UC San Diego

UC San Diego Electronic Theses and Dissertations

Title

Systems Biology of Liver Regeneration and Pathologies

Permalink

https://escholarship.org/uc/item/05d214b4

Author

Min, Jun SungJun

Publication Date

2015

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

UNIVERSITY OF CALIFORNIA, SAN DIEGO

Systems Biology of Liver Regeneration and Pathologies

A dissertation submitted in partial satisfaction of the requirements for the degree Doctor of Philosophy

in

Bioengineering

by

Jun SungJun Min

Committee in charge:

Professor Shankar Subramaniam, Chair
Professor Pedro Cabrales
Professor Daniel Tartakovsky
Professor Shyni Varghese
Professor Yingxiao Wang

2015

Copyright

Jun SungJun Min, 2015

All rights reserved.

The Dissertation of Jun SungJun Min is approved, and it is acceptable

in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____

_____
Chair

University of California, San Diego

2015

DEDICATION

To my friends and family

With their love and support

TABLE OF CONTENTS

Signature Page iii

Dedication iv

Table of Contents v

List of Figures vii

List of Tables viii

List of Supplemental Figures ix

List of Supplemental Tables x

Acknowledgements xi

Vita xiv

Abstract of the Dissertation xv

Introduction 1

CHAPTER 1: SYSTEMS BIOLOGY OF LIVER REGENERATION WITH TRANSCRIPTOMIC AND METABOLOMIC ANALYSES 6

- Abstract 6
- Introduction 7
- Methods 10
- Results 16
- Discussion 30

Supplementary Materials 39

Acknowledgements 63

CHAPTER 2: TRANSCRIPTOMIC AND INTEGRATIVE ANALYSES OF BILIARY ATRESIA 64

Introduction 64

Methods 66

Results 70

Discussion 88

Supplementary Materials 94

Acknowledgements 116

CHAPTER 3: TARGET SEQUENCING, EXOME SEQUENCING, AND NETWORK ANALYSES OF BILIARY ATRESIA 117

Introduction 117

Methods 121

Results 127

Discussion 139

Supplementary Materials 146

Acknowledgements 166

Conclusion 167

References 170

LIST OF FIGURES

Figure 1.1 Venn diagrams of differentially regulated genes after PHx under the p-value cutoff of 0.05 and the FDR cutoff of 0.1 and over-represented biological functions and pathways during the priming phase 18
Figure 1.2 Temporal network analysis in Cytoscape 21
Figure 1.3 Acute phase genes after partial hepatectomy 22
Figure 1.4 Temporal gene expression of the acute phase proteins and their correlation with cytokine profiles 23
Figure 1.5 Transcriptomic and metabolic profiles for cholesterol metabolism 26
Figure 1.6 Metabolic fold changes at 3 hours and the heatmap of lipid metabolic genes 29
Figure 1.7 Correlation heatmap of gene-metabolite pairs in the sterol pathway 30
Figure 1.8 Proposed mechanism of the priming phase of complement-induced liver regeneration 37

Figure 2.1 Workflows for the transcriptomic and the integrative analyses 70
Figure 2.2 Distribution of RNAseq read counts in BA 71
Figure 2.3 RNAseq dispersion plot 72
Figure 2.4 Differentially regulated genes in enriched biological categories 78
Figure 2.5 Enriched KEGG pathways 80
Figure 2.6 Sequence features of significant GWAS variants 86

Figure 3.1 Novel systems biology approach for the reconstruction of the BA network 120
Figure 3.2 Linkage disequilibrium analysis of exon #7 of MAN1A2 130
Figure 3.3 Whole exome network 134
Figure 3.4 Proposed biliary atresia network 136
Figure 3.5 Common biological functions in the proposed biliary atresia network 137

LIST OF TABLES

Table 2.1	List of differentially regulated genes	73
Table 2.2	Enriched Gene Ontology terms	77
Table 2.3	Enriched KEGG pathways	80
Table 2.4	Differentially regulated genes in the complement and coagulation cascade	81
Table 2.5	Differential regulation of exons in MAN1A2	82
Table 2.6	Differential alternate splicing in MAN1A2	83
Table 2.7	Significant pairs of differentially regulated genes and BA-associated SNPs	84
Table 2.8	eQTL results from the second integrative analysis	87
Table 2.9	Functional prediction of unassociated SNPS	88
Table 3.1	Recent biliary atresia GWAS studies	118
Table 3.2	Average alignment metrics for target and whole exome sequencing	127
Table 3.3	Novel SNPs from target sequencing	128
Table 3.4	Novel missense SNPs from whole exome	131
Table 3.5	Top 5 common transcription factors from the BA network	138

LIST OF SUPPLEMENTAL FIGURES

Figure S1.1 Correlation plot between RNAseq and qPCR 43
Figure S1.2 Metabolic fold changes at 3 hours 43
Figure S1.3 Metabolic profiles of uric acid and 2,3-diphosphoglycerate..... 62

Figure S3.1 The first cluster of the MCODE algorithm on the whole exome network..... 154
Figure S3.2 Proposed mechanism for the role of ARF6 in BA..... 165

LIST OF SUPPLEMENTAL TABLES

Table S1.1 RNAseq pooling plan 39
Table S1.2 Alignment results for RNAseq analysis 39
Table S1.3 Alignment performance of OSA and TOPHAT 40
Table S1.4 Complete calculations for differentially regulated genes 40
Table S1.5 List of metabolites measured using mass spectrometry 41
Table S1.6 List of differentially regulated genes under the FDR cutoff of 0.1 at 0 hours 44
Table S1.7 List of differentially regulated genes under the FDR cutoff of 0.1 at 0.5 hours 46
Table S1.8 List of differentially regulated genes under the FDR cutoff of 0.1 at 1 hour 49
Table S1.9 List of differentially regulated genes under the FDR cutoff of 0.1 at 3 hours 52
Table S1.10 Metabolite measurements of phospholipids for KO conditions 58
Table S1.11 Metabolite measurements of phospholipids for WT conditions 59
Table S1.12 Metabolite measurements of cholesterol esters for KO conditions 60
Table S1.13 Metabolite measurements of cholesterol esters for WT conditions 61

Table S2.1 Enriched Panther signaling pathways 94
Table S2.2 Enriched BIOCARTA signaling pathways 94
Table S2.3 Enriched GO:BP 94
Table S2.4 Enriched GO:MF 95
Table S2.5 Significant SNPs from the TDT analysis 95
Table S2.6 RTqPCR primer sequences for MAN1A2 exons 101
Table S2.7 RTqPCR results 101
Table S2.8 List of differentially regulated genes from the RNAseq data .. 102

Table S3.1 Numbered SNPs in the proposed biliary atresia network 146
Table S3.2 List of BA patient samples used in each data sources 148
Table S3.3 Predicted functions of novel variants 150
Table S3.4 Common transcription factors from the BA network 151
Table S3.5 List of genes mapped from the common SNPs in the whole exome data and the internal list of GWAS 154

ACKNOWLEDGEMENTS

I would like to thank my advisor, Professor Shankar Subramaniam, who has supported me with invaluable insight and guidance throughout my Ph.D. years. I would like to also acknowledge the rest of my committee members, Professors Daniel Tartakovsky, Pedro Cabrales, Yingxiao Wang, and Shyni Varghese, who have been generous with their time and advice.

I am grateful for help I received from friends, family, and various members of Subramaniam lab at UCSD. I would like to acknowledge my friends, Dinh Diep (Zhang lab) and Joseph Sugie (Sung lab), for fruitful scientific discussions and feedback on my research. I would like to also acknowledge my brother, Peter Min (Engler lab), who has helped with the writing of this dissertation. Last but not least, I would like to acknowledge the members of Subramaniam lab, especially Dr. Shakti Gupta and Dr. Mano Maurya, who have provided positive mentorship since my undergraduate years in the lab.

Chapter 1, in full, is a re-editing of materials currently being prepared for submission for publication in Jun Min, Robert DeAngelis, Charles Evans, Mano Maurya, Shakti Gupta, Charles Burant, John Lambris, and Shankar Subramaniam. Systems analysis of complement-induced priming phase of liver regeneration, **in preparation**. The dissertation author is the primary investigator and author of this paper. The dissertation author is responsible for

designing the overall experiments, performing all analytic methods, and writing of the paper.

Chapter 2, in part, is a re-editing of materials published in Mylarappa Ningappa, Jun Min, Brandon W. Higgs, Chethan Ashokkumar, Sarangarajan Ranganathan, and Rakesh Sindhi, Genome-wide association studies in biliary atresia. Wiley interdisciplinary reviews. Systems biology and medicine. 2015.7: 267-273. The dissertation author contributed to the writing of the review paper. Chapter 2, in part, is also a re-editing of materials currently being prepared for submission for publication in Juhoon So, Ningappa Mylarappa, Jun Min, Brandon Higgs, Qing Sun, Hakon Hakonarson, Shankar Subramaniam, Donghun Shin and Rakesh Sindhi. The role of MAN1A2 in biliary atresia, **in preparation**. The dissertation author is responsible for analyzing data and performing systems biology methods.

Chapter 3, in part, is a re-editing of materials currently being prepared for submission for publication in Ningappa Mylarappa*, Jun Min*, Brandon Higgs, Qing Sun, Hakon Hakonarson, Donghun Shin, Shankar Subramaniam, and Rakesh Sindhi. Systems analysis of biliary atresia through integration of high-throughput biological data, **in preparation**. The dissertation author is the co-first author of this paper and is responsible for designing and performing all analytic methods and writing of the paper. Chapter 3, in part, is also a re-editing of materials published in Mylarappa Ningappa, Juhoon So, Joseph Glessner, Chethan Ashokkumar, Sarangarajan Ranganathan, Jun Min, Brandon W. Higgs, Qing Sun, Kimberly Haberman, Lori Schmitt, Silvia

Vilarinho, Pramod K. Mistry, Gerard Vockley, Anil Dhawan, George K. Gittes, Hakon Hakonarson, Ronald Jaffe, Shankar Subramaniam, Donghun Shin, and Rakesh Sindhi. The role of ARF6 in biliary atresia. PLOS ONE. 2015. The dissertation author was responsible for analyzing data and performing systems biology methods to derive a comprehensive mechanism, which is depicted in a diagram.

