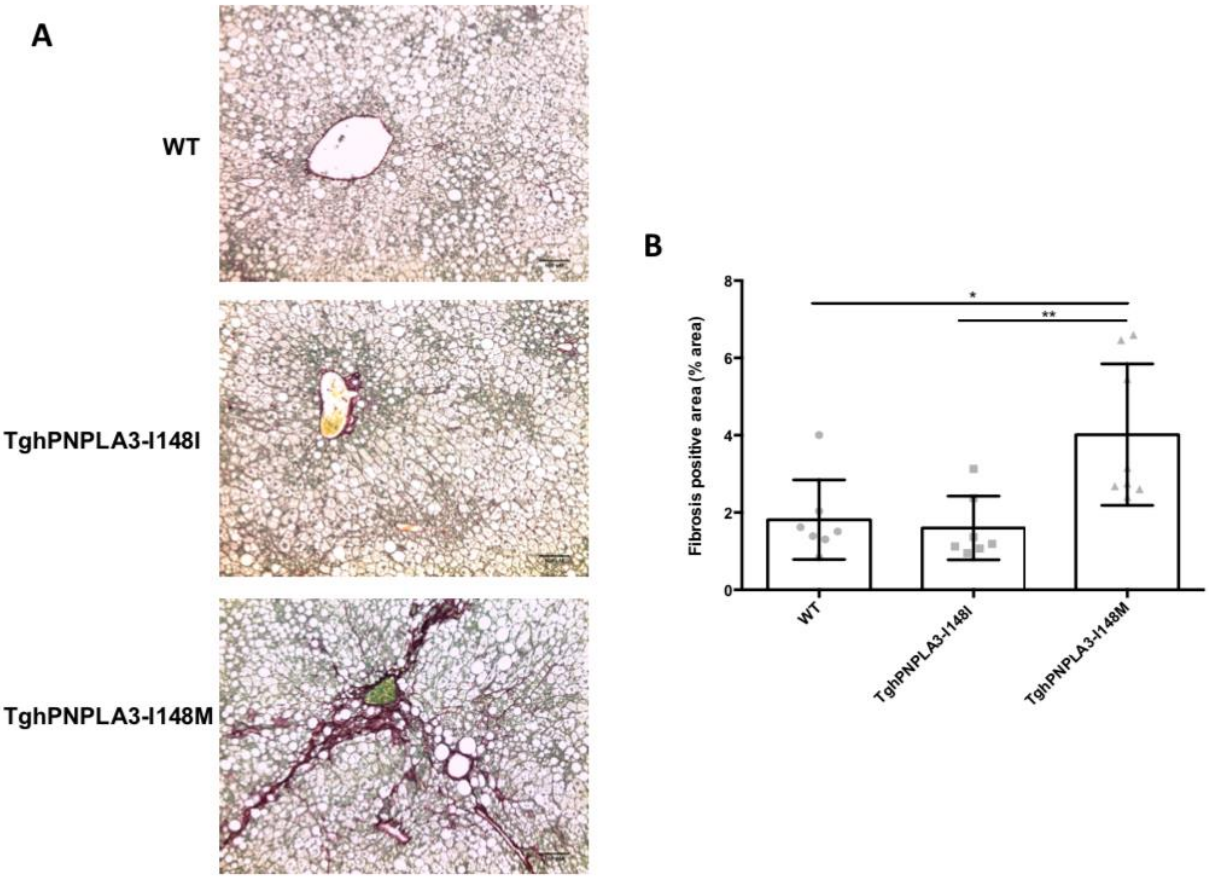


Figure 4.8 Sirius red staining of mice fed with an HFFC diet for 20 weeks

(A) Representative images of Sirius red staining. (B) Percentage of fibrosis positive area in TghPNPLA3–I148I, TghPNPLA3–I148M, and non–transgenic wide type mice. Error bar represents standard deviation (SD). *: Tukey adjusted $p<0.05$; **: Tukey adjusted $p<0.01$.

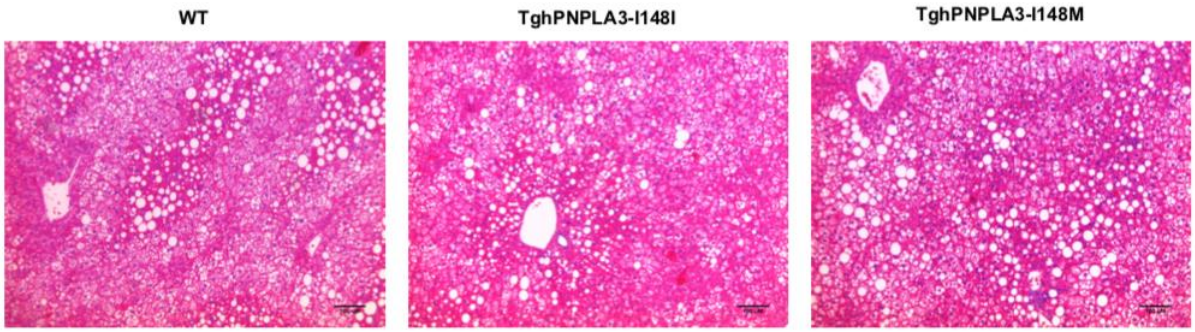


Figure 4.9 H&E staining of mice fed with an HFFC diet for 20 weeks

H&E staining of liver sections of TghPNPLA3–I148I, TghPNPLA3–I148M, and non–transgenic wide type mice.

We next evaluated the effect of PNPLA3 I148M on glucose homeostasis over a 20–week follow–up. As shown in **Figure 4.10 A**, we observed a significant genotype–time interaction on the fasting glucose level (two–way repeated measure ANOVA, $p<0{\cdot}0001$), suggesting that the effect of PNPLA3 I148M on glucose levels depended on disease progression. At week 16 and 18, the TghPNPLA3–I148M mice displayed significantly higher fasting glucose levels than their TghPNPLA3–I148I littermates ($p=0{\cdot}0048$ and $0{\cdot}0082$, respectively). There is also a significant interaction between genotype and time on fasting insulin levels between the TghPNPLA3 I148I and TghPNPLA3 I148M groups ($p=0{\cdot}0009$). However, there is no significant difference between the TghPNPLA3–I148M and non–transgenic wildtype mice. Beginning at week 12, the insulin levels between the latter two groups remain unchanged (Figure 4.10 B). Results of Glucose tolerance test (GTT) showed that TghPNPLA3–I148M mice experienced a reduced clearance of blood glucose as compared to the TghPNPLA3–I148I controls ($p=0{\cdot}012$) (Figure 4.10 C). However, we did not observe a significant difference in the response to insulin challenge in the Insulin tolerance test (ITT) (Figure 4.10 D).

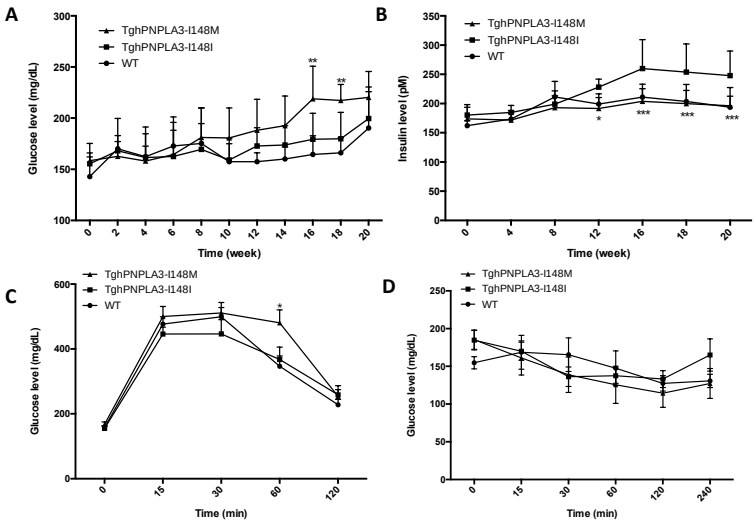


Figure 4.10 Effect of PNPLA3 I148M mutant on glucose and insulin levels with an HFFC diet

(A) Change in glucose levels over time of TghPNPLA3–I148I, TghPNPLA3–I148M, and non–transgenic wide type mice fed with a high–fat, high–fructose, high–cholesterol (HFFC) diet for 20 weeks. (B) Change in insulin levels for 20 weeks. (C) glucose tolerance test (GTT) and (D) insulin tolerance test (ITT) were performed at the 16th week of HFFC diet feeding. Error bar represents standard deviation (SD). The significance level of the comparison between TghPNPLA3–I148I and TghPNPLA3–I148M was indicated as follows: *: Tukey adjusted p<0.05; **: Tukey adjusted p<0.01; ***: Tukey adjusted p<0.001.

The body weight change of the mice on HFFC diet over time was shown in Figure 4.11 A. The TghPNPLA3–I148M mice were significantly lighter than the TghPNPLA3–I148I controls beginning at week 15 (all p<0·05). Magnetic resonance imaging (MRI) examination at week 20 showed that less total fat was accumulated in the TghPNPLA3–I148M mice than the TghPNPLA3–I148I controls (p=0·012), while the lean mass of non–fat tissues was not significantly different (Figure 4.11 B). Further investigation on the composition of the isolated fat showed that there was a significantly more epididymal white adipose tissue (EWAT) accumulation relative to the total peripheral adipose tissue in the TghPNPLA3–I148M mice than their TghPNPLA3–I148I littermates or non–transgenic controls (p=0·034 and 0·0015, respectively) (Figure 4.11 C). Examining the plasma lipid profile showed a significant decrease in total cholesterol levels beginning at week 16 for TghPNPLA3–I148M mice as compared to the

TghPNPLA3–I148I controls (Figure 4.11 D). No significant difference in TG levels between the three groups was observed (Figure 4.11 E).