VITA

2010 Bachelor of Science, University of California, San Diego

2015 Doctor of Philosophy, University of California, San Diego

RECENT PUBLICATIONS

J. Min, R. DeAngelis, C. Evans, M. Maurya, S. Gupta, C. Burant, J. Lambris, and S. Subramaniam, “Systems analysis of complement-induced priming phase of liver regeneration,” **in preparation**

J. So*, N. Mylarappa*, **J. Min**, B. Higgs, Q. Sun, H. Hakonarson, S. Subramaniam, D. Shin and R. Sindhi, “The role of MAN1A2 in biliary atresia,” **in preparation**

N. Mylarappa*, **J. Min***, B. Higgs, Q. Sun, H. Hakonarson, D. Shin, S. Subramaniam, and R. Sindhi, “Systems analysis of biliary atresia through integration of high-throughput biological data,” **in preparation**

N. Mylarappa, **J. Min**, B. W. Higgs, C. Ashokkumar, S. Ranganathan, and R. Sindhi. 2015. Genome-wide association studies in biliary atresia. Wiley interdisciplinary reviews. Systems biology and medicine 7: 267-273.

N. Mylarappa, J. So, J. Glessner, C. Ashokkumar, S. Ranganathan, **J. Min**, B. W. Higgs, Q. Sun, K. Haberman, L. Schmitt, S. Vilarinho, P. K. Mistry, G. Vockley, A. Dhawan, G. K. Gittes, H. Hakonarson, R. Jaffe, S. Subramaniam, D. Shin, and R. Sindhi. 2015. The Role of ARF6 in Biliary Atresia. PloS one 10

ABSTRACT OF THE DISSERTATION

Systems Biology of Liver Regeneration and Pathologies

by

Jun SungJun Min

Doctor of Philosophy in Bioengineering

University of California, San Diego, 2015

Professor Shankar Subramaniam, Chair

The liver is the largest internal organ, accounting for approximately 2-3% of the average body weight, and is involved in a variety of important functions such as digestion, metabolism, detoxification, production of vital proteins, coagulation, and immune response. Understanding the complex anatomy and physiology of the liver has been a long-standing challenge for

scientists and physicians who struggle to identify underlying causes for many types of liver disease. There are over 100 types of liver disease with different risk factors that can lead to cirrhosis and liver failure. To better understand the physiology of the liver and the pathogenesis of various types of liver disease, I have applied novel systems biology approaches to investigate the mechanisms of liver regeneration and pathologies.

In Chapter 1, I investigated the priming phase of liver regeneration in the complement-knockout mice. The complement system, part of the innate immune system, has been recently shown to successfully promote the early phase, or priming phase, of liver regeneration, during which complex regulation of signaling pathways and other molecular events occur. To better understand the role of the complement system in this complex biological process, I analyzed transcriptomic and metabolomic measurements during the several time points of the priming phase of liver regeneration to identify novel biomarkers and relevant biological pathways. I also supplemented the results with protein-protein interaction network, correlation analysis, and literature knowledge to derive a comprehensive mechanism.

In Chapters 2 and 3, I investigated the complex pathogenesis of biliary atresia, a rare disease of the liver and the bile ducts. Biliary atresia has unknown etiologies and multiple disease forms based on the heterogeneous phenotypes that are difficult to diagnose. To better elucidate the pathogenesis of this disease, I analyzed different high-throughput molecular data such as genome-wide association study, mRNA sequencing, target genome

sequencing, and whole exome sequencing. I also integrated novel biomarkers from these datasets with a protein-protein interaction network to reconstruct a disease specific network that is both highly interpretable and enriched in important biological functions, including inflammation, immunity, fibrosis, and development.

INTRODUCTION

Within the interdisciplinary field of bioengineering, systems biology has become more popular due to the technological advancements in next-generation sequencing that allows high-throughput quantification of DNA, RNA, miRNA, exome, methylome and epigenome (1). Systems biology is the study of systems of biological components that cannot be explained by the simple sum of their components' functions. It is used to approach complex, non-Mendelian disease with effective quantitative methods that can unveil the intricate web of interactions (2). For example, we can analyze multiple high-throughput molecular data such as DNA and RNA to not only identify novel significant biomarkers but also to integrate the results in a network with known protein-protein interactions (3). Furthermore, we can identify over- or under-represented biological functions and pathways based on the group of functionally similar genes or genomic variants that are differentially regulated in the disease group compared to the normal group (4). While most studies still involve in-depth investigations of individual parts such as particular genes or proteins, more studies are beginning to employ a systems biology approach to significantly improve our understanding of the comprehensive mechanism of a disease (5, 6).

The rising application of systems biology stems from the increasing popularity of next-generation sequencing (NGS) datasets. NGS technology allows massively parallel sequencing, during which millions of DNA fragments

from multiple samples are sequenced for high-throughput analysis (7). NGS analysis typically involves aligning millions of short reads to a genome or a transcriptome and identifying variants or genes that are differentially regulated in a disease group (8). The analysis can then be followed by systems biology methods to provide further insights. While genome-wide association study (GWAS) and microarray studies still provide comparable high-throughput quantification of biological data, NGS experiments remove the design bias of an array-based system and can generate more detailed data, especially on the regions that are not targetable by probes (9). Recently, the cost and the availability of NGS platforms have been steadily improving (1, 10), which allows more opportunities to employ systems biology approaches to analyze large-scale data.

One popular method in systems biology is enrichment analysis that analyzes the common functions from the groups of genes and their interactions. Enrichment analysis identifies the classes of genes that are over- or under-represented in a large set of differentially regulated genes; the classes can be known biological functions, pathways or processes (4). For example, enrichment analysis can identify inflammation as a key mechanism of a disease if the number of differentially regulated inflammatory genes exceeds a certain statistical threshold.

Network analysis using protein-protein interactions is another common method in systems biology (3). Proteins are the main molecules that execute biological responses by interacting with other proteins in signaling pathways.

These interactions can either be experimentally tested or computationally predicted to derive a large set of protein-protein interactions, which can then be assembled into a protein-protein interaction network. In this network, proteins are represented by nodes and their interactions represented by edges. This large protein-protein interaction network can be used to identify other genes that may interact with differentially regulated genes at the protein level. This allows us to refine the list of potential biomarkers and biological pathways of a disease.

In this dissertation, I used these common systems biology and novel integrative methods to analyze liver regeneration and pathologies. Liver is the largest internal organ in the human body and plays a central role in metabolic homeostasis and compound detoxification (11). Liver also produces enzymes, cofactors, and serum proteins such as albumin, complement components, and acute phase proteins (12). Because of these essential functions, injuries or infection to the liver resulting from cell loss, autoimmune diseases, toxins, or surgical resection of the liver tissues can lead to severe symptoms and often life-threatening conditions. For example, an infection in the liver can cause inflammation that can progress to fibrosis or cirrhosis (13). Hepatitis B can also lead to chronic liver failure if not treated with medications or liver transplantation (14). Fortunately, the liver has the ability to regenerate especially following toxic injury or surgical resection of the liver with the help of the complement system and diverse cell signaling and metabolic pathways (15).

Although much research has been done on delineating which genes, proteins, or metabolites could contribute to the pathogenesis of various liver-related conditions, the detailed mechanisms have not been fully established. For example, the priming phase of liver regeneration is mediated by well-known transcription factors and immediate-early genes but the relationship between metabolism, acute inflammation, the complement system, and the regulation of gene expression has not been well-studied (16). In addition, the detailed etiologies and pathogenesis of biliary atresia remain unknown even though different theories have been proposed with a few BA marker genes and significant genomic variants that were discovered (17).

In Chapter 1, I investigated the priming phase of liver regeneration and its relationship to the complement system. Due to the complex time-sensitive nature of signaling events that can initiate and progress the priming phase of liver regeneration (16), I obtained the transcriptomic and metabolic data from the liver tissues of mice at various time points during the first 3 hours. I performed differential and enrichment analyses from the transcriptomic data to identify important groups of genes and over-represented biological functions and pathways. I also performed network analysis through the integration of the transcriptomic data with the protein-protein interaction network as well as correlation analysis between multiple sets of data. Based on the results of several analyses, I proposed a comprehensive mechanism for the complement-induced priming phase of liver regeneration.

In Chapter 2, I performed transcriptomic and integrative analyses using RNAseq and GWAS data from the liver tissue and blood samples of young biliary atresia (BA) patients to investigate the pathogenesis of BA, a rare liver disease that is characterized by the disruption of the biliary system (17). For the RNAseq analysis, I performed differential and enrichment analyses, similar to the methods in Chapter 1. For the integrative analysis, I devised integrative methods to analyze both RNAseq and GWAS data to identify the pairs of differentially regulated genes and their nearby BA-associated variants. I also analyzed the predicted functions of significant genomic variants that did not lead to different transcriptional regulation of their nearby genes.

In Chapter 3, I first performed target and whole exome sequencing analyses to complement the results from Chapter 2 to derive a comprehensive list of BA-related genes and their associated single nucleotide polymorphisms (SNPs). To ensure a certain degree of statistical and biological confidence, I chose only the highly-common variants among the BA patients. I then used many of the relevant variants and their target genes discovered from multiple analyses to reconstruct a BA network by using protein-protein interactions. This BA network conveyed the complex pathogenesis through the interplay of significant biomarkers and known biological functions for different forms of BA.

CHAPTER 1: SYSTEMS BIOLOGY OF LIVER REGENERATION WITH TRANSCRIPTOMIC AND METABOLOMIC ANALYSES

Abstract

Liver regeneration is a well-orchestrated and unique process in the liver that allows mature hepatocytes to re-enter the cell cycle to proliferate to replace lost or damaged cells. This process is often impaired in fatty or diseased livers, leading to cirrhosis and other deleterious phenotypes. Prior research has established the role of the complement system and its effector proteins in the progression of liver regeneration; however, a detailed mechanistic understanding of involvement of complement in regeneration is yet to be established. In this Chapter, I have examined the role of the complement system during the priming phase of liver regeneration through transcriptomic and metabolomic analyses. Based on the analyses, I showed that the complement system activates c-fos and promotes tumor necrosis factor α signaling pathway, which then activates acute phase genes such as serum amyloid proteins and orosomucoids. The complement system also regulates efflux and metabolism of cholesterol, an important metabolite for cell cycle and proliferation.

Introduction

Liver is the second largest organ in the body with the unique ability to regenerate itself from as little as 25% of its original mass (18, 19). This regenerative property is essential in supporting the liver's ongoing central role in many biological processes such as complex homeostasis and compound detoxification. However, liver regeneration is impaired in diseased, aged, or fatty livers (20-24). Therefore, a detailed understanding of the mechanisms underlying liver regeneration is necessary for the development of therapies to enhance or restore the regenerative property of the diseased livers. Despite the continuous efforts to unravel the mechanisms of liver regeneration over the past decades, the comprehensive mechanism still has not been fully mapped (25).

Liver regeneration occurs in three main phases: priming, proliferation, termination (26). During the priming phase of liver regeneration, which lasts around 4 hours in mice, the majority of quiescent hepatocytes rapidly re-enter the cell cycle with the help of various cytokines such as tumor necrosis factor α (TNFα) and interleukin 6 (IL-6) (16, 26). These cytokines are mainly produced by nearby liver macrophages, also known as Kupffer cells, that are activated by lipopolysaccharide (LPS) and complement effector proteins such as complement 3a (C3a) and complement 5a (C5a) (27). The proliferation phase occurs when hepatocytes undergo mitosis, and the expansion of the remaining liver occurs with the help of growth factors and metabolic signaling

(15). In rodents, most of the increase in the liver mass occurs by day 3 after partial hepatectomy (PHx), a surgical resection of 2/3 of the liver, with the complete mass restoration achieved by 5 to 7 days (15, 28). Finally, the termination phase occurs with the regulation of various pathways that can alter the hepatic mass (29). Of the main stages of liver regeneration, the priming phase is of great interest because the normally quiescent hepatocytes re-enter the cell cycle to proliferate in response to an injury or an infection (15). A better understanding of this phase of liver regeneration could provide key insights into the complex pathways that activate cellular proliferation and advance our knowledge in regenerative medicine.

Liver regeneration is often studied using the PHx model developed in 1931 (30). This PHx model in rodents requires a surgical resection of approximately 2/3 of the liver which initiates liver regeneration. Another popular model to study liver regeneration is carbon tetrachloride (CCl_4) model (31), in which the process is induced by CCl_4 toxin to mimic the toxic injury involving necrosis and acute inflammation. However, the results from this model are often complicated to interpret properly due to the difficulty of differentiating the effect of the regenerative process from the non-regenerative response to the toxin.

The complement system is part of the innate immune system that has recently been introduced as one of the key regulators of liver regeneration (27, 32-34). In prior studies, complement-knockout mice were used to demonstrate the importance of the complement effector proteins, C3a and C5a, in

mediating successful liver regeneration in mice (33, 34). For example, mice deficient in complement component 3 (C3) and complement component 5 (C5) genes exhibited severe damage to parenchyma, increased necrosis and hepatocyte degeneration, and higher mortality rate than the wild-type during liver regeneration (33, 34). Various cytokines and phosphoproteins also showed significant differences between the knockout and the wild-type mice, especially during the first few hours after PHx (34). Based on these studies, an overall mechanism of the complement-induced liver regeneration with a focus on intercellular signaling, has been proposed (27).

In this Chapter, I report a more complete mechanism of the priming phase of liver regeneration from a concomitant analysis of the transcriptional and metabolic states during the process of regeneration. In order to elucidate the mechanisms involving the complement system proteins, I carried out a comparative analysis, of the transcriptional and metabolic changes as a time course during the priming phase of regeneration, on both wild-type and C3-knockout mice with PHx and sham surgery. The analysis of the transcriptional and metabolic changes during this phase revealed the role of acute phase proteins and the modulation of the regeneration by sterol metabolism. I present the detailed analyses in the sections below.

Methods

Animal studies

PHx experiments, according to the method of Higgins et.al. (30), were performed to remove 2/3 of the liver from 13 to 16 weeks-old young male mice of either C57BL/6 wild-type (WT) or C3-/- (KO) origin. Experiments were also performed on mice with sham surgery to serve as negative controls. These sham experiments were necessary because surgical inflammation can influence the expression of cell-cycle and proliferative genes, which will interfere with the analysis of the priming phase of liver regeneration that is dependent on these biological processes. After the PHx experiments, the remaining parts of the liver were collected after 0.5, 1, and 3 hours to capture the temporal changes during the priming phase of liver regeneration. The livers at 0 hours for both the WT and the KO were also collected to analyze the baseline difference. 3 biological replicates were used for each PHx and sham experiment to account for biological variability.

RNAseq experiments

From the collected livers at various time points, RNA was extracted and purified with Qiagen Allprep Kit to prepare for cDNA synthesis and gene expression analysis. A pooling scheme was devised to reduce the number of samples to 21 (Supplementary Material). The pooled RNA samples were run on Bioanalyzer to check for their RNA integrity. Illumina Truseq cDNA library

construction kit was used to synthesize cDNA libraries after poly-A selection and fragmentation. Then, the cDNA fragments were size-selected and inserted into the flow cell of Hiseq 2000 at the Biogem facility of University of California, San Diego (UCSD). The sequencing option was single-end 50-bp with 7 samples in each of the 3 lanes.

RNAseq pipeline

A new RNAseq pipeline was developed with the existing tools to effectively analyze the transcriptomic data. First, Omicsoft Sequence Aligner (OSA) was used to align the RNAseq reads to the mouse genome and the transcriptome with the default parameters (35). Then, HTseq-count was used to quantify the number of aligned reads associated with each gene and transcript (36). All uniquely mapped reads were counted, but ambiguous reads that mapped to several different genes were ignored. DESeq, a popular R package for RNAseq analysis, was used to derive the list of differentially expressed genes across paired conditions (37). Different combinations of parameters and filtering schemes were optimized to produce the highest number of statistically relevant genes. For example, the genes in the lowest 40% quantile of the total read counts across all samples were removed to increase the power of the statistical analysis while minimizing the removal of differentially regulated genes. The counted reads were normalized with the DEseq's default method, while the variance was estimated for each condition. The statistical tests were performed on each of the 7 core groups between

PHx and sham: KO vs. WT at 0 hours, KO PHx vs. KO sham at 0.5 hours, KO PHx vs. KO sham at 1 hour, KO PHx vs. KO sham at 3 hours, WT PHx vs. WT sham at 0.5 hours, WT PHx vs. WT sham at 1 hour, and WT PHx vs. WT sham at 3 hours. Once the p-values and the false-discovery rate (FDR) values from Benjamini-Hochberg method were calculated for each gene in the core groups, the genes under either the FDR of 0.1 or the p-value of 0.05 were chosen for further analyses.

In addition to the statistical tests performed on PHx vs. sham, the sequential fold change cutoff of 1.5 was applied to create the list of differentially up- or down-regulated genes in the KO with respect to the WT. More specifically, the union of differentially regulated genes from the statistical tests in the KO and the WT was obtained at each time point. Then, from the combined list of genes, only the genes with the KO/WT fold change difference of 1.5 or higher were selected. Both the KO and the WT PHx expression levels were normalized with respect to sham from the previous DESeq analysis. As a result, the final list of differentially regulated genes, which also show significant difference with respect to sham, was generated. For the differential analysis at 0 hours, only the DESeq results with the p-value cutoff of 0.05 were utilized between the KO and the WT because no sham data existed. The complete workflow to derive the list of differentially regulated genes at each time point is shown in the Supplementary Material.

qPCR experiments

The genes, C-fos, c-jun, and TIS21, which showed the greatest changes over time with respect to sham control during the priming phase of liver regeneration from the published study, were selected for qPCR validation (16). Glyceraldehyde 3-phosphate dehydrogenase (GAPDH) was used as the housekeeping gene. The primer sequences for the genes were obtained from the published studies and Primerbank, Harvard online database for PCR primers that have been validated for mouse genes (38, 39). The primer sequences are listed as follows: 5'-CCTTCGGATTCTCCGTTTCTCT-3' (forward) and 5'-TGGTGAAGACCGTGTCAGGA-3' (reverse) for c-fos; 5'-CCTTCTACGACGATGCCCTC-3' (forward) and 5'-GGTTCAAGGTCATGC TCTGTTT-3' (reverse) for c-jun; 5'-ATGAGCCACGGGAAGAGAAC-3' (forward) and 5'-GCCCTACTGAAAACCTTG AGTC-3' (reverse) for TIS21; 5'-AGGTCGGTGTGAACGGATTTG-3' (forward) and 5'-TGTAGACCATGTAGTT GAGGTCA -3' (reverse) for GAPDH. Prior to the qPCR experiments, a set of validation experiments were performed to test the primers and check their PCR efficiencies for 4-log dilution. The qPCR experiments were performed in 2 steps. First, the cDNA library was created from the purified RNAs using High Capacity cDNA Reverse Transcription Kit from Applied Biosystems. Then, the cDNA library was mixed with each of the 4 primers and Fast SYBR Green Master Mix from Applied Biosystems. Real-time fluorescent measurements were taken from the qPCR machine, Eppendorf RealPlex. For each biological replicate, 3 technical replicates were used.

Enrichment Analysis

DAVID analysis (40) was performed to identify significantly enriched Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways (41) and biological processes from Gene Ontology (GO) terms (42) under the modified Fisher-exact p-value cutoff of 0.05. Heatmap was derived for the enrichment results with Circos software (43). Negative $\log_2$ of enrichment p-value was used as the scale for the heatmap.

Network Analysis

An integrative analysis using protein-protein interactions and gene expression data was performed by using a custom mouse interaction network derived in Cytoscape, a network visualization tool (44). This mouse interaction network was derived from the known mouse protein-protein interactions and transcription factor-to-target interactions using online databases such as STRING, BIOGRID, and TRANSFAC (45-48). From the STRING database that uses a scoring scheme between 0 and 1 based on both predicted and experimentally validated protein-protein interactions, a 0.9 cutoff was used to extract highly feasible protein-protein interactions. Autosome clustering from the Clustermaker plugin of Cytoscape was used to identify the clusters of genes that are well-connected within the mouse interaction network (49, 50). Autosome clustering uses the unsupervised training of a self-organizing map, a type of artificial neural network in machine learning (50). This algorithm used correlation from the gene expression data as node weights to find the clusters

of genes that are different from each other. Among these clusters that are separated from one another, the largest clusters with enriched biological functions or KEGG pathways were selected as relevant networks for further analyses.

Metabolomics Studies

Mass spectrometry measurements were made for 143 metabolites for all 3 biological replicates for each genotype, time point, and control (sham, resected). The measured metabolites included cholesterol esters, phospholipids, and metabolites in the sterol pathway (Supplementary Material). The PHx values were first normalized with respect to their sham control. Then, the final fold change between the KO and the WT was calculated by comparing their normalized PHx values. The mass spectrometry measurements of important metabolites were used together with the transcriptomic data to identify novel mechanisms and derive mechanistic insights.

Correlational Analysis

Pearson correlation analyses were performed in Excel. The first correlation analysis compared the cytokine and the transcriptomic profiles. Correlation between the previously published cytokine data and the acute phase genes was performed with a time-delay; the cytokine measurements at 0, 0.5, and 1 hours were compared with the gene expression measurements

at 0.5, 1, and 3 hours, respectively. The second correlation analysis compared the transcriptomic and the metabolic data in the sterol pathway at the same time points. The expanded pathway for sterol lipids was taken from LIPID Metabolites and Pathways Strategy (LIPID MAPS) (51, 52). In the sterol pathway, multiple genes regulated certain metabolites; among these genes, only the gene with the highest correlation with its paired metabolite was selected. The second correlation analysis used three time points, 0.5, 1, 3 hours, for the KO and the WT and 4 time points, 0, 0.5, 1, 3 hours, for the KO/WT fold change. The correlation results in the sterol pathway were visualized through a heat map.