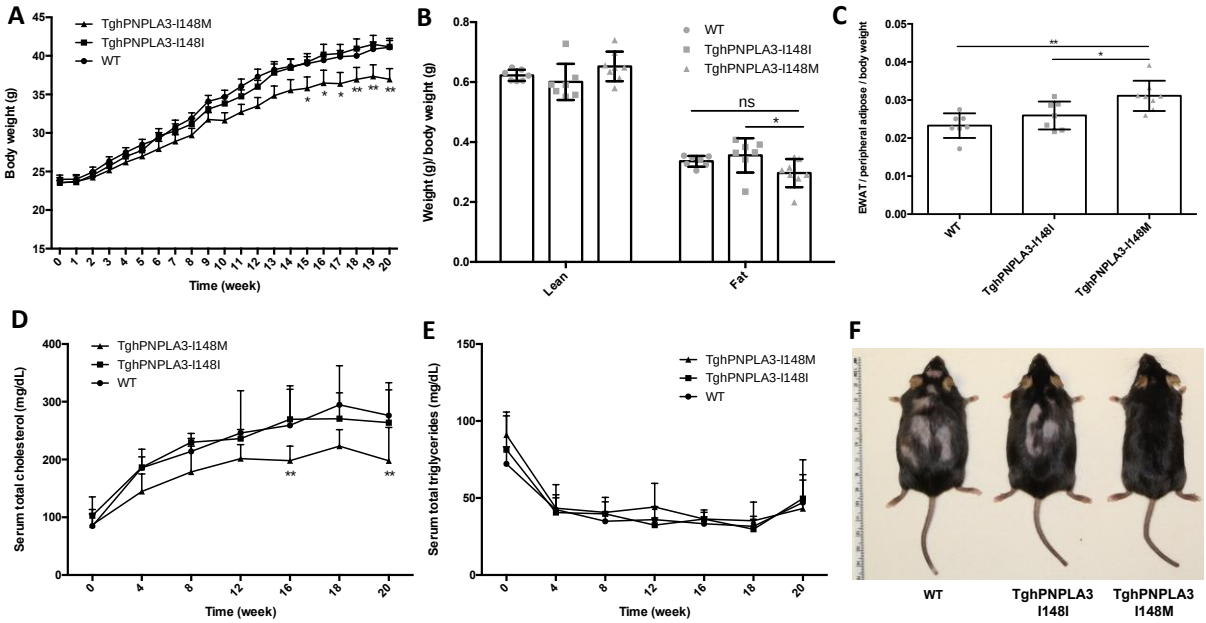


Figure 4.11 Effect of PNPLA3 I148M mutant on body weight, fat composition, and lipid profiles with an HFFC diet

(A) Change in body weight of TghPNPLA3–I148I, TghPNPLA3–I148M, and non–transgenic wide type mice fed with a high–fat, high–fructose, high–cholesterol (HFFC) diet for 20 weeks. (B) Body composition analysis by magnetic resonance imaging (MRI). Fat tissue weight and non–fat lean mass were normalized by the body weight (C) Epididymal white adipose tissue (EWAT) accumulation at the 20th week of HFFC diet feeding. EWAT accumulation was calculated as weight of EWAT divided the total peripheral adipose tissue weight and then normalized by the body weight. (D) Change in serum total cholesterol levels over 20 weeks (E) Change in serum total triglycerides levels over 20 weeks. (D) The representative image of TghPNPLA3–I148I, TghPNPLA3–I148M, and non–transgenic wide type mouse at the 20th week of HFFC diet feeding. Error bar represents standard deviation (SD). The significance level of the comparison between TghPNPLA3–I148I and TghPNPLA3–I148M was indicated as follows: *: Tukey adjusted $P<0.05$; **: Tukey adjusted $P<0.01$; ns: not significant.

4.5 Discussion

This is the first study to delineate the causal inter–relationship between NAFLD, T2D, and obesity using bidirectional MR. This relationship was also experimentally validated using murine models, which further demonstrated a causal relationship between hepatic steatosis or NASH and the susceptibility to T2D and obesity. We found that while genetically driven NAFLD is a causal

risk factor for T2D, it protects against overall obesity (indexed by BMI). However, NAFLD SNPs causally increase the risk for central obesity. On the other hand, genetically driven T2D and central or general obesity are all causal for fatty liver disease. Our study thus suggests that genetically driven NAFLD (likely "lean NAFLD") and the "metabolic NAFLD" attributed to T2D and/or obesity may be different diseases. Similarly, "NAFLD–driven T2D" may be different from T2D caused by other factors as well. Given the active drug development in these areas, it is important to clarify the right disease subtypes to be targeted. Our findings hence have important implications to clinical management, biomedical research and the development of precision medicine for the three diseases.

It has been long regarded that NAFLD and NASH are the central manifestations of T2D and obesity. Due to the lack of data in the natural history of NAFLD/NASH in humans, it remains largely unclear with regard to the causal relationship between NAFLD and T2D or obesity. Without addressing this issue, it is difficult to answer many key questions, e.g. shall we manage the NAFLD in preventing/treating T2D and obesity or vice versa? More importantly, while there is significant heterogeneity in the etiology of each of the three diseases, what would be the right management strategy for the right patient? Delineating this causal relationship is key to precision disease prevention and treatment. With the findings in our study, it is now possible to further dissect the three diseases into important subtypes. First, we observed that genetically driven NAFLD is a causal risk factor for T2D. Our study confirmed the findings of a previous small–scale MR analysis (Dongiovanni et al., 2018), and is in line with the results from numerous observational studies, including a recent meta–analysis of 19 observational studies with 296,439 individuals which indicated that patients with NAFLD had a two–fold risk of developing T2D

than those without NAFLD (Mantovani et al., 2018). More importantly, our animal model study on two different diets also consistently validated that PNPLA3–I148M–driven hepatosteatosis or NASH increases the risk for hyperglycemia. Therefore, based on our study T2D is likely to be divided into at least two subgroups, i.e. a genetically driven NAFLD–associated T2D ("NAFLD driven–T2D") and T2D due to other etiologies. This suggests that at least a significant proportion of T2D patients or pre–diabetic individuals, i.e. those who carry the PNPLA3–I148M or TM6SF2 variants and develop NAFLD should benefit from an intervention aiming at reducing hepatic steatosis or more advanced liver perturbations. This is particularly consistent with findings of multiple longitudinal studies in Asian populations where nondiabetic individuals with hepatic steatosis at baseline have 2.78–fold increased risk for T2D after 5 years of follow–up; while nondiabetic individuals with steatosis at baseline but not at the follow–up did not demonstrate an increased risk for T2D (Lee et al., 2019; Sung, Wild, & Byrne, 2013). On the other hand, our findings also demonstrated that T2D is a risk factor for at least a subset of NAFLD, hence, a "T2D–driven NAFLD" that may be a comorbidity of diabetes mellitus. Previous population–based studies suggested that hyperglycemia could be a factor for progression to liver fibrosis, and T2D can enhance the liver stiffness (see review(Lee et al., 2019)). T2D may also modify the genetic risk of PNPLA3 I148M for NAFLD as well. High glucose level can increase the expression of PNPLA3 via regulating carbohydrate–response element–binding protein (ChREBP) (Y. Huang et al., 2010), which is a necessary step for the accumulation of PNPLA3–I148M protein on the surface of lipid droplet in hepatocytes (J. Z. Li et al., 2012; W. Liu et al., 2016; Smagris et al., 2015). This accumulation further alters the dynamics of hydrolysis of triglycerides and lead to hepatic steatosis (Y. Wang, Kory, BasuRay,

Cohen, & Hobbs, 2019). Therefore, individuals who have a high risk for T2D should consider early intervention to further prevent liver injuries.

Our study also revealed an interesting relationship between NAFLD and obesity. Our analysis suggests that genetically driven liver–specific fat storage or deposition may remodel the fat distribution in the whole body. Although epidemiologically obesity is highly correlated with NAFLD, the genetically instrumented NAFLD is actually not a causal risk factor for overall obesity. Rather, it protects against the overall BMI elevation. However, genetically driven NAFLD causally increases risks for central obesity, characterized by an increased waist–to–hip ratio in humans. Our observation is consistent with numerous observational studies on the associations between NAFLD, visceral fat or WHR and BMI (Pang et al., 2019; Radmard et al., 2016). Our mouse study also accurately recapitulated these relationships. Interestingly, both MR and animal studies indicated that genetic NAFLD causally leads to decreased total cholesterol (TC) but not triglycerides (TG). Therefore, central obesity, at least in part, may possess a subtype attributed to genetically–driven NAFLD. This is particularly significant, as both TC and TG are generally correlated with visceral fat as demonstrated in many studies (Luo et al., 2014; Odamaki et al., 1999; Sadeghi et al., 2013). The dissociation in our study suggesting that the NAFLD–driven central obesity is likely a unique subtype.

More importantly, our study corroborates the widely discussed hypothesis about “lean NAFLD” and “obese NAFLD”. Both cross–sectional and longitudinal analyses over the past few years have observed significant differences in disease manifestation, progression and clinical outcomes between lean NAFLD and obese NAFLD (Feldman et al., 2017; Hagstrom et al., 2018; Lu et al.,

2018). Our study echoes this observation and indicates that genetically driven NAFLD may be more likely progressed to lean NAFLD (but not necessarily reduces visceral fat accumulation) while does not promote the development of overall obesity. This is also consistent with the observation that PNPLA3 I148M allele is more significantly associated with lean NAFLD (Feldman et al., 2017; Lu et al., 2018). Taken together, our study suggests that so–called "lean NAFLD" is not just a phenomenon, but is a disease phenotype based on a certain causal mechanism. On the other hand, our MR analysis demonstrated that both genetically instrumented BMI and WHRadjBMI causally increase the risk for NAFLD, suggesting adiposity can be a risk factor for at least a subset of NAFLD, i.e. "obese NAFLD", which may synergize the effects of genetic risk factors on the development of NASH or more severe liver injuries (Stender et al., 2017), but could also play an independent role. In the latter case, NAFLD is thus more likely a comorbidity of obesity. Therefore, while the management of the lean NAFLD may require more attention to the liver, reducing weight and BMI would be critical to the prevention of obese NAFLD. Collectively, it is now clearer that, depending on the contributing load of genetic risk to each of NAFLD, T2D, and obesity, these diseases should be further dissected into subgroups based on their causal inter–relationship: e.g. NAFLD should be further classified into "genetically–driven NAFLD" or "hepatic NAFLD", and "metabolic NAFLD" or "systemic NAFLD, while T2D should be also considered at least to be divided into subgroups that are attributed to the hepatic perturbations and extrahepatic modifications (e.g. pancreas), respectively. There should be a NAFLD–driven central obesity as well. The causal relationships among the three diseases and main findings of the study were summarized in Figure 4.12. In future studies, it would be critical to explore how to distinguish these subgroups and identify the

high–risk individuals, so that prevention or treatment strategies for each condition can be developed accordingly.

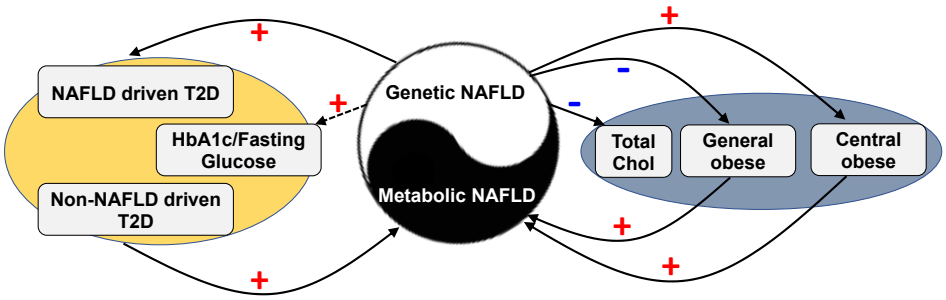


Figure 4.12 Schematic presentation of the causal relationships among NAFLD, T2D, and obesity. "+": positive relationship; "–": negative relationship; dashed line represents the suggestive causal relationship.