Results

Transcriptional Changes

To explore the changes in the priming phase of complement-induced liver regeneration, I first analyzed the time-series transcriptomic data in the C3 knockout (KO) and the C57BL/6 wild-type (WT) through RNAseq experiments across multiple conditions. The alignment result from the RNAseq experiments produced the following statistics: aside from one outlier, the samples had an average count of 15 million uniquely mapped reads, 3 million of which were splice reads (Supplementary Material). Correlation between the RNAseq and

the qPCR data for the selected genes, c-fos, c-jun, and TIS21, was high (r^2=0.76).

To identify differentially regulated genes, I performed differential analysis using DEseq on PHx vs. sham for every genotype and time point. Then, a fold change cutoff of 1.5 was applied between the KO and the WT to produce the final list of differentially regulated genes. The total numbers of differentially regulated genes are shown in **Figure 1.1ab**. The results showed a trend of increasing number of differentially regulated genes as the priming phase progresses, which suggests that a small set of genes, such as immediate early genes and transcription factors, are influenced first, while the majority of their downstream genes are primed later.

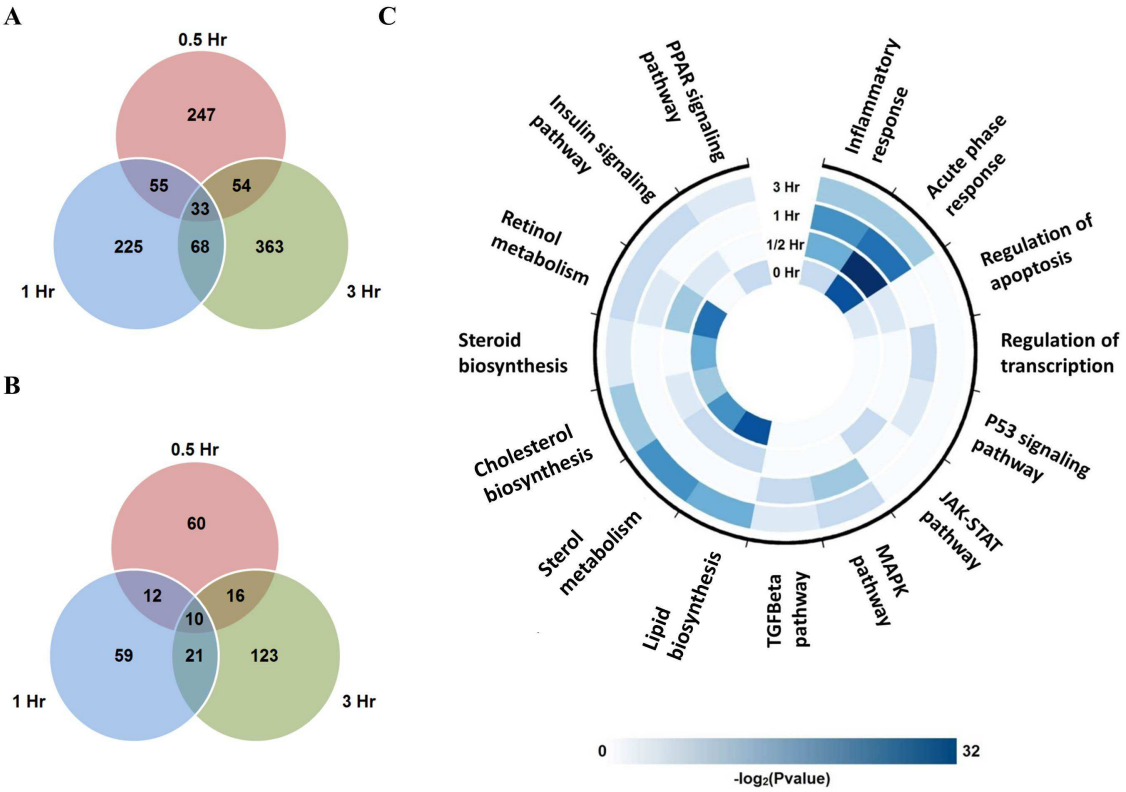


Figure 1.1 Venn diagrams of differentially regulated genes after PHx under the p-value cutoff of 0.05 (*A*) and the FDR cutoff of 0.1 (*B*) and over-represented biological functions and pathways during the priming phase (*C*). *A* and *B*, all data are shown as the number of differentially regulated genes between the KO and the WT at each time point (n=8 (0.5 hr) and n=6 (1-3 hr) mice) *C*, Enrichment analysis was performed at each time point using DAVID from the list of differentially regulated genes under the p-value cutoff of 0.05. Negative $\log_2$ of enrichment p-value was used as the scale for the heatmap (enrichment p≤0.05, a modified Fisher exact test from DAVID)

Activation of multiple signaling pathways

The complement system has been shown to increase the production of IL-6 and TNFα that play a key role in liver regeneration by activating pathways such as NF-κB pathway in Kupffer cells (27). Therefore, I hypothesized and observed that the signaling pathways activated by these cytokines in hepatocytes would be deactivated in the KO mice. For example, enrichment

analysis on the final list of differentially regulated genes for GO terms in biological process and KEGG pathways revealed that the complement system promotes acute-phase response and transcription and activates cell cycle-related pathways, such as mitogen-activated protein kinases (MAPK), and TGFβ pathways, between 1 and 3 hours (**Figure 1.1c**). I also observed metabolic changes occurring in the KO transcriptome; retinol metabolism, cholesterol biosynthetic process, and insulin signaling pathway were significantly enriched. KEGG pathways including peroxisome proliferator-activated receptors (PPAR), janus kinase-signal transducer and activator of transcription (JAK-STAT), and several metabolic pathways were significantly enriched in the KO (**Figure 1.1c**).

Temporal network analysis

In order to investigate the complex temporal changes in the transcriptome, I created networks in Cytoscape at each time point (**Figure 1.2**). These networks were derived from the gene expression data and the custom mouse interaction network composed of protein-protein and transcription factor interactions. Autosome clustering was applied to create clusters, or networks of genes, that were enriched in biological processes. The largest cluster that was enriched in transcription was selected as the most significant network. This network revealed interesting and complex biological insights. For example, based on the transcriptional changes between 0 and 3 hours, c-fos is progressively downregulated in the KO. In addition, while SOCS3 and

STAT3 were briefly upregulated for the first 1 hour, some of their neighbor genes were affected at 3 hours. The complex changes in the transcriptional network across multiple time points further demonstrate the intricate regulation the complement system has on multiple transcription factors and their interacting genes during the priming phase of liver regeneration.

Complement and the acute phase response

Acute phase response was one of the most significantly over-represented biological processes in the KO across all time points from the enrichment analysis (**Figure 1.1c**). Therefore, I analyzed the transcriptional regulation of the acute phase genes from the acute phase response signaling pathway from QIAGEN's Ingenuity Target explorer (53) (**Figure 1.3**). Among these acute phase genes, serum amyloid As and orosomucoid genes not only showed significant transcriptional regulation, but also expressed similar temporal pattern; their KO expressions steadily decreased from 0.5 to 3 hours, whereas their WT expressions increased (**Figure 1.4a-d**).

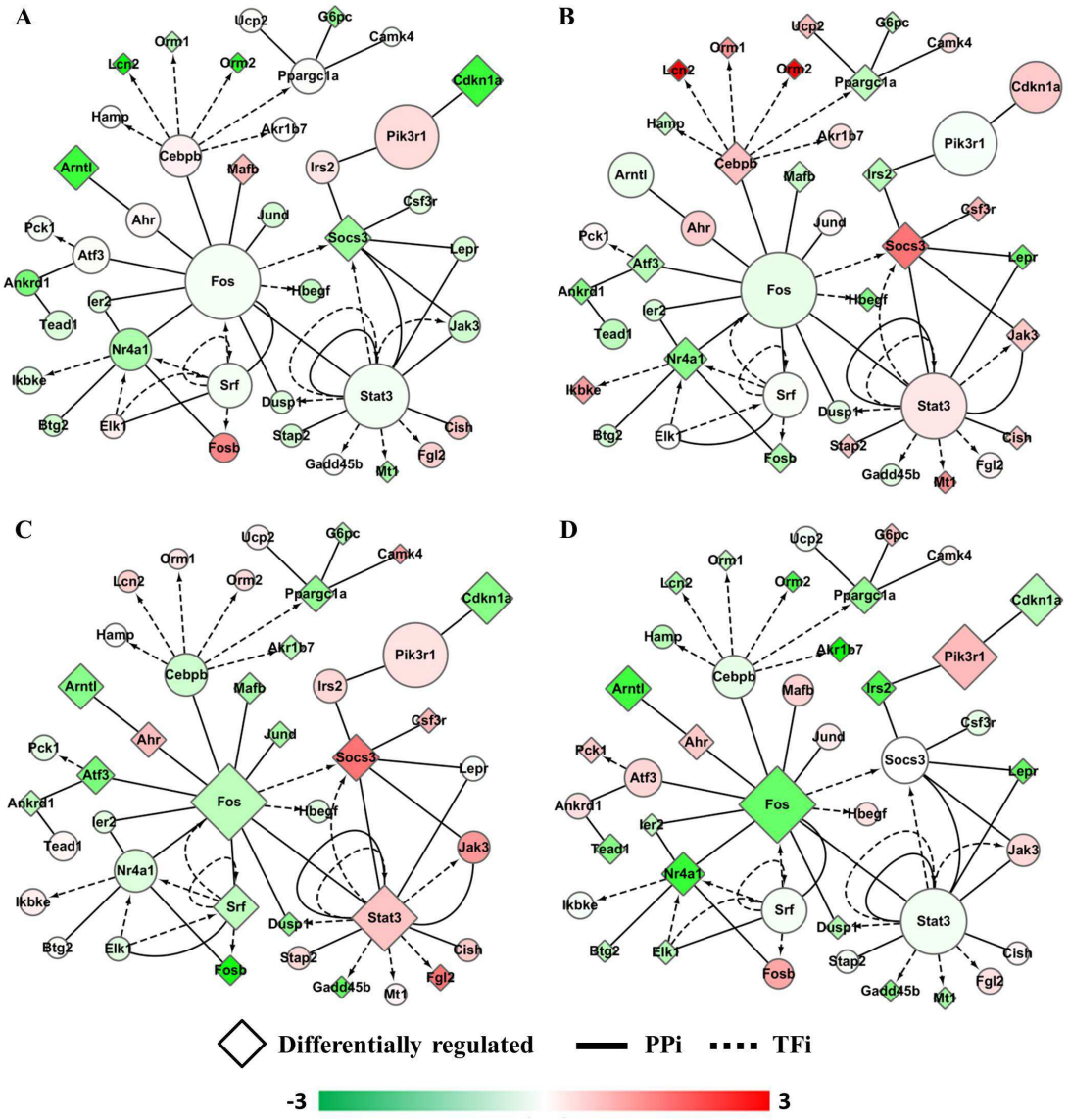


Figure 1.2 Temporal network analysis in Cytoscape. The networks represent the biggest module derived from Autosome clustering at 4 time points: 0 hr (*A*), 0.5 hr (*B*), 1 hr (*C*), 3 hr (*D*). This particular network of 41 genes shows significant enrichment in transcription (FDR<0.1, DAVID). The red color indicates upregulation, whereas the green indicates downregulation, with respect to the WT. The size of a node is based on the number of directly connected genes. The solid line indicates protein-protein interaction (PPi), whereas the directed dashed line indicates transcription factor-target interaction (TFi). The diamond shape represents differentially regulated genes, either under the p-value of 0.05 or the FDR of 0.1, at that particular time point between the KO and the WT.

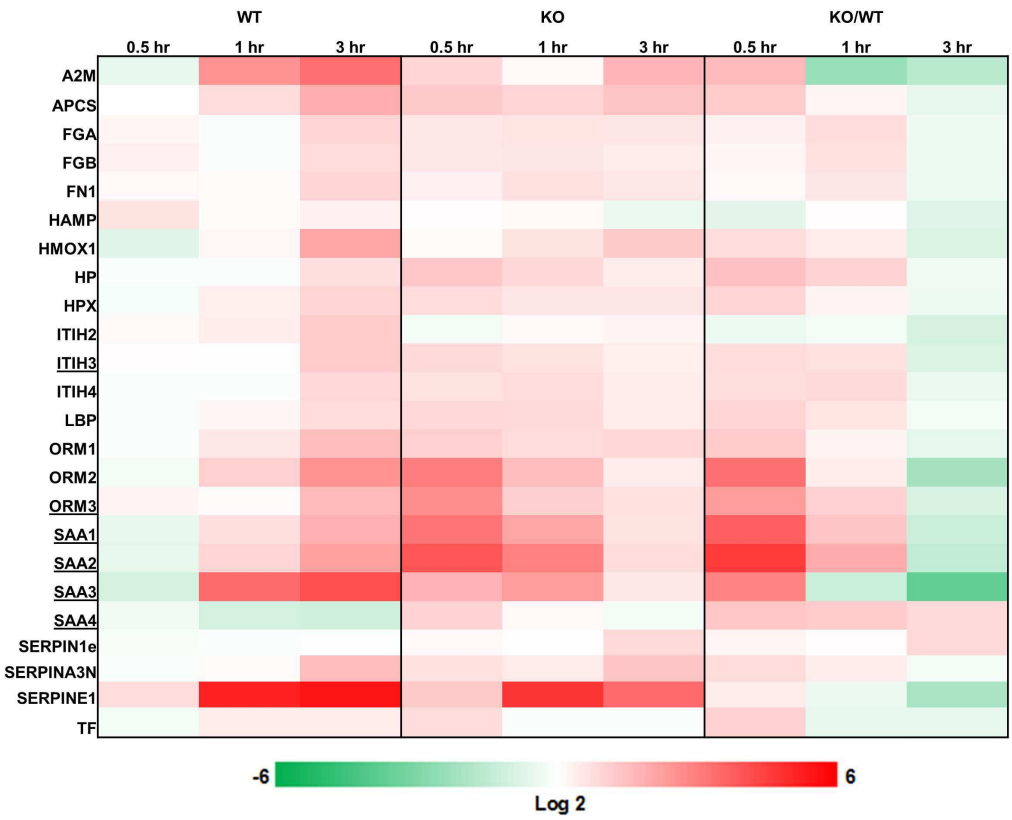
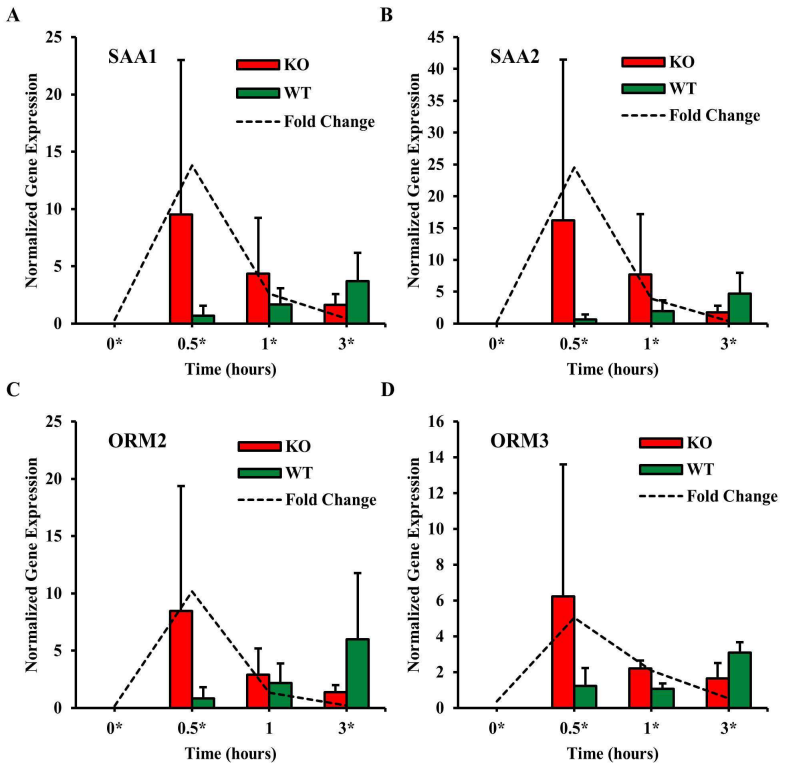


Figure 1.3 Acute phase genes after partial hepatectomy. The heatmap of transcriptional changes in the acute phase genes are plotted for KO, WT, and KO/WT fold change. Log_2 color scale is used. Underlined genes are differentially regulated at all three time points.



E

Cytokine	WT	KO
TNFα	0.999	0.986
KC	0.998	-0.996
IL-1b	0.981	-0.989
IL-10	0.981	-0.383
GM-CSF	0.958	0.911
IL-17	0.880	1.000
MIP-1a	0.715	0.984
RANTES	0.676	1.000
IL-6	0.675	-1.000
IL-5	0.134	-0.950
IL-4	0.126	0.987
G-CSF	0.038	0.947
IL-2	-0.675	0.954
IL-12	-0.820	0.996
IFN-g	-0.921	0.996
IL-1a	-0.935	0.981
IL-3	-0.965	1.000

Figure 1.4 Temporal gene expression of the acute phase proteins (*A-D*) and their correlation with cytokine profiles (*E*). Temporal gene expressions of four acute phase proteins, SAA1(*A*), SAA2(*B*), ORM2(*C*), ORM3(*D*), are plotted. The red bar graph represents the KO data, while the green bar graph represents the WT data. PHx gene expression values from both genotypes are normalized with respect to their sham data. The black line represents the KO/WT fold change. *next to hour points in the x-axis indicates the statistical significance of $p<0.05$ and 1.5 fold change. *E*, Pearson correlation coefficients were calculated for the linear relationship between the measured cytokines from the previous study and SAA1, representing the acute phase genes. Only the PHx measurements were used for both data.

To investigate the source of the trends observed in the gene expression profiles of the acute phase proteins, I compared their PHx measurements with the cytokine measurements from the previous study under the same experimental conditions. Here, the time-delay was necessary to represent the response delay from the cytokine signals to affect their downstream genes. Therefore, I compared the cytokine measurements at 0, 0.5, and 1 hours with

the gene expression measurements at 0.5, 1, and 3 hours, respectively. The temporal profiles of TNFα level showed time-delayed correlation with the temporal profiles of SAA1, representing the acute phase genes; the Pearson correlation coefficients for both the KO and the WT measurements between TNFα and SAA1 with the time-delay were both positive and very close to 1, indicating strong linear relationship (**Figure 1.4e)**. Furthermore, the temporal profiles of other cytokines, including IL-6, did not correlate with those of the acute phase genes as well as TNFα did in both genotypes. In addition to the strong positive correlation observed between TNFα and the acute phase genes, the correlation between c-fos and the acute phase genes was also high, with the Pearson correlation coefficients of 0.89 for the KO and 0.97 for the WT.

Complement and cholesterol metabolism and efflux

In addition to the changes in the early acute phase genes, I observed metabolic changes in the KO transcriptome from the enriched GO terms (**Figure 1.2**). At 0 and 3 hours, various lipid metabolic pathways related to cholesterol metabolism were significantly affected in the KO mice. Therefore, I hypothesized that cholesterol metabolism and efflux may play a significant role during the priming phase of liver regeneration. In the liver, cholesterol homeostasis is closely monitored and regulated by liver X receptor (LXR) that is activated by oxysterols (54). Cholesterol 25-hydroxylase (CH25H), the gene responsible for the synthesis of the main agonist for LXR, 25-

hydroxycholesterol, was significantly downregulated with respect to the WT (**Figure 1.5a**) (55). Other genes related to cholesterol efflux such as ATP-binding cassette sub-family G member 5 (ABCG5) and ATP-binding cassette sub-family G member 8 (ABCG8) also showed downregulation (**Figure 1.5bc**) (56). Upregulation of HMG-CoA reductase (HMGCR) that results in increased cholesterol biosynthesis was observed at 3 hours (**Figures 1.5d**) (57).