Despite the significance of our findings, the detailed molecular mechanism underlying these causal relationships remains to be further investigated. For the causal role of NAFLD in increasing the T2D risk, our MR analysis found a weak causal relationship between genetic NAFLD and fasting glucose and fasting insulin levels. A few hypotheses underlying this NAFLD–T2D relationship were proposed recently, for example, Fetuin–B can cause glucose intolerance, which was deemed as a very convincing mediating mechanism (El-Ashmawy & Ahmed, 2019; Lin et al., 2019; Lonardo et al., 2019; Meex & Watt, 2017; Peter et al., 2018). In addition, recent MR analyses have demonstrated that central obesity, indexed by increased WHR, is causal to hyperglycemia, whereas the overall obesity is causal to hyperinsulinemia (T. Wang et al., 2016; T. Wang et al., 2018), further highlighting the complex interrelationship among these three diseases, as well as their etiological heterogeneity. Whether genetically instrumented NAFLD independently increases the susceptibility to T2D and central obesity or actually affects one after another remains further investigation. Inversely, the mechanisms underlying the causal role of T2D in increasing risk for NAFLD are also incompletely understood, with gut microbiota and expanded, inflamed, dysfunctional adipose tissue as popular

hypotheses at present (Lonardo et al., 2019). However, it remains to be further explored whether these potential mechanisms involve the PNPLA3 or TM6SF2 related signaling. Similarly, the causal mechanism underlying the association between genetically driven NAFLD and reduced BMI and total cholesterol but increased abdominal fat accumulation is also largely unclear. Our study warrants continued investigation into these mechanisms.

Our study also has important indications to biomedical research for the three diseases, which is especially pivotal to targets identification and drug development. For instance, the causal pathways leading to "hepatic" or lean NAFLD perhaps are distinct from those of the "metabolic NAFLD". Selecting the right model would be hence very critical to the success of drug development. Similarly, as mentioned above, determining the correct model for T2D or obesity based on the potential causes may be also key to warrant effective research.

Our study has several limitations. First, we only selected individuals of European descent to avoid the potential confounding effects of population structure. Thus, the findings in our study need to be validated in other ethnic groups; Second, although PNPLA3 and TM6SF2 (NCAN) were two strongest and the most compelling causal variants for NAFLD, the limited number of SNPs prohibits the use of MR–PRESSO and modified Q’ statistics to test the presence of pleiotropic effects. In fact, TM6SF2 variant has been found to be associated with multiple metabolic factors especially lipoprotein lipid profile, therefore may possess potential pleiotropy effects. In addition, TM6SF2 rs58542926 has a relatively lower allele frequency (7% in Europeans) and is also missed from multiple commonly used genotyping platforms. These facts limit its potential as an ideal proxy for an MR analysis. Although our GWAS using the UK

Biobank produced more loci that are potentially associated with NAFLD, after a stringent selection, the PNPLA3 and TM6SF2/NCAN loci polymorphisms remain to be the most reliable markers as an instrument for NAFLD. Therefore, it is possible that the causal relationships we have identified in this analysis are only limited to these two genes. In addition, the UK biobank samples only have limited clinical information. We can only focus on the ICD codes for "fatty liver disease" as a phenotype. Although we tried our best to remove other potential confounding factors, e.g. individuals with other known liver disease or hepatitis, the phenotype may not exactly reflect the strictly defined "non–alcoholic fatty liver disease". However, both PNPLA3 and TM6SF2 loci were successfully identified as top hits in this study (Table 4.8), suggesting it does reflect the genetic patterns underlying NAFLD. Therefore, the reverse MR to test the causal role of T2D or obesity in increasing risks for NAFLD are likely reliable. Future studies should consider using a large sample set with well–defined NAFLD/NASH phenotypes to further validate these relationships. Moreover, due to the complexity of the biological systems, the existence of the reciprocal feedback loops might mask the truly bidirectional causal relationship. Future studies should consider using alternative approaches such as structural equation modeling to estimate the effect of feedback loops (Evans & Davey Smith, 2015). Further studies on the contributing load of genetic risk to each of the three diseases and their interactions with the environmental factors are necessary to apply the findings at the individual level.

In summary, our study combined both a bidirectional MR analysis with the largest–to–date sample sets and animal models to delineate the causal relationships between NAFLD, T2D, and obesity. Our findings provided strong evidence for disease subphenotyping and corroborated key hypotheses previously generated from observational studies on these three important diseases.

These identifications warranted new directions to the development of precision preventive and therapeutic strategies for the three diseases.

Table 4.8 Characteristics of the associations of PNPLA3 rs738409, NCAN rs2228603, and TM6SF2 rs58542926 with NAFLD in UKBB samples

Gene	**SNP**	**logOR**	**SE**	**P value**
PNPLA3	rs738409	0.33	0.053	2.09E–10
NCAN	rs2228603	0.32	0.081	6.83E–05
TM6SF2	rs58542926	0.39	0.082	2.69E–06

OR: Odds ratio; SE: standard error.

REFERENCES

Ahola-Olli, A. V., Wurtz, P., Havulinna, A. S., Aalto, K., Pitkanen, N., Lehtimaki, T., . . . Raitakari, O. T. (2017). Genome-wide Association Study Identifies 27 Loci Influencing Concentrations of Circulating Cytokines and Growth Factors. *American Journal Of Human Genetics, 100*(1), 40-50. doi:10.1016/j.ajhg.2016.11.007

Almasio, P. L., Licata, A., Maida, M., Macaluso, F. S., Costantino, A., Alessi, N., . . . Craxi, A. (2016). Clinical Course and Genetic Susceptibility of Primary Biliary Cirrhosis: Analysis of a Prospective Cohort. *Hepat Mon, 16*(11), e31681. doi:10.5812/hepatmon.31681

Amin, N., van Duijn, C. M., & Janssens, A. C. (2009). Genetic scoring analysis: a way forward in genome wide association studies? *European Journal Of Epidemiology, 24*(10), 585-587. doi:10.1007/s10654-009-9387-y

Baralle, F. E., & Giudice, J. (2017). Alternative splicing as a regulator of development and tissue identity. *Nat Rev Mol Cell Biol, 18*(7), 437-451. doi:10.1038/nrm.2017.27

Bataller, R., & Brenner, D. A. (2005). Liver fibrosis. *Journal Of Clinical Investigation, 115*(2), 209-218. doi:10.1172/jci24282

Bell, C. G., & Beck, S. (2009). Advances in the identification and analysis of allele-specific expression. *Genome Med, 1*(5), 56. doi:10.1186/gm56

Benyon, R. C., & Iredale, J. P. (2000). Is liver fibrosis reversible? *Gut, 46*(4), 443-446.

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics, 30*(15), 2114-2120. doi:10.1093/bioinformatics/btu170

Bowden, J., Davey Smith, G., & Burgess, S. (2015). Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *Int J Epidemiol, 44*(2), 512-525. doi:10.1093/ije/dyv080

Bowden, J., Davey Smith, G., Haycock, P. C., & Burgess, S. (2016). Consistent Estimation in Mendelian Randomization with Some Invalid Instruments Using a Weighted Median Estimator. *Genet Epidemiol, 40*(4), 304-314. doi:10.1002/gepi.21965

Bowden, J., Del Greco, M. F., Minelli, C., Davey Smith, G., Sheehan, N. A., & Thompson, J. R. (2016). Assessing the suitability of summary data for two-sample Mendelian randomization analyses using MR-Egger regression: the role of the I2 statistic. *Int J Epidemiol, 45*(6), 1961-1974. doi:10.1093/ije/dyw220

Bowden, J., Del Greco, M. F., Minelli, C., Zhao, Q., Lawlor, D. A., Sheehan, N. A., . . . Davey Smith, G. (2018). Improving the accuracy of two-sample summary-data Mendelian randomization: moving beyond the NOME assumption. *Int J Epidemiol*. doi:10.1093/ije/dyy258

Bowden, J., Spiller, W., Del Greco, M. F., Sheehan, N., Thompson, J., Minelli, C., & Davey Smith, G. (2018). Improving the visualization, interpretation and analysis of two-sample summary data Mendelian randomization via the Radial plot and Radial regression. *Int J Epidemiol, 47*(6), 2100. doi:10.1093/ije/dyy265

Brinegar, A. E., Xia, Z., Loehr, J. A., Li, W., Rodney, G. G., & Cooper, T. A. (2017). Extensive alternative splicing transitions during postnatal skeletal muscle development are required for calcium handling functions. *Elife, 6*. doi:10.7554/eLife.27192

Buensuceso, A. V., & Deroo, B. J. (2013). The ephrin signaling pathway regulates morphology and adhesion of mouse granulosa cells in vitro. *Biology of Reproduction, 88*(1), 25. doi:10.1095/biolreprod.112.100123

Burgess, S., Butterworth, A., & Thompson, S. G. (2013). Mendelian Randomization Analysis With Multiple Genetic Variants Using Summarized Data. *Genetic Epidemiology, 37*(7), 658-665. doi:10.1002/gepi.21758

Burgess, S., Davies, N. M., & Thompson, S. G. (2016). Bias due to participant overlap in two-sample Mendelian randomization. *Genet Epidemiol, 40*(7), 597-608. doi:10.1002/gepi.21998

Cardoso-Moreira, M., Halbert, J., Valloton, D., Velten, B., Chen, C., Shao, Y., . . . Kaessmann, H. (2019). Gene expression across mammalian organ development. *Nature*. doi:10.1038/s41586-019-1338-5

Castel, S. E., Levy-Moonshine, A., Mohammadi, P., Banks, E., & Lappalainen, T. (2015). Tools and best practices for data processing in allelic expression analysis. *Genome Biology, 16*, 195. doi:10.1186/s13059-015-0762-6

Castel, S. E., Levy-Moonshine, A., Mohammadii, P., Banks, E., & Lappalainenii, T. (2015). Tools and best practices for data processing in allelic expression analysis. *Genome Biology, 16*. doi:ARTN 195 10.1186/s13059-015-0762-6

Chalasani, N., Younossi, Z., Lavine, J. E., Charlton, M., Cusi, K., Rinella, M., . . . Sanyal, A. J. (2018). The diagnosis and management of nonalcoholic fatty liver disease: Practice guidance from the American Association for the Study of Liver Diseases. *Hepatology, 67*(1), 328-357. doi:10.1002/hep.29367

Chatterjee, N., Shi, J., & Garcia-Closas, M. (2016). Developing and evaluating polygenic risk prediction models for stratified disease prevention. *Nat Rev Genet, 17*(7), 392-406. doi:10.1038/nrg.2016.27

Chen, J., Rozowsky, J., Galeev, T. R., Harmanci, A., Kitchen, R., Bedford, J., . . . Gerstein, M. (2016). A uniform survey of allele-specific binding and expression over 1000-Genomes-Project individuals. *Nature Communications, 7*, 11101. doi:10.1038/ncomms11101

Chen, J. J., Patel, A., Sodani, K., Xiao, Z. J., Tiwari, A. K., Zhang, D. M., . . . Chen, Z. S. (2013). bba, a synthetic derivative of 23-hydroxybutulinic acid, reverses multidrug resistance by inhibiting the efflux activity of MRP7 (ABCC10). *PLoS One, 8*(9), e74573. doi:10.1371/journal.pone.0074573

Chen, Q. R., Braun, R., Hu, Y., Yan, C., Brunt, E. M., Meerzaman, D., . . . Buetow, K. (2013). Multi-SNP analysis of GWAS data identifies pathways associated with nonalcoholic fatty liver disease. *PloS One, 8*(7), e65982. doi:10.1371/journal.pone.0065982

Clapper, J. R., Hendricks, M. D., Gu, G., Wittmer, C., Dolman, C. S., Herich, J., . . . Roth, J. D. (2013). Diet-induced mouse model of fatty liver disease and nonalcoholic steatohepatitis reflecting clinical disease progression and methods of assessment. *Am J Physiol Gastrointest Liver Physiol, 305*(7), G483-495. doi:10.1152/ajpgi.00079.2013

Consortium, G. T., Laboratory, D. A., Coordinating Center -Analysis Working, G., Statistical Methods groups-Analysis Working, G., Enhancing, G. g., Fund, N. I. H. C., . . . Montgomery, S. B. (2017). Genetic effects on gene expression across human tissues. *Nature, 550*(7675), 204-213. doi:10.1038/nature24277

Das, S., Forer, L., Schonherr, S., Sidore, C., Locke, A. E., Kwong, A., . . . Fuchsberger, C. (2016). Next-generation genotype imputation service and methods. *Nat Genet, 48*(10), 1284-1287. doi:10.1038/ng.3656

De La Vega, F. M., & Bustamante, C. D. (2018). Polygenic risk scores: a biased prediction? *Genome Med, 10*(1), 100. doi:10.1186/s13073-018-0610-x

De Silva, N. M. G., Borges, M. C., Hingorani, A., Engmann, J., Shah, T., Zhang, X., . . . Lawlor, D. A. (2019). Liver Function and Risk of Type 2 Diabetes: Bidirectional Mendelian Randomization Study. *Diabetes*. doi:10.2337/db18-1048

Denk, H., Abuja, P. M., & Zatloukal, K. (2018). Animal models of NAFLD from the pathologist's point of view. *Biochim Biophys Acta Mol Basis Dis*. doi:10.1016/j.bbadis.2018.04.024

Dobin, A., & Gingeras, T. R. (2015). Mapping RNA-seq Reads with STAR. *Curr Protoc Bioinformatics, 51*, 11 14 11-19. doi:10.1002/0471250953.bi1114s51

Dongiovanni, P., Stender, S., Pietrelli, A., Mancina, R. M., Cespiati, A., Petta, S., . . . Valenti, L. (2018). Causal relationship of hepatic fat with liver damage and insulin resistance in nonalcoholic fatty liver. *J Intern Med, 283*(4), 356-370. doi:10.1111/joim.12719

Dunham, I., Kundaje, A., Aldred, S. F., Collins, P. J., Davis, C., Doyle, F., . . . Consortium, E. P. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature, 489*(7414), 57-74. doi:10.1038/nature11247

Dupuis, J., Langenberg, C., Prokopenko, I., Saxena, R., Soranzo, N., Jackson, A. U., . . . Barroso, I. (2010). New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. *Nat Genet, 42*(2), 105-116. doi:10.1038/ng.520

El-Ashmawy, H. M., & Ahmed, A. M. (2019). Serum fetuin-B level is an independent marker for nonalcoholic fatty liver disease in patients with type 2 diabetes. *Eur J Gastroenterol Hepatol*. doi:10.1097/MEG.0000000000001354

Elsharkawy, A. M., Oakley, F., & Mann, D. A. (2005). The role and regulation of hepatic stellate cell apoptosis in reversal of liver fibrosis. *Apoptosis, 10*(5), 927-939. doi:10.1007/s10495-005-1055-4

Eslam, M., Valenti, L., & Romeo, S. (2018). Genetics and epigenetics of NAFLD and NASH: Clinical impact. *J Hepatol, 68*(2), 268-279. doi:10.1016/j.jhep.2017.09.003

Evans, D. M., & Davey Smith, G. (2015). Mendelian Randomization: New Applications in the Coming Age of Hypothesis-Free Causality. *Annu Rev Genomics Hum Genet, 16*, 327-350. doi:10.1146/annurev-genom-090314-050016

Fabbrini, E., Sullivan, S., & Klein, S. (2010). Obesity and nonalcoholic fatty liver disease: biochemical, metabolic, and clinical implications. *Hepatology, 51*(2), 679-689. doi:10.1002/hep.23280

Feldman, A., Eder, S. K., Felder, T. K., Kedenko, L., Paulweber, B., Stadlmayr, A., . . . Aigner, E. (2017). Clinical and Metabolic Characterization of Lean Caucasian Subjects With Non-alcoholic Fatty Liver. *Am J Gastroenterol, 112*(1), 102-110. doi:10.1038/ajg.2016.318

Flutre, T., Wen, X., Pritchard, J., & Stephens, M. (2013). A statistical framework for joint eQTL analysis in multiple tissues. *PLoS Genet, 9*(5), e1003486. doi:10.1371/journal.pgen.1003486

Franke, A., McGovern, D. P., Barrett, J. C., Wang, K., Radford-Smith, G. L., Ahmad, T., . . . Parkes, M. (2010). Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nature Genetics, 42*(12), 1118-1125. doi:10.1038/ng.717

Gamazon, E. R., Innocenti, F., Wei, R., Wang, L., Zhang, M., Mirkov, S., . . . Liu, W. (2013). A genome-wide integrative study of microRNAs in human liver. *BMC Genomics, 14*, 395. doi:10.1186/1471-2164-14-395

Gamazon, E. R., Wheeler, H. E., Shah, K. P., Mozaffari, S. V., Aquino-Michaels, K., Carroll, R. J., . . . Im, H. K. (2015). A gene-based association method for mapping traits using reference transcriptome data. *Nat Genet, 47*(9), 1091-1098. doi:10.1038/ng.3367

Gandal, M. J., Zhang, P., Hadjimichael, E., Walker, R. L., Chen, C., Liu, S., . . . Geschwind, D. H. (2018). Transcriptome-wide isoform-level dysregulation in ASD, schizophrenia, and bipolar disorder. *Science, 362*(6420). doi:10.1126/science.aat8127

Genomes Project, C., Auton, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., Kang, H. M., . . . Abecasis, G. R. (2015). A global reference for human genetic variation. *Nature, 526*(7571), 68-74. doi:10.1038/nature15393

Gentles, A. J., Newman, A. M., Liu, C. L., Bratman, S. V., Feng, W., Kim, D., . . . Alizadeh, A. A. (2015). The prognostic landscape of genes and infiltrating immune cells across human cancers. *Nat Med, 21*(8), 938-945. doi:10.1038/nm.3909

Gerstberger, S., Hafner, M., & Tuschl, T. (2014). A census of human RNA-binding proteins. *Nat Rev Genet, 15*(12), 829-845. doi:10.1038/nrg3813

Gorden, A., Yang, R. Z., Yerges-Armstrong, L. M., Ryan, K. A., Speliotes, E., Borecki, I. B., . . . Consortium, G. (2013). Genetic Variation at NCAN Locus Is Associated with Inflammation and Fibrosis in Non-Alcoholic Fatty Liver Disease in Morbid Obesity. *Human Heredity, 75*(1), 34-43. doi:10.1159/000346195

Gordillo, M., Evans, T., & Gouon-Evans, V. (2015). Orchestrating liver development. *Development, 142*(12), 2094-2108. doi:10.1242/dev.114215

Gusev, A., Ko, A., Shi, H., Bhatia, G., Chung, W., Penninx, B. W., . . . Pasaniuc, B. (2016). Integrative approaches for large-scale transcriptome-wide association studies. *Nat Genet, 48*(3), 245-252. doi:10.1038/ng.3506

Hagstrom, H., Nasr, P., Ekstedt, M., Hammar, U., Stal, P., Hultcrantz, R., & Kechagias, S. (2018). Risk for development of severe liver disease in lean patients with nonalcoholic fatty liver disease: A long-term follow-up study. *Hepatol Commun, 2*(1), 48-57. doi:10.1002/hep4.1124

Hansen, K. D., Irizarry, R. A., & Wu, Z. (2012). Removing technical variability in RNA-seq data using conditional quantile normalization. *Biostatistics, 13*(2), 204-216. doi:10.1093/biostatistics/kxr054

Hardy, T., Oakley, F., Anstee, Q. M., & Day, C. P. (2016). Nonalcoholic Fatty Liver Disease: Pathogenesis and Disease Spectrum. *Annu Rev Pathol, 11*, 451-496. doi:10.1146/annurev-pathol-012615-044224

Hartwig, F. P., Davies, N. M., Hemani, G., & Davey Smith, G. (2016). Two-sample Mendelian randomization: avoiding the downsides of a powerful, widely applicable but potentially fallible technique. *Int J Epidemiol, 45*(6), 1717-1726. doi:10.1093/ije/dyx028

Harvey, C. T., Moyerbrailean, G. A., Davis, G. O., Wen, X. Q., Luca, F., & Pique-Regi, R. (2015). QuASAR: quantitative allele-specific analysis of reads. *Bioinformatics, 31*(8), 1235-1242. doi:10.1093/bioinformatics/btu802

Heinrichs, D., Knauel, M., Offermanns, C., Berres, M. L., Nellen, A., Leng, L., . . . Wasmuth, H. E. (2011). Macrophage migration inhibitory factor (MIF) exerts antifibrotic effects in experimental liver fibrosis via CD74. *Proceedings of the National Academy of Sciences of the United States of America, 108*(42), 17444-17449. doi:10.1073/pnas.1107023108

Hernandez-Gea, V., & Friedman, S. L. (2011). Pathogenesis of liver fibrosis. *Annual Review of Pathology, 6*, 425-456. doi:10.1146/annurev-pathol-011110-130246

Hinds, D. A., McMahon, G., Kiefer, A. K., Do, C. B., Eriksson, N., Evans, D. M., . . . Tung, J. Y. (2013). A genome-wide association meta-analysis of self-reported allergy identifies shared and allergy-specific susceptibility loci. *Nature Genetics, 45*(8), 907-911. doi:10.1038/ng.2686

Horani, A., Muhanna, N., Pappo, O., Melhem, A., Alvarez, C. E., Doron, S., . . . Safadi, R. (2007). Beneficial effect of glatiramer acetate (Copaxone) on immune modulation of experimental hepatic fibrosis. *American Journal of Physiology Gastrointestinal and Liver Physiology, 292*(2), G628-638. doi:10.1152/ajpgi.00137.2006