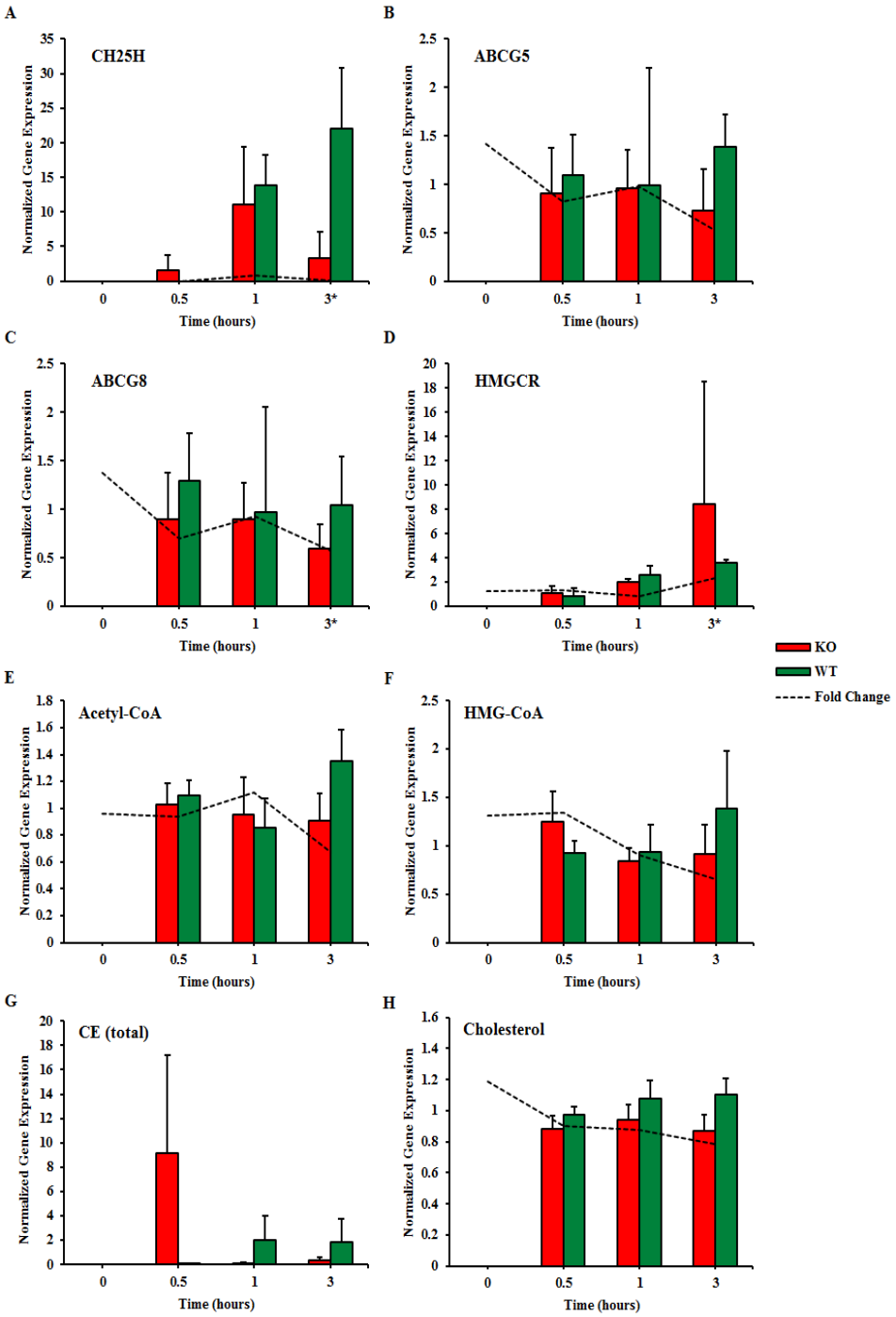


Figure 1.5 Transcriptomic (*A-D*) and metabolic (*E-H*) profiles for cholesterol metabolism: Cholesterol 25-hydroxylase (CH25H) (*A*), ABCC1(*B*), ABCG8(*C*), HMGCR(*D*), Acetyl-CoA(*E*), HMG-CoA(*F*), CE total(*G*), and cholesterol(*H*). The red bar graph represents the KO data, while the green bar graph represents the WT data. PHx gene expression values from both genotypes are normalized with respect to their sham data. The black line represents the KO/WT fold change. *next to hour points in the x-axis indicates the statistical significance of $p<0.05$ and 1.5 fold change.

To observe the predicted metabolic changes based on the transcriptomic results, I analyzed 143 metabolites using mass spectrometry. Some notable metabolites related to the cholesterol homeostasis were acetyl coenzyme A (acetyl-CoA), 3-hydroxy-3-methylglutaryl-coenzyme A (HMG-CoA), and cholesterol esters (CE) (**Figure 1.5efg**). Although the changes in these metabolites at 3 hours were not statistically significantly different between the KO and the WT due to the high variance often observed in metabolic data, their KO/WT fold changes were greater than 1.5. Despite the metabolic and transcriptomic changes related to cholesterol metabolism and efflux, the cholesterol level was moderately consistent across both genotypes at all the time points, which suggests other mechanisms that regulate the cholesterol levels exist (**Figure 1.5h**).

Overall metabolic demands

Cellular proliferation can cause significant remodeling of metabolic signals to meet the new bioenergetic needs of the growing cells, like in cancer (58). Similarly for hepatocytes during liver regeneration, I observed overall decrease in the metabolic measurements in the KO mice at 3 hours; 70% of the metabolites showed lower KO measurements at 3 hours compared to their WT measurements (**Figure 1.6a**). Furthermore, 69% of the metabolites that were not related to cholesterol homeostasis also showed lower KO measurements (Supplementary Material). Lastly, peroxisome proliferator-activated receptor gamma coactivator 1-alpha (PPARGC1A) and solute carrier

family 37 (glucose-6-phosphate transporter) member 1 (SLC37A1), which are involved in energy metabolism, were differentially regulated at all three time points. These results suggest that metabolic demands in the KO are not being met during the later stage of the priming phase.

To further evaluate metabolic changes, I analyzed the list of lipid metabolic genes through a heatmap (**Figure 1.6b**). The list of metabolites was taken from LIPID MAPS website (52). Several cytochrome P450 genes, including Cyp26a1, Cyp2a4, and Cyp4a14, were differentially regulated at all three time points. In addition, PPARGC1A, also known as PGC-1alpha, is a key regulator of energy metabolism. Similarly, SLC37A1 may also regulate energy metabolism by transporting glycerol-3-phosphate between cellular compartments (59, 60).

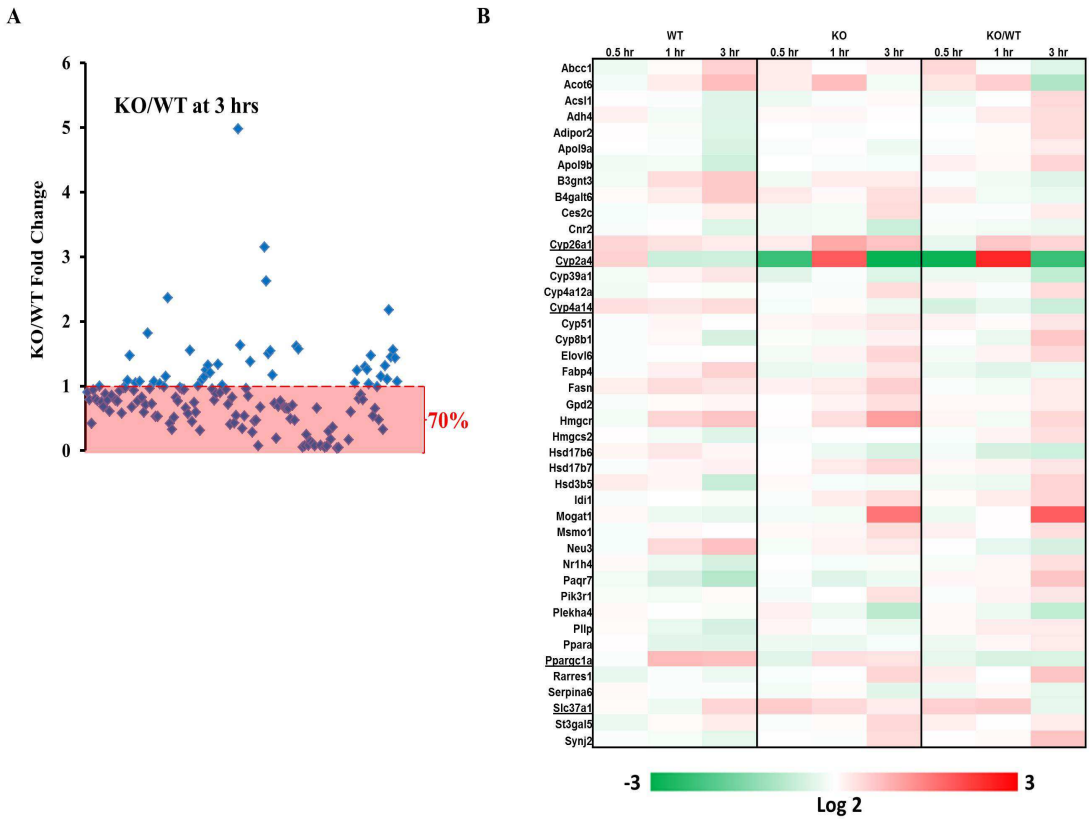


Figure 1.6 Metabolic fold changes at 3 hours (*A*) and the heatmap of lipid metabolic genes (*B*). *A,* KO/WT fold changes at 3 hours are shown for all 143 metabolites. Each blue dot represents the fold change of a single metabolite. *B,* The gene expression profiles that show significantly different changes at either all three time points or only at 3 hours are plotted. Log_2 color scale is used. Underlined genes are differentially regulated at all three time points.

Transcriptomic and metabolic changes in the sterol pathway

Correlation analysis between the transcriptomic and the metabolic data was performed for the gene-metabolite pair in the sterol pathway to observe if the transcriptomic changes would translate into metabolic changes. The majority of the measured gene-metabolite pairs in the sterol pathway showed high correlation (**Figure 1.7**). However, cholesterol esters showed significantly different correlation between the genotypes, which suggests that the mechanism involved in the KO may be different from that of the WT.

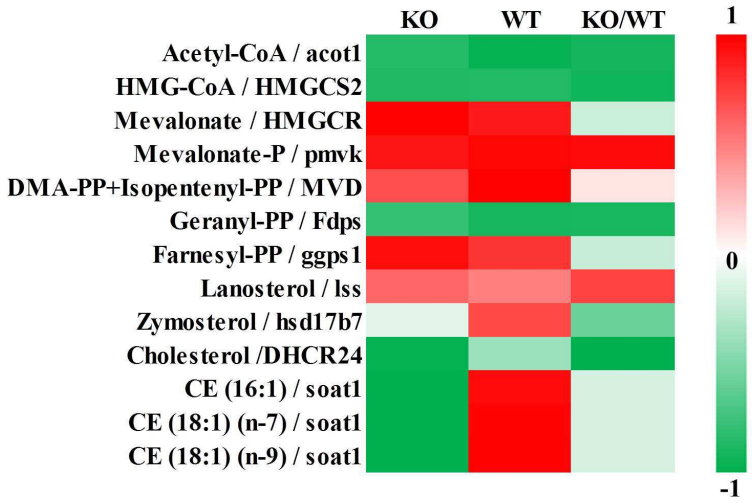


Figure 1.7 Correlation heatmap of gene-metabolite pairs in the sterol pathway. Pearson correlation coefficient was calculated for the measured gene-metabolite pairs in the sterol pathway. If multiple genes are known to regulate a metabolite, the pair with the highest correlation was chosen. The numbers of time points used for correlation calculation were 3, 3, and 4 for KO, WT, and KO/WT categories, respectively.