Hruby, A., & Hu, F. B. (2015). The Epidemiology of Obesity: A Big Picture. *Pharmacoeconomics, 33*(7), 673-689. doi:10.1007/s40273-014-0243-x

Hu, H., Miao, Y. R., Jia, L. H., Yu, Q. Y., Zhang, Q., & Guo, A. Y. (2019). AnimalTFDB 3.0: a comprehensive resource for annotation and prediction of animal transcription factors. *Nucleic Acids Research, 47*(D1), D33-D38. doi:10.1093/nar/gky822

Hu, M., Phan, F., Bourron, O., Ferre, P., & Foufelle, F. (2017). Steatosis and NASH in type 2 diabetes. *Biochimie, 143*, 37-41. doi:10.1016/j.biochi.2017.10.019

Huang, Q. Q., Ritchie, S. C., Brozynska, M., & Inouye, M. (2018). Power, false discovery rate and Winner's Curse in eQTL studies. *Nucleic Acids Research, 46*(22), e133. doi:10.1093/nar/gky780

Huang, X., Liu, G., Guo, J., & Su, Z. (2018). The PI3K/AKT pathway in obesity and type 2 diabetes. *Int J Biol Sci, 14*(11), 1483-1496. doi:10.7150/ijbs.27173

Huang, Y., de Boer, W. B., Adams, L. A., MacQuillan, G., Rossi, E., Rigby, P., . . . Jeffrey, G. P. (2013). Image analysis of liver collagen using sirius red is more accurate and correlates better with serum fibrosis markers than trichrome. *Liver Int, 33*(8), 1249-1256. doi:10.1111/liv.12184

Huang, Y., He, S., Li, J. Z., Seo, Y. K., Osborne, T. F., Cohen, J. C., & Hobbs, H. H. (2010). A feed-forward loop amplifies nutritional regulation of PNPLA3. *Proc Natl Acad Sci U S A, 107*(17), 7892-7897. doi:10.1073/pnas.1003585107

Huse, S. M., Gruppuso, P. A., Boekelheide, K., & Sanders, J. A. (2015). Patterns of gene expression and DNA methylation in human fetal and adult liver. *Bmc Genomics, 16*, 981. doi:10.1186/s12864-015-2066-3

Innocenti, F., Cooper, G. M., Stanaway, I. B., Gamazon, E. R., Smith, J. D., Mirkov, S., . . . Brown, C. D. (2011). Identification, replication, and functional fine-mapping of expression quantitative trait loci in primary human liver tissue. *PLoS Genetics, 7*(5), e1002078. doi:10.1371/journal.pgen.1002078

International Multiple Sclerosis Genetics Consortium, Wellcome Trust Case Control Consortium 2, Sawcer, S., Hellenthal, G., Pirinen, M., Spencer, C. C., . . . Compston, A. (2011). Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. *Nature, 476*(7359), 214-219. doi:10.1038/nature10251

Jain, A., & Tuteja, G. (2019). TissueEnrich: Tissue-specific gene enrichment analysis. *Bioinformatics, 35*(11), 1966-1967. doi:10.1093/bioinformatics/bty890

Jin, Y., Birlea, S. A., Fain, P. R., Gowan, K., Riccardi, S. L., Holland, P. J., . . . Spritz, R. A. (2010). Variant of TYR and autoimmunity susceptibility loci in generalized vitiligo. *New England Journal of Medicine, 362*(18), 1686-1697. doi:10.1056/NEJMoa0908547

Jun, G., Flickinger, M., Hetrick, K. N., Romm, J. M., Doheny, K. F., Abecasis, G. R., . . . Kang, H. M. (2012). Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data. *Am J Hum Genet, 91*(5), 839-848. doi:10.1016/j.ajhg.2012.09.004

Kamiya, A., Inoue, Y., & Gonzalez, F. J. (2003). Role of the hepatocyte nuclear factor 4alpha in control of the pregnane X receptor during fetal liver development. *Hepatology, 37*(6), 1375-1384. doi:10.1053/jhep.2003.50212

Khera, A. V., Chaffin, M., Aragam, K. G., Haas, M. E., Roselli, C., Choi, S. H., . . . Kathiresan, S. (2018). Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat Genet, 50*(9), 1219-1224. doi:10.1038/s41588-018-0183-z

Khera, A. V., Chaffin, M., Wade, K. H., Zahid, S., Brancale, J., Xia, R., . . . Kathiresan, S. (2019). Polygenic Prediction of Weight and Obesity Trajectories from Birth to Adulthood. *Cell, 177*(3), 587-596 e589. doi:10.1016/j.cell.2019.03.028

Klein, K., Jungst, C., Mwinyi, J., Stieger, B., Krempler, F., Patsch, W., . . . Kullak-Ublick, G. A. (2010). The human organic anion transporter genes OAT5 and OAT7 are transactivated by hepatocyte nuclear factor-1alpha (HNF-1alpha). *Mol Pharmacol, 78*(6), 1079-1087. doi:10.1124/mol.110.065201

Kleiner, D. E., Brunt, E. M., Van Natta, M., Behling, C., Contos, M. J., Cummings, O. W., . . . Nonalcoholic Steatohepatitis Clinical Research Network. (2005). Design and validation of a histological scoring system for nonalcoholic fatty liver disease. *Hepatology, 41*(6), 1313-1321. doi:10.1002/hep.20701

Knowles, D. A., Davis, J. R., Edgington, H., Raj, A., Fave, M. J., Zhu, X., . . . Battle, A. (2017a). Allele-specific expression reveals interactions between genetic variation and environment. *Nature Methods, 14*(7), 699-702. doi:10.1038/nmeth.4298

Knowles, D. A., Davis, J. R., Edgington, H., Raj, A., Fave, M. J., Zhu, X. W., . . . Battle, A. (2017b). Allele-specific expression reveals interactions between genetic variation and environment. *Nature Methods, 14*(7), 699-+. doi:10.1038/nmeth.4298

Kopp, F., & Mendell, J. T. (2018). Functional Classification and Experimental Dissection of Long Noncoding RNAs. *Cell, 172*(3), 393-407. doi:10.1016/j.cell.2018.01.011

Kozlitina, J., Smagris, E., Stender, S., Nordestgaard, B. G., Zhou, H. H., Tybjaerg-Hansen, A., . . . Cohen, J. C. (2014). Exome-wide association study identifies a TM6SF2 variant that confers susceptibility to nonalcoholic fatty liver disease. *Nat Genet, 46*(4), 352-356. doi:10.1038/ng.2901

Lattouf, R., Younes, R., Lutomski, D., Naaman, N., Godeau, G., Senni, K., & Changotade, S. (2014). Picrosirius red staining: a useful tool to appraise collagen networks in normal and pathological tissues. *Journal of Histochemistry and Cytochemistry, 62*(10), 751-758. doi:10.1369/0022155414545787

Lauridsen, B. K., Stender, S., Kristensen, T. S., Kofoed, K. F., Kober, L., Nordestgaard, B. G., & Tybjaerg-Hansen, A. (2018). Liver fat content, non-alcoholic fatty liver disease, and ischaemic heart disease: Mendelian randomization and meta-analysis of 279 013 individuals. *Eur Heart J, 39*(5), 385-393. doi:10.1093/eurheartj/ehx662

Lawlor, D. A., Harbord, R. M., Sterne, J. A. C., Timpson, N., & Smith, G. D. (2008). Mendelian randomization: Using genes as instruments for making causal inferences in epidemiology. *Statistics in Medicine, 27*(8), 1133-1163. doi:10.1002/sim.3034

Lee, Y. H., Cho, Y., Lee, B. W., Park, C. Y., Lee, D. H., Cha, B. S., & Rhee, E. J. (2019). Nonalcoholic Fatty Liver Disease in Diabetes. Part I: Epidemiology and Diagnosis. *Diabetes Metab J, 43*(1), 31-45. doi:10.4093/dmj.2019.0011

Leung, T. M., Tipoe, G. L., Liong, E. C., Lau, T. Y., Fung, M. L., & Nanji, A. A. (2008). Endothelial nitric oxide synthase is a critical factor in experimental liver fibrosis. *International Journal of Experimental Pathology, 89*(4), 241-250. doi:10.1111/j.1365-2613.2008.00590.x

Li, B., & Dewey, C. N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *Bmc Bioinformatics, 12*, 323. doi:10.1186/1471-2105-12-323

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., . . . Genome Project Data Processing, S. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics, 25*(16), 2078-2079. doi:10.1093/bioinformatics/btp352

Li, J. Z., Huang, Y., Karaman, R., Ivanova, P. T., Brown, H. A., Roddy, T., . . . Hobbs, H. H. (2012). Chronic overexpression of PNPLA3I148M in mouse liver causes hepatic steatosis. *J Clin Invest, 122*(11), 4130-4144. doi:10.1172/JCI65179

Li, M., Santpere, G., Imamura Kawasawa, Y., Evgrafov, O. V., Gulden, F. O., Pochareddy, S., . . . Sestan, N. (2018). Integrative functional genomic analysis of human brain development and neuropsychiatric risks. *Science, 362*(6420). doi:10.1126/science.aat7615

Li, S., Tan, H. Y., Wang, N., Zhang, Z. J., Lao, L., Wong, C. W., & Feng, Y. (2015). The Role of Oxidative Stress and Antioxidants in Liver Diseases. *International Journal of Molecular Sciences, 16*(11), 26087-26124. doi:10.3390/ijms161125942

Li, T. T., Huang, J., Jiang, Y., Zeng, Y., He, F. C., Zhang, M. Q., . . . Zhang, X. G. (2009). Multi-stage analysis of gene expression and transcription regulation in C57/B6 mouse liver development. *Genomics, 93*(3), 235-242. doi:10.1016/j.ygeno.2008.10.006

Li, Y. R., Li, J., Zhao, S. D., Bradfield, J. P., Mentch, F. D., Maggadottir, S. M., . . . Hakonarson, H. (2015). Meta-analysis of shared genetic architecture across ten pediatric autoimmune diseases. *Nature Medicine, 21*, 1018-1027. doi:10.1038/nm.3933 https://www.nature.com/articles/nm.3933#supplementary-information

Lin, M., Liu, C., Liu, Y., Wang, D., Zheng, C., Shi, X., . . . Li, Z. (2019). Fetuin-B Links Nonalcoholic Fatty Liver Disease to Chronic Kidney Disease in Obese Chinese Adults: A Cross-Sectional Study. *Ann Nutr Metab, 74*(4), 287-295. doi:10.1159/000499843

Liu, J. Z., van Sommeren, S., Huang, H., Ng, S. C., Alberts, R., Takahashi, A., . . . Weersma, R. K. (2015). Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nature Genetics, 47*(9), 979-986. doi:10.1038/ng.3359

Liu, W., Anstee, Q. M., Wang, X., Gawrieh, S., Gamazon, E. R., Athinarayanan, S., . . . Chalasani, N. (2016). Transcriptional regulation of PNPLA3 and its impact on susceptibility to nonalcoholic fatty liver Disease (NAFLD) in humans. *Aging (Albany NY), 9*(1), 26-40. doi:10.18632/aging.101067

Liu, Y. L., Reeves, H. L., Burt, A. D., Tiniakos, D., McPherson, S., Leathart, J. B., . . . Anstee, Q. M. (2014). TM6SF2 rs58542926 influences hepatic fibrosis progression in patients with non-alcoholic fatty liver disease. *Nat Commun, 5*, 4309. doi:10.1038/ncomms5309

Liu, Z., Chalasani, N., Lin, J., Gawrieh, S., He, Y., Tseng, Y. J., & Liu, W. (2019). Integrative omics analysis identifies macrophage migration inhibitory factor signaling pathways underlying human hepatic fibrogenesis and fibrosis. *Journal of Bio-X Research, 2*(1), 16-24. doi:10.1097/jbr.0000000000000026

Liu, Z., Zhang, Y., Graham, S., Pique-Regi, R., Dong, X. C., Chen, Y. E., . . . Liu, W. (2019). doi:10.1101/657734

Loh, P. R., Palamara, P. F., & Price, A. L. (2016). Fast and accurate long-range phasing in a UK Biobank cohort. *Nat Genet, 48*(7), 811-816. doi:10.1038/ng.3571

Lonardo, A., Lugari, S., Ballestri, S., Nascimbeni, F., Baldelli, E., & Maurantonio, M. (2019). A round trip from nonalcoholic fatty liver disease to diabetes: molecular targets to the rescue? *Acta Diabetol, 56*(4), 385-396. doi:10.1007/s00592-018-1266-0

Loomis, A. K., Kabadi, S., Preiss, D., Hyde, C., Bonato, V., St Louis, M., . . . Sattar, N. (2016). Body Mass Index and Risk of Nonalcoholic Fatty Liver Disease: Two Electronic Health Record Prospective Studies. *J Clin Endocrinol Metab, 101*(3), 945-952. doi:10.1210/jc.2015-3444

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

Lu, F. B., Hu, E. D., Xu, L. M., Chen, L., Wu, J. L., Li, H., . . . Chen, Y. P. (2018). The relationship between obesity and the severity of non-alcoholic fatty liver disease: systematic review and meta-analysis. *Expert Rev Gastroenterol Hepatol, 12*(5), 491-502. doi:10.1080/17474124.2018.1460202

Luo, Y., Ma, X., Shen, Y., Hao, Y., Hu, Y., Xiao, Y., . . . Jia, W. (2014). Positive relationship between serum low-density lipoprotein cholesterol levels and visceral fat in a Chinese nondiabetic population. *PLoS One, 9*(11), e112715. doi:10.1371/journal.pone.0112715

MacParland, S. A., Liu, J. C., Ma, X. Z., Innes, B. T., Bartczak, A. M., Gage, B. K., . . . McGilvray, I. D. (2018). Single cell RNA sequencing of human liver reveals distinct intrahepatic macrophage populations. *Nature Communications, 9*(1), 4383. doi:10.1038/s41467-018-06318-7

Mahajan, A., Taliun, D., Thurner, M., Robertson, N. R., Torres, J. M., Rayner, N. W., . . . McCarthy, M. I. (2018). Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. *Nat Genet, 50*(11), 1505-1513. doi:10.1038/s41588-018-0241-6

Mangia, A., Gentile, R., Cascavilla, I., Margaglione, M., Villani, M. R., Stella, F., . . . Andriulli, A. (1999). HLA class II favors clearance of HCV infection and progression of the chronic liver damage. *Journal Of Hepatology, 30*(6), 984-989.

Mann, D. A., & Smart, D. E. (2002). Transcriptional regulation of hepatic stellate cell activation. *Gut, 50*(6), 891-896.

Manning, A. K., Hivert, M. F., Scott, R. A., Grimsby, J. L., Bouatia-Naji, N., Chen, H., . . . Langenberg, C. (2012). A genome-wide approach accounting for body mass index identifies genetic variants influencing fasting glycemic traits and insulin resistance. *Nat Genet, 44*(6), 659-669. doi:10.1038/ng.2274

Mantovani, A., Byrne, C. D., Bonora, E., & Targher, G. (2018). Nonalcoholic Fatty Liver Disease and Risk of Incident Type 2 Diabetes: A Meta-analysis. *Diabetes Care, 41*(2), 372-382. doi:10.2337/dc17-1902

Marsillach, J., Camps, J., Ferre, N., Beltran, R., Rull, A., Mackness, B., . . . Joven, J. (2009). Paraoxonase-1 is related to inflammation, fibrosis and PPAR delta in experimental liver disease. *BMC Gastroenterol, 9*, 3. doi:10.1186/1471-230X-9-3

Medvedeva, Y. A., Lennartsson, A., Ehsani, R., Kulakovskiy, I. V., Vorontsov, I. E., Panahandeh, P., . . . Drablos, F. (2015). EpiFactors: a comprehensive database of human epigenetic factors and complexes. *Database (Oxford), 2015*, bav067. doi:10.1093/database/bav067

Meex, R. C. R., & Watt, M. J. (2017). Hepatokines: linking nonalcoholic fatty liver disease and insulin resistance. *Nat Rev Endocrinol, 13*(9), 509-520. doi:10.1038/nrendo.2017.56

Mignone, F., Gissi, C., Liuni, S., & Pesole, G. (2002). Untranslated regions of mRNAs. *Genome Biology, 3*(3). doi:ARTN 0004 10.1186/gb-2002-3-3-reviews0004

Novo, E., Cannito, S., Zamara, E., Valfre di Bonzo, L., Caligiuri, A., Cravanzola, C., . . . Parola, M. (2007). Proangiogenic cytokines as hypoxia-dependent factors stimulating migration of human hepatic stellate cells. *American Journal of Pathology, 170*(6), 1942-1953.

Ober, E. A., & Lemaigre, F. P. (2018). Development of the liver: Insights into organ and tissue morphogenesis. *Journal of Hepatology, 68*(5), 1049-1062. doi:10.1016/j.jhep.2018.01.005

Odamaki, M., Furuya, R., Ohkawa, S., Yoneyama, T., Nishikino, M., Hishida, A., & Kumagai, H. (1999). Altered abdominal fat distribution and its association with the serum lipid profile in non-diabetic haemodialysis patients. *Nephrol Dial Transplant, 14*(10), 2427-2432. doi:10.1093/ndt/14.10.2427

Oguri, T., Ozasa, H., Uemura, T., Bessho, Y., Miyazaki, M., Maeno, K., . . . Ueda, R. (2008). MRP7/ABCC10 expression is a predictive biomarker for the resistance to paclitaxel in non-small cell lung cancer. *Mol Cancer Ther, 7*(5), 1150-1155. doi:10.1158/1535-7163.MCT-07-2088

Orozco, G., Viatte, S., Bowes, J., Martin, P., Wilson, A. G., Morgan, A. W., . . . Eyre, S. (2014). Novel rheumatoid arthritis susceptibility locus at 22q12 identified in an extended UK genome-wide association study. *Arthritis Rheumatol, 66*(1), 24-30. doi:10.1002/art.38196

Osswald, C., Baumgarten, K., Stumpel, F., Gorboulev, V., Akimjanova, M., Knobeloch, K. P., . . . Koepsell, H. (2005). Mice without the regulator gene Rsc1A1 exhibit increased Na+-(D)-Glucose cotransport in small intestine and develop obesity. *Molecular and Cellular Biology, 25*(1), 78-87. doi:10.1128/Mcb.25.1.78-87.2005

Panasevich, M. R., Meers, G. M., Linden, M. A., Booth, F. W., Perfield, J. W., 2nd, Fritsche, K. L., . . . Rector, R. S. (2018). High-fat, high-fructose, high-cholesterol feeding causes severe NASH and cecal microbiota dysbiosis in juvenile Ossabaw swine. *Am J Physiol Endocrinol Metab, 314*(1), E78-E92. doi:10.1152/ajpendo.00015.2017

Pang, Y., Kartsonaki, C., Turnbull, I., Guo, Y., Chen, Y., Clarke, R., . . . Chen, Z. (2019). Adiposity in relation to risks of fatty liver, cirrhosis and liver cancer: a prospective study of 0.5 million Chinese adults. *Sci Rep, 9*(1), 785. doi:10.1038/s41598-018-36460-7

Panousis, N. I., Gutierrez-Arcelus, M., Dermitzakis, E. T., & Lappalainen, T. (2014). Allelic mapping bias in RNA-sequencing is not a major confounder in eQTL studies. *Genome Biology, 15*(9). doi:ARTN 467 10.1186/s13059-014-0467-2

Park, J. H., Gail, M. H., Weinberg, C. R., Carroll, R. J., Chung, C. C., Wang, Z., . . . Chatterjee, N. (2011). Distribution of allele frequencies and effect sizes and their interrelationships for common genetic susceptibility variants. *Proc Natl Acad Sci U S A, 108*(44), 18026-18031. doi:10.1073/pnas.1114759108

Paternoster, L., Tilling, K., & Smith, G. D. (2017). Genetic epidemiology and Mendelian randomization for informing disease therapeutics: Conceptual and methodological challenges. *Plos Genetics, 13*(10). doi:ARTN e1006944 10.1371/journal.pgen.1006944

Patsopoulos, N. A., Bayer Pharma MS Genetics Working Group, Steering Committees of Studies Evaluating IFNβ-1b and a CCR1-Antagonist, ANZgene Consortium, GeneMSA, International Multiple Sclerosis Genetics Consortium, . . . de Bakker, P. I. (2011). Genome-wide meta-analysis identifies novel multiple sclerosis susceptibility loci. *Annals Of Neurology, 70*(6), 897-912. doi:10.1002/ana.22609

Pellicoro, A., Ramachandran, P., Iredale, J. P., & Fallowfield, J. A. (2014). Liver fibrosis and repair: immune regulation of wound healing in a solid organ. *Nature Reviews: Immunology, 14*(3), 181-194. doi:10.1038/nri3623

Persico, M., Masarone, M., Damato, A., Ambrosio, M., Federico, A., Rosato, V., . . . Vecchione, C. (2017). "Non alcoholic fatty liver disease and eNOS dysfunction in humans". *BMC Gastroenterology, 17*(1), 35. doi:10.1186/s12876-017-0592-y

Peter, A., Kovarova, M., Staiger, H., Machann, J., Schick, F., Konigsrainer, A., . . . Stefan, N. (2018). The hepatokines fetuin-A and fetuin-B are upregulated in the state of hepatic steatosis and may differently impact on glucose homeostasis in humans. *Am J Physiol Endocrinol Metab, 314*(3), E266-E273. doi:10.1152/ajpendo.00262.2017