Discussion

Transcriptional regulation of cell cycle-related pathways

The complement system regulates several significant genes and pathways related to cell cycle and proliferation, two major processes involved in liver regeneration (15). For example, the complement system activates MAPK, p53, JAK-STAT and TGFβ pathways across different time points. In addition, the network analysis shows progressive downregulation of c-fos, an immediate-early gene involved in proliferation, in the KO. Signal transducer and activator of transcription 3 (STAT3) and suppressor of cytokine signaling 3 (SOCS3), which are known regulators of liver regeneration, show brief but

significant downregulation at one of the time points; STAT3 has been linked with cell survival and DNA synthesis during the acute phase, while SOCS3 is a negative regulator of liver regeneration by inhibiting JAK-STAT pathway (61, 62). Other genes in the transcriptional network did not show clear biological insights due to the observed complexity of the temporal transcriptomic changes.

Transcriptional regulation of the acute phase genes by TNFα

The gene expression profiles of the acute phase proteins such as SAA1, SAA2, ORM2, and ORM3 showed that the complement system progressively activates acute phase response during the priming phase of liver regeneration. TNFα, which is regulated by C3a and C5a of the complement system, is the most likely candidate cytokine for regulating the expression of these acute phase genes because the time-delayed correlation between TNFα and the acute phase genes was very high for both the WT and the KO (27). In addition, other cytokines, including IL-6, did not result in moderate or high correlation. Furthermore, several studies have shown that TNFα can regulate the acute phase genes through MAPK pathway (63, 64). Since the correlation between c-fos and the acute phase genes was also high for both the WT and the KO, TNFα may regulate the expression of the acute phase genes through the induction of c-fos and other immediate early genes.

The acute phase genes and cholesterol efflux

The acute phase response has been linked with liver cirrhosis, a severe phenotype in advanced liver diseases with impaired liver regeneration (65-67). For example, ORMs and SAAs have been implicated as potential biomarkers for liver cirrhosis in human (66, 67). These acute phase proteins showed significantly different transcriptional regulation during the priming phase. I further hypothesized that the acute phase proteins may regulate cholesterol that is required for cell cycle progression and cell growth. More specifically, the acute phase proteins may regulate cholesterol efflux through remodeling of high-density lipoprotein (HDL) (68). For example, SAA can remodel native HDL to acute-phase HDL by replacing apolipoprotein A1 as the major apolipoprotein during the acute phase response (69, 70). Acute-phase HDL with the dominant SAA proteins possesses lower capacity to promote cholesterol efflux than the native HDL; in other words, acute-phase HDL has the higher capacity to keep cholesterol within the cells than the native HDL (71, 72). Since the transcriptomic regulations of SAA proteins steadily decreased over time in the KO mice, less SAA proteins were available to remodel native HDL to acute-phase HDL. Higher concentration of native HDL during the first 3 hours is expected to promote higher cholesterol efflux and keep less cholesterol within the cells.

Another evidence for the relationship between SAA-rich acute-phase HDL and the change in cholesterol efflux is total cholesterol ester (CE) delivery. The result from the previous study showed that total CE delivery was

significantly higher for acute-phase HDL than for native HDL (71). The metabolic data was consistent with this knowledge; since the concentration of acute-phase HDL would decrease over time in the KO, the total CE concentration also decreased significantly from 0.5 hours to 1 or 3 hours. Although the variance of total CE concentration at 0.5 hours was high, the majority of CEs with different saturation ratios showed similar profiles.

Regulation of cholesterol metabolism and efflux

Cholesterol is a well-known cell cycle regulator that is tightly controlled by the cells through LXR (73). I hypothesized that if cholesterol level was to decrease as a result of higher cholesterol efflux during the first 3 hours, then hepatocytes would respond promptly by transcriptionally increasing cholesterol biosynthesis and reducing its catabolism and secretion. Several results support this hypothesis. For example, significant downregulation of cholesterol 25-hydroxylase (CH25H) that is responsible for synthesizing 25-hydroxycholesterol was observed at 3 hours. This oxysterol is the main agonist of LXR; it can deactivate LXR pathway and downregulate genes related to cholesterol efflux such as ATP-binding cassette sub-family G member 5 (ABCG5) and ATP-binding cassette sub-family G member 8 (ABCG8) (55, 56). Hepatocytes also stimulated the expression of HMGCR at 3 hours to promote cholesterol biosynthesis to further restore cholesterol availability (57). Upregulation of HMGCR can be explained by the reduced concentration of TNFα, an inhibitor of insulin signaling pathway, because

insulin can strongly stimulate HMGCR synthesis (74, 75). Lastly, another evidence for increased cholesterol availability within hepatocytes in the KO is enriched steroid and bile acid biosynthesis at 3 hours from the enrichment analysis of the transcriptomic data.

As a result of transcriptional regulation of cholesterol metabolism and efflux, which can restore the cholesterol availability lowered by the native HDL during the first 3 hours, the cholesterol level remained relatively stable in both genotypes across all time points. This highlights hepatocytes' tight regulation of cholesterol to ensure successful progression of cell cycle during the priming phase of liver regeneration. Although the resulting cholesterol level did not change much across the genotypes during the priming phase, it may change during the later live regeneration stages, such as the proliferation phase, where significant metabolic changes are known to occur. Moreover, most of the metabolites and their associated genes in the sterol pathway showed high correlation, suggesting that the transcriptomic changes are being translated into metabolic changes.

Metabolic demands during early liver regeneration

Besides cholesterol metabolism, the complement system may help hepatocytes meet other metabolic demands. For example, the majority of the 143 measured metabolites showed lower KO measurements compared to the WT at 3 hours. This result did not change when the metabolites that were related to cholesterol metabolism were excluded. Based on these results, I

hypothesize that when hepatocytes are focused on synthesizing cholesterol and lowering its secretion in the KO mice, fewer resources are made available to meet the other metabolic needs in preparation for prolonged cellular proliferation. Similar phenomenon occurs in cancer cells when they dramatically alter the metabolic circuitry to meet the bioenergetic and biosynthetic demands of increased proliferation (58, 76). The transcriptomic data also revealed that PPARGC1A, the gene involved in energy metabolism, and SLC37A1, the gene involved in transporting glycerol-3-phosphate, are differentially regulated at all three time points. Regenerative process is highly dependent on increased proliferation, and liver regeneration should be no exception.

Limitations of the results

There are several limitations of the results. The first limitation is the small sample size of the pooled RNAseq experiments due to the design of the study having many different experimental conditions. To offset the possibility of Type 1 errors from transcriptomic analysis, the union of two different sets of differentially regulated genes and the sequential fold change cutoff of 1.5 were used to derive the list of differentially regulated genes with higher statistical confidence. The second limitation is the lack of proteomic data to confirm the results based on the transcriptomic data. Unfortunately, transcriptomic changes do not always correlate with proteomic changes due to the instability of the protein and the complex regulatory steps that exist between

transcription and translation. Regardless, future proteomic analysis can supplement the findings of the study. The third limitation is the assumption that the majority of the cells in the liver during early liver regeneration are hepatocytes. About 80% of liver mass is occupied by hepatocytes but the remaining cells, including Kupffer cells, endothelial cells, stem cells, and satellite cells, could also regulate the expression of significant genes identified from the study (77, 78). However, separating hepatocytes from the population of liver cells can be experimentally difficult and is known to induce unnecessary stress signals that can disrupt important priming signals in hepatocytes (79). Lastly, the complement activation may have occurred in the absence of C3. There have been reports of novel complement activation pathways that can generate complement effector proteins even in the C3-/- mice (80, 81). Therefore, the results in the Chapter may be limited to C3-/- biology rather than the comprehensive complement-deficient biology.