Plagnol, V., Howson, J. M., Smyth, D. J., Walker, N., Hafler, J. P., Wallace, C., . . . Todd, J. A. (2011). Genome-wide association analysis of autoantibody positivity in type 1 diabetes cases. *PLoS Genetics, 7*(8), e1002216. doi:10.1371/journal.pgen.1002216

Poon, H., Quirk, C., DeZiel, C., & Heckerman, D. (2014). Literome: PubMed-scale genomic knowledge base in the cloud. *Bioinformatics, 30*(19), 2840-2842. doi:10.1093/bioinformatics/btu383

Prokopenko, I., Poon, W., Magi, R., Prasad, B. R., Salehi, S. A., Almgren, P., . . . Lyssenko, V. (2014). A central role for GRB10 in regulation of islet function in man. *PLoS Genet, 10*(4), e1004235. doi:10.1371/journal.pgen.1004235

Pruim, R. J., Welch, R. P., Sanna, S., Teslovich, T. M., Chines, P. S., Gliedt, T. P., . . . Willer, C. J. (2010). LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics, 26*(18), 2336-2337. doi:10.1093/bioinformatics/btq419

Pulit, S. L., Stoneman, C., Morris, A. P., Wood, A. R., Glastonbury, C. A., Tyrrell, J., . . . Lindgren, C. M. (2019). Meta-analysis of genome-wide association studies for body fat distribution in 694 649 individuals of European ancestry. *Hum Mol Genet, 28*(1), 166-174. doi:10.1093/hmg/ddy327

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., . . . Sham, P. C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *American Journal Of Human Genetics, 81*(3), 559-575. doi:10.1086/519795

Radmard, A. R., Rahmanian, M. S., Abrishami, A., Yoonessi, A., Kooraki, S., Dadgostar, M., . . . Merat, S. (2016). Assessment of Abdominal Fat Distribution in Non-Alcoholic Fatty Liver Disease by Magnetic Resonance Imaging: a Population-based Study. *Arch Iran Med, 19*(10), 693-699. doi:0161910/AIM.005

Radulovic, K., & Niess, J. H. (2015). CD69 Is the Crucial Regulator of Intestinal Inflammation: A New Target Molecule for IBD Treatment? *Journal of Immunology Research*. doi:Artn 497056 10.1155/2015/497056

Rao, C. V., Asch, A. S., & Yamada, H. Y. (2017). Frequently mutated genes/pathways and genomic instability as prevention targets in liver cancer. *Carcinogenesis, 38*(1), 2-11. doi:10.1093/carcin/bgw118

Richardson, T. G., Harrison, S., Hemani, G., & Davey Smith, G. (2019). An atlas of polygenic risk score associations to highlight putative causal relationships across the human phenome. *Elife, 8*. doi:10.7554/eLife.43657

Romeo, S., Kozlitina, J., Xing, C., Pertsemlidis, A., Cox, D., Pennacchio, L. A., . . . Hobbs, H. H. (2008). Genetic variation in PNPLA3 confers susceptibility to nonalcoholic fatty liver disease. *Nat Genet, 40*(12), 1461-1465. doi:10.1038/ng.257

Roytman, M., Kichaev, G., Gusev, A., & Pasaniuc, B. (2018). Methods for fine-mapping with chromatin and expression data. *PLoS Genet, 14*(2), e1007240. doi:10.1371/journal.pgen.1007240

Ruijtenberg, S., & van den Heuvel, S. (2016). Coordinating cell proliferation and differentiation: Antagonism between cell cycle regulators and cell type-specific gene expression. *Cell Cycle, 15*(2), 196-212. doi:10.1080/15384101.2015.1120925

Sadeghi, M., Pourmoghaddas, Z., Hekmatnia, A., Sanei, H., Tavakoli, B., Tchernof, A., . . . Sarrafzadegan, N. (2013). Abdominal fat distribution and serum lipids in patients with and without coronary heart disease. *Arch Iran Med, 16*(3), 149-153. doi:013163/AIM.006

Saha, S. K., Parachoniak, C. A., Ghanta, K. S., Fitamant, J., Ross, K. N., Najem, M. S., . . . Bardeesy, N. (2014). Mutant IDH inhibits HNF-4alpha to block hepatocyte differentiation and promote biliary cancer. *Nature, 513*(7516), 110-114. doi:10.1038/nature13441

Saxena, R., Hivert, M. F., Langenberg, C., Tanaka, T., Pankow, J. S., Vollenweider, P., . . . Watanabe, R. M. (2010). Genetic variation in GIPR influences the glucose and insulin responses to an oral glucose challenge. *Nat Genet, 42*(2), 142-148. doi:10.1038/ng.521

Schneider, V. A., Graves-Lindsay, T., Howe, K., Bouk, N., Chen, H. C., Kitts, P. A., . . . Church, D. M. (2017). Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Res, 27*(5), 849-864. doi:10.1101/gr.213611.116

Schrem, H., Klempnauer, J., & Borlak, J. (2002). Liver-enriched transcription factors in liver function and development. Part I: The hepatocyte nuclear factor network and liver-specific gene expression. *Pharmacological Reviews, 54*(1), 129-158. doi:DOI 10.1124/pr.54.1.129

Schuppan, D. (1990). Structure of the extracellular matrix in normal and fibrotic liver: collagens and glycoproteins. *Seminars In Liver Disease, 10*(1), 1-10. doi:10.1055/s-2008-1040452

Shabalin, A. A. (2012). Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics, 28*(10), 1353-1358. doi:10.1093/bioinformatics/bts163

Shanker, S., Paulson, A., Edenberg, H. J., Peak, A., Perera, A., Alekseyev, Y. O., . . . Nicolet, C. M. (2015). Evaluation of commercially available RNA amplification kits for RNA sequencing using very low input amounts of total RNA. *J Biomol Tech, 26*(1), 4-18. doi:10.7171/jbt.15-2601-001

Shao, L., Xing, F., Xu, C., Zhang, Q., Che, J., Wang, X., . . . Zhang, Q. (2019). Patterns of genome-wide allele-specific expression in hybrid rice and the implications on the genetic basis of heterosis. *Proc Natl Acad Sci U S A, 116*(12), 5653-5658. doi:10.1073/pnas.1820513116

Sheldon, R. D., Laughlin, M. H., & Rector, R. S. (2014). Reduced hepatic eNOS phosphorylation is associated with NAFLD and type 2 diabetes progression and is prevented by daily exercise in hyperphagic OLETF rats. *J Appl Physiol (1985), 116*(9), 1156-1164. doi:10.1152/japplphysiol.01275.2013

Si-Tayeb, K., Lemaigre, F. P., & Duncan, S. A. (2010). Organogenesis and Development of the Liver. *Developmental Cell, 18*(2), 175-189. doi:10.1016/j.devcel.2010.01.011

Smagris, E., BasuRay, S., Li, J., Huang, Y., Lai, K. M., Gromada, J., . . . Hobbs, H. H. (2015). Pnpla3I148M knockin mice accumulate PNPLA3 on lipid droplets and develop hepatic steatosis. *Hepatology, 61*(1), 108-118. doi:10.1002/hep.27242

Speliotes, E. K., Yerges-Armstrong, L. M., Wu, J., Hernaez, R., Kim, L. J., Palmer, C. D., . . . Consortium, G. (2011). Genome-wide association analysis identifies variants associated with nonalcoholic fatty liver disease that have distinct effects on metabolic traits. *PLoS Genet, 7*(3), e1001324. doi:10.1371/journal.pgen.1001324

Srivastava, A., Shukla, V., Tiwari, D., Gupta, J., Kumar, S., & Kumar, A. (2018). Targeted therapy of chronic liver diseases with the inhibitors of angiogenesis. *Biomedicine & Pharmacotherapy, 105*, 256-266. doi:10.1016/j.biopha.2018.05.102

Staiger, D., & Stock, J. H. (1997). Instrumental variables regression with weak instruments. *Econometrica, 65*(3), 557-586. doi:Doi 10.2307/2171753

Stender, S., Kozlitina, J., Nordestgaard, B. G., Tybjaerg-Hansen, A., Hobbs, H. H., & Cohen, J. C. (2017). Adiposity amplifies the genetic risk of fatty liver disease conferred by multiple loci. *Nat Genet, 49*(6), 842-847. doi:10.1038/ng.3855

Strawbridge, R. J., Dupuis, J., Prokopenko, I., Barker, A., Ahlqvist, E., Rybin, D., . . . Florez, J. C. (2011). Genome-wide association identifies nine common variants associated with fasting proinsulin levels and provides new insights into the pathophysiology of type 2 diabetes. *Diabetes, 60*(10), 2624-2634. doi:10.2337/db11-0415

Sung, K. C., Wild, S. H., & Byrne, C. D. (2013). Resolution of fatty liver and risk of incident diabetes. *J Clin Endocrinol Metab, 98*(9), 3637-3643. doi:10.1210/jc.2013-1519

Sveinbjornsson, G., Albrechtsen, A., Zink, F., Gudjonsson, S. A., Oddson, A., Másson, G., . . . Stefansson, K. (2016). Weighting sequence variants based on their annotation increases power of whole-genome association studies. *Nature Genetics, 48*, 314-317. doi:10.1038/ng.3507 https://www.nature.com/articles/ng.3507#supplementary-information

Sweet, D. H. (2005). Organic anion transporter (Slc22a) family members as mediators of toxicity. *Toxicol Appl Pharmacol, 204*(3), 198-215. doi:10.1016/j.taap.2004.10.016

Tao, R., Xiong, X., DePinho, R. A., Deng, C. X., & Dong, X. C. (2013). Hepatic SREBP-2 and cholesterol biosynthesis are regulated by FoxO3 and Sirt6. *J Lipid Res, 54*(10), 2745-2753. doi:10.1194/jlr.M039339

Tehranchi, A., Hie, B., Dacre, M., Kaplow, I., Pettie, K., Combs, P., & Fraser, H. B. (2019). Fine-mapping cis-regulatory variants in diverse human populations. *Elife, 8*. doi:10.7554/eLife.39595

Thomas, H. (2018). Liver: Delineating the role of angiogenesis in liver fibrosis. *Nat Rev Gastroenterol Hepatol, 15*(1), 6. doi:10.1038/nrgastro.2017.168

Tian, L., Khan, A., Ning, Z., Yuan, K., Zhang, C., Lou, H., . . . Xu, S. (2018). Genome-wide comparison of allele-specific gene expression between African and European populations. *Hum Mol Genet, 27*(6), 1067-1077. doi:10.1093/hmg/ddy027

Ullah, A. Z. D., Oscanoa, J., Wang, J., Nagano, A., Lemoine, N. R., & Chelala, C. (2018). SNPnexus: assessing the functional relevance of genetic variation to facilitate the promise of precision medicine. *Nucleic Acids Research, 46*(W1), W109-W113. doi:10.1093/nar/gky399

van de Geijn, B., McVicker, G., Gila, Y., & Pritchard, J. K. (2015). WASP: allele-specific software for robust molecular quantitative trait locus discovery. *Nature Methods, 12*(11), 1061-1063. doi:10.1038/Nmeth.3582

van de Geijn, B., McVicker, G., Gilad, Y., & Pritchard, J. K. (2015). WASP: allele-specific software for robust molecular quantitative trait locus discovery. *Nature Methods, 12*(11), 1061-1063. doi:10.1038/nmeth.3582

Van Herck, M. A., Vonghia, L., & Francque, S. M. (2017). Animal Models of Nonalcoholic Fatty Liver Disease-A Starter's Guide. *Nutrients, 9*(10). doi:10.3390/nu9101072

Verbanck, M., Chen, C. Y., Neale, B., & Do, R. (2018). Detection of widespread horizontal pleiotropy in causal relationships inferred from Mendelian randomization between complex traits and diseases. *Nat Genet, 50*(5), 693-698. doi:10.1038/s41588-018-0099-7

Visscher, P. M., Brown, M. A., McCarthy, M. I., & Yang, J. (2012). Five years of GWAS discovery. *American Journal Of Human Genetics, 90*(1), 7-24. doi:10.1016/j.ajhg.2011.11.029

Voight, B. F., Peloso, G. M., Orho-Melander, M., Frikke-Schmidt, R., Barbalic, M., Jensen, M. K., . . . Kathiresan, S. (2012). Plasma HDL cholesterol and risk of myocardial infarction: a mendelian randomisation study. *Lancet, 380*(9841), 572-580. doi:10.1016/S0140-6736(12)60312-2

Wald, A. (1940). The fitting of straight lines if both variables are subject to error. *Annals of Mathematical Statistics, 11*, 284-300. doi:DOI 10.1214/aoms/1177731868

Wang, D., Liu, S., Warrell, J., Won, H., Shi, X., Navarro, F. C. P., . . . Gerstein, M. B. (2018). Comprehensive functional genomic resource and integrative model for the human brain. *Science, 362*(6420). doi:10.1126/science.aat8464

Wang, L., Athinarayanan, S., Jiang, G., Chalasani, N., Zhang, M., & Liu, W. (2015). Fatty acid desaturase 1 gene polymorphisms control human hepatic lipid composition. *Hepatology, 61*(1), 119-128. doi:10.1002/hep.27373

Wang, N., Chen, C., Zhao, L., Chen, Y., Han, B., Xia, F., . . . Lu, Y. (2018). Vitamin D and Nonalcoholic Fatty Liver Disease: Bi-directional Mendelian Randomization Analysis. *EBioMedicine, 28*, 187-193. doi:10.1016/j.ebiom.2017.12.027

Wang, T., Ma, X., Tang, T., Jin, L., Peng, D., Zhang, R., . . . Jia, W. (2016). Overall and central obesity with insulin sensitivity and secretion in a Han Chinese population: a Mendelian randomization analysis. *Int J Obes (Lond), 40*(11), 1736-1741. doi:10.1038/ijo.2016.155

Wang, T., Zhang, R., Ma, X., Wang, S., He, Z., Huang, Y., . . . Jia, W. (2018). Causal Association of Overall Obesity and Abdominal Obesity with Type 2 Diabetes: A Mendelian Randomization Analysis. *Obesity (Silver Spring), 26*(5), 934-942. doi:10.1002/oby.22167

Wang, X., Liu, Z., Wang, K., Wang, Z., Sun, X., Zhong, L., . . . Liu, W. (2016). Additive Effects of the Risk Alleles of PNPLA3 and TM6SF2 on Non-alcoholic Fatty Liver Disease (NAFLD) in a Chinese Population. *Front Genet, 7*, 140. doi:10.3389/fgene.2016.00140

Wang, Y., Kory, N., BasuRay, S., Cohen, J. C., & Hobbs, H. H. (2019). PNPLA3, CGI-58, and Inhibition of Hepatic Triglyceride Hydrolysis in Mice. *Hepatology*. doi:10.1002/hep.30583

Wei, R., Yang, F., Urban, T. J., Li, L., Chalasani, N., Flockhart, D. A., & Liu, W. (2012). Impact of the Interaction between 3'-UTR SNPs and microRNA on the Expression of Human Xenobiotic Metabolism Enzyme and Transporter Genes. *Front Genet, 3*, 248. doi:10.3389/fgene.2012.00248

Welsh, P., Polisecki, E., Robertson, M., Jahn, S., Buckley, B. M., de Craen, A. J. M., . . . Sattar, N. (2010). Unraveling the Directional Link between Adiposity and Inflammation: A Bidirectional Mendelian Randomization Approach. *Journal of Clinical Endocrinology & Metabolism, 95*(1), 93-99. doi:10.1210/jc.2009-1064

Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., . . . Parkinson, H. (2014). The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Research, 42*(Database issue), D1001-1006. doi:10.1093/nar/gkt1229

Wheeler, E., Leong, A., Liu, C. T., Hivert, M. F., Strawbridge, R. J., Podmore, C., . . . Meigs, J. B. (2017). Impact of common genetic determinants of Hemoglobin A1c on type 2 diabetes risk and diagnosis in ancestrally diverse populations: A transethnic genome-wide meta-analysis. *PLoS Med, 14*(9), e1002383. doi:10.1371/journal.pmed.1002383

Willer, C. J., Schmidt, E. M., Sengupta, S., Peloso, G. M., Gustafsson, S., Kanoni, S., . . . Global Lipids Genetics, C. (2013). Discovery and refinement of loci associated with lipid levels. *Nat Genet, 45*(11), 1274-1283. doi:10.1038/ng.2797

Wilson, A. L., Schrecengost, R. S., Guerrero, M. S., Thomas, K. S., & Bouton, A. H. (2013). Breast cancer antiestrogen resistance 3 (BCAR3) promotes cell motility by regulating actin cytoskeletal and adhesion remodeling in invasive breast cancer cells. *PloS One, 8*(6), e65678. doi:10.1371/journal.pone.0065678

Woo Seo, D., Yeop You, S., Chung, W. J., Cho, D. H., Kim, J. S., & Su Oh, J. (2015). Zwint-1 is required for spindle assembly checkpoint function and kinetochore-microtubule attachment during oocyte meiosis. *Scientific Reports, 5*, 15431. doi:10.1038/srep15431

Xu, G. F., Liu, B. Y., Sun, Y. B., Du, Y., Snetselaar, L. G., Hu, F. B., & Bao, W. (2018). Prevalence of diagnosed type 1 and type 2 diabetes among US adults in 2016 and 2017: population based study. *Bmj-British Medical Journal, 362*. doi:ARTN k1497 10.1136/bmj.k1497

Yamaoka, K., Nouchi, T., Marumo, F., & Sato, C. (1993). Alpha-smooth-muscle actin expression in normal and fibrotic human livers. *Digestive Diseases and Sciences, 38*(8), 1473-1479.

Yang, X., Chen, Q., Sun, L., Zhang, H., Yao, L., Cui, X., . . . Chang, Y. (2017). KLF10 transcription factor regulates hepatic glucose metabolism in mice. *Diabetologia, 60*(12), 2443-2452. doi:10.1007/s00125-017-4412-2

Yavorska, O. O., & Burgess, S. (2017). MendelianRandomization: an R package for performing Mendelian randomization analyses using summarized data. *Int J Epidemiol, 46*(6), 1734-1739. doi:10.1093/ije/dyx034

Yoshiji, H., Kuriyama, S., Yoshii, J., Ikenaka, Y., Noguchi, R., Hicklin, D. J., . . . Fukui, H. (2003). Vascular endothelial growth factor and receptor interaction is a prerequisite for murine hepatic fibrogenesis. *Gut, 52*(9), 1347-1354.

Younossi, Z., Anstee, Q. M., Marietti, M., Hardy, T., Henry, L., Eslam, M., . . . Bugianesi, E. (2018). Global burden of NAFLD and NASH: trends, predictions, risk factors and prevention. *Nat Rev Gastroenterol Hepatol, 15*(1), 11-20. doi:10.1038/nrgastro.2017.109

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

Yu, Y., Ping, J., Chen, H., Jiao, L. X., Zheng, S. Y., Han, Z. G., . . . Huang, J. A. (2010). A comparative analysis of liver transcriptome suggests divergent liver function among human, mouse and rat. *Genomics, 96*(5), 281-289. doi:10.1016/j.ygeno.2010.08.003

Yucesoy, B., Kaufman, K. M., Lummus, Z. L., Weirauch, M. T., Zhang, G., Cartier, A., . . . Bernstein, D. I. (2015). Genome-Wide Association Study Identifies Novel Loci Associated With Diisocyanate-Induced Occupational Asthma. *Toxicological Sciences, 146*(1), 192-201. doi:10.1093/toxsci/kfv084

Zanger, U. M., & Schwab, M. (2013). Cytochrome P450 enzymes in drug metabolism: regulation of gene expression, enzyme activities, and impact of genetic variation. *Pharmacol Ther, 138*(1), 103-141. doi:10.1016/j.pharmthera.2012.12.007

Zhang, X., Wan, Y., Zhang, S., Lu, L., Chen, Z., Liu, H., . . . Cai, W. (2015). Nonalcoholic fatty liver disease prevalence in urban school-aged children and adolescents from the Yangtze River delta region: a cross-sectional study. *Asia Pacific Journal of Clinical Nutrition, 24*(2), 281-288. doi:10.6133/apjcn.2015.24.2.13

Zhao, H., Huang, Y., Shi, J., Dai, Y., Wu, L., & Zhou, H. (2018). ABCC10 Plays a Significant Role in the Transport of Gefitinib and Contributes to Acquired Resistance to Gefitinib in NSCLC. *Front Pharmacol, 9*, 1312. doi:10.3389/fphar.2018.01312

Zhou, Q., Liu, M. W., Xia, X., Gong, T. Q., Feng, J. W., Liu, W. L., . . . Qin, J. (2017). A mouse tissue transcription factor atlas. *Nature Communications, 8*. doi:ARTN 15089 10.1038/ncomms15089

Zhou, W., Nielsen, J. B., Fritsche, L. G., Dey, R., Gabrielsen, M. E., Wolford, B. N., . . . Lee, S. (2018). Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nat Genet, 50*(9), 1335-1341. doi:10.1038/s41588-018-0184-y

Zhou, Y., Zhu, G., Charlesworth, J. C., Simpson, S., Jr., Rubicz, R., Goring, H. H., . . . ANZgene consortium. (2016). Genetic loci for Epstein-Barr virus nuclear antigen-1 are associated with risk of multiple sclerosis. *Multiple Sclerosis, 22*(13), 1655-1664. doi:10.1177/1352458515626598

Zhu, Z., Zhang, F., Hu, H., Bakshi, A., Robinson, M. R., Powell, J. E., . . . Yang, J. (2016). Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat Genet, 48*(5), 481-487. doi:10.1038/ng.3538