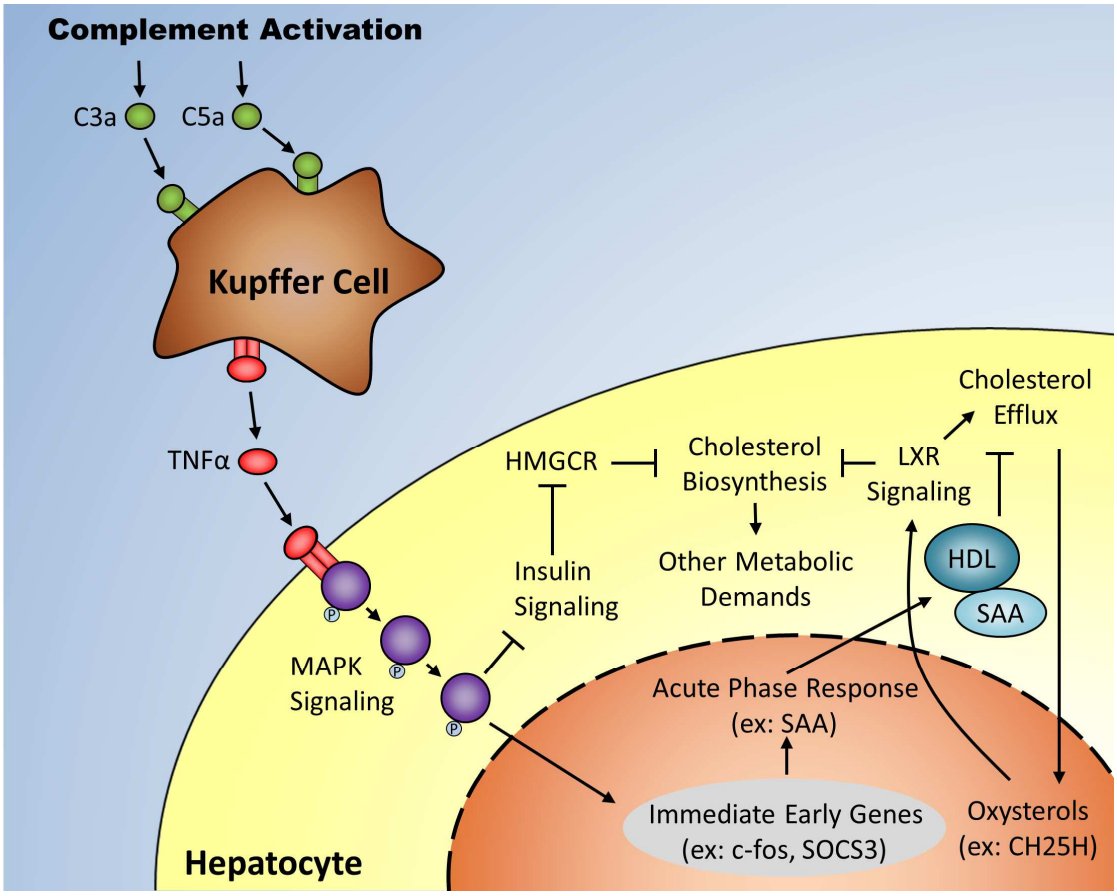


Figure 1.8 Proposed mechanism of the priming phase of complement-induced liver regeneration. The complement activation induced by liver injury or PHx increases the concentration of complement effector proteins, C3a and C5a, which bind to the nearby Kupffer cells to release TNFα. TNFα, then, binds to the receptors on hepatocytes to initiate MAPK pathway whose downstream targets are immediate early genes such as c-fos and SOCS3. This leads to activation of the acute phase proteins such as SAA that can replace the lipoprotein of HDL to form acute-phase HDL. Acute-phase HDL promotes lower cholesterol efflux than the native HDL which causes hepatocytes to respond by activating LXR through oxysterols to reduce cholesterol biosynthesis for stable cholesterol level. TNFα can also inhibit insulin signaling, potentially through MAPK pathway, which then reduces the expression of HMGCR, a key enzyme in cholesterol biosynthesis. The reduced cholesterol biosynthesis allows hepatocytes to use their cellular resources to meet other metabolic demands required for upcoming prolonged proliferation of liver regeneration.

Conclusion

In Chapter 1, I have performed transcriptomic and metabolomic analyses across multiple time points to investigate the mechanism of the

priming phase of complement-induced liver regeneration. Based on the significant results, I have also proposed the comprehensive mechanism of the role of the complement system (**Figure 1.8**). The proposed mechanism highlights the intricate interaction between the complement system, acute phase proteins, cholesterol metabolism, and the priming phase of live regeneration. Future studies including proteomic analysis and investigation of the complement system's role on the later stages of liver regeneration could supplement the findings of this Chapter.

Supplementary Materials

Table S1.1 RNAseq pooling plan. The pooling plan for the RNAseq experiments is shown for all experimental conditions. Since the sham experiments were used as controls for surgical inflammation, only one replicate was used for both KO and WT and 1 and 3 hour points.

Time (hr)	KO		WT	
	PHx	Sham	PHx	Sham
0	1		1	
0.5	2	2	2	2
1	2	1	2	1
3	2	1	2	1

Table S1.2 Alignment results for RNAseq analysis. The table shows fairly consistent alignment results for most of the samples except for one, which had more than twice the reads of any sample.

Sample #	Surgery	Genotype	Time	Uniquely mapped reads	Total reads	Map %
1	Phx	WT	0.5hr	11,317,953	18,636,730	60.73%
2	Phx	WT	0.5hr	14,411,221	24,543,641	58.72%
3	Phx	WT	1hr	34,198,372	56,649,953	60.37%
4	Phx	WT	1hr	12,326,581	20,767,866	59.35%
5	Phx	WT	3hr	13,988,596	23,235,245	60.20%
6	Phx	WT	3hr	15,080,287	25,044,863	60.21%
7	Phx	KO	0.5hr	12,975,059	22,073,690	58.78%
8	Phx	KO	0.5hr	14,057,704	23,429,316	60.00%
9	Phx	KO	1hr	13,762,045	22,544,397	61.04%
10	Phx	KO	1hr	13,926,643	24,471,223	56.91%
11	Phx	KO	3hr	14,522,209	23,553,303	61.66%
12	Phx	KO	3hr	16,556,822	27,852,034	59.45%
13	Sham	WT	0.5hr	13,273,857	23,724,113	55.95%
14	Sham	WT	0.5hr	12,384,498	22,076,570	56.10%
15	Sham	WT	1,3hr	13,288,092	22,291,985	59.61%
16	Sham	KO	0.5hr	12,905,077	22,871,284	56.42%
17	Sham	KO	0.5hr	14,353,902	22,930,972	62.60%
18	Sham	KO	1hr	10,389,784	17,845,983	58.22%
19	Sham	KO	3hr	14,089,053	22,128,453	63.67%
20	WT		0hr	14,788,944	25,472,095	58.06%
21	KO		0hr	14,535,693	24,504,930	59.32%
	Without sample 3		Average	13,646,701	22,999,935	59.35%
			Stdev	1,370,334	2,241,971	2.07%

Table S1.3 Alignment performance of OSA and TOPHAT. OSA's alignment result was compared to that of TOPHAT.

	Uniquely mapped reads	Total reads	Map %	Splice reads	Total number of genes with nonzero expression
OSA	12,504,575	18,636,730	67.10%	2,355,473	18318
TOPHAT	12,176,988	18,636,730	65.34%	2,016,064	17314

Table S1.4 Complete calculations for differentially regulated genes. The complete equations for calculating and classifying differentially regulated genes for all time points are shown.

Time (hr)	KO DE genes	WT DE genes	Final DE genes
0	n/a	n/a	$p.value\left(\frac{KO.0hr}{WT.0hr}\right) > 0.05$
0.5	$p.value\left(\frac{KO.0.5hr.PHx}{KO.0.5hr.sham}\right) > 0.05$	$p.value\left(\frac{WT.0.5hr.PHx}{WT.0.5hr.sham}\right) > 0.05$	From the union of KO DE genes and WT DE genes at 0.5 hours, $\frac{\frac{KO.0.5hr.PHx}{KO.0.5hr.sham}}{\frac{WT.0.5hr.PHx}{WT.0.5hr.sham}} > 1.5 \text{ or } < 1/1.$
1	$p.value\left(\frac{KO.1hr.PHx}{KO.1hr.sham}\right) > 0.05$	$p.value\left(\frac{WT.1hr.PHx}{WT.1hr.sham}\right) > 0.05$	From the union of KO DE genes and WT DE genes at 1 hour, $\frac{\frac{KO.0.5hr.PHx}{KO.0.5hr.sham}}{\frac{WT.0.5hr.PHx}{WT.0.5hr.sham}} > 1.5 \text{ or } < 1/1.$
3	$p.value\left(\frac{KO.3hr.PHx}{KO.3hr.sham}\right) > 0.05$	$p.value\left(\frac{WT.3hr.PHx}{WT.3hr.sham}\right) > 0.05$	From the union of KO DE genes and WT DE genes at 3 hours, $\frac{\frac{KO.0.5hr.PHx}{KO.0.5hr.sham}}{\frac{WT.0.5hr.PHx}{WT.0.5hr.sham}} > 1.5 \text{ or } < 1/1.$

Table S1.5 List of metabolites measured using mass spectrometry. The entire list of metabolites measured in the metabolic data is shown. The number of "?" indicates the uncertainty in the identity of a metabolite based on the mass spectrometry measurements.

alpha tocophereol	L-glutamine
Cholesterol	? phosphoglycolic acid
?? 5-cholesten-3-beta-7-alpha-diol	? phosphoenolpyruvic acid
beta-sitosterol	?? 1,4-dideoxy-1,4-imino-D-arabinitol
?? 3,7,12-trihydroxycholan-24-oic acid	?? 2-amino-2-methyl-1,3-propanediol
5-beta-cholestan-3beta-ol	[1023] pyrophosphate [14.993]
Campesterol	?? D-lyxose
?? Lanosterol	taurine
?? Zymosterol	? Xylitol
lactic acid	? beta-glycerolphosphate
L-mimosine 1	diglycerol
L-valine	glycerol 1-phosphate
Urea	3-phosphoglyceric acid
phosphoric acid	L-ornithine
DL-isoleucine	citric acid
L-proline	?? 5-aminovaleric acid
?? phosphonomycin	? psicose 1
Glycine	D-glucose
succinic acid	?? DL-glyceraldehyde
glyceric acid	? 2,3-dihydroxybiphenyl
L-alanine	D-allose
fumaric acid	L-lysine
L-serine	galacturonic acid
Antiarol	tyrosine
L-threonine	gluconic acid
?? malonic acid	? Palatinose
?? 5,6-dihydro-5-methyluracil	? methyl-beta-D-galactopyranoside
Beta- alanine	isopropyl beta-D-1-thiogalactopyranoside
?? DL-3-aminoisobutyric acid	galactonic acid
? iminodiacetic acid	cytindine-5'-monophosphate
? DL-threo-beta-hydroxyaspartic acid	xanthine
D-malic acid	myo-inositol
?? L-methionine	D-ribose-5-phosphate
aspartic acid	uric acid
?? Amobarbital	? erythrose 4-phosphate
? trans-3-hydroxy-L-proline	? D (+)altrose

Table S1.5 The list of metabolites measured. (Continued)

? trans-3-hydroxy-L-proline 2	? fructose-1,6-diphosphate
L-glutamic acid	?? 2,3-diphosphoglycerate
threonic acid	L-tryptophan
? linoleic acid	*CE 20:5*
D-glucose-6-phosphate	*CE 24:0*
Uridine	*CE 22:4*
Inosine	*CE 24:1*
?? n-acetylneuraminic acid	*CE 22:5*
Sucrose	*CE 22:6*
Adenosine	CE Total
Maltose	*Phospholipid (PL) 14:0*
? Sophorose	*PL 14:1*
1-stearoyl-rac-glycerol	*PL 16:0*
?? trehalose	*PL 16:1*
Isomaltose	*PL 18:0*
?? lactobionic acid	*PL 18:1 (n-9)*
uridine 5'-monophosphate	*PL 18:1 (n-7)*
adenosine-5-monophosphate	*PL 18:2*
? 4-O-Methylphloracetophenone	*PL 18:3 (n-6)*
Acetyl-CoA	*PL 20:0*
HMG-CoA	*PL 18:3 (n-3)*
Mevalonate	*PL 20:1*
Mevalonate-P	*PL 20:2*
DMA-PP+Isopentenyl-PP	*PL 20:3*
Geranyl-PP	*PL 22:0*
Farnesyl-PP	*PL 20:4*
GeranylGeranyl-PP	PL Total
Cholesterol Ester (CE) 14:0	*CE 18:3 (n-3)*
CE 14:1	*CE 20:4*
CE 16:0	*CE 22:1*
CE 16:1	*CE 20:1*
CE 18:0	*CE 20:2*
CE 18:1 (n-9)	*CE 20:3*
CE 18:1 (n-7)	*CE 22:0*
CE 18:2	*CE 20:0*
CE 18:3 (n-6)	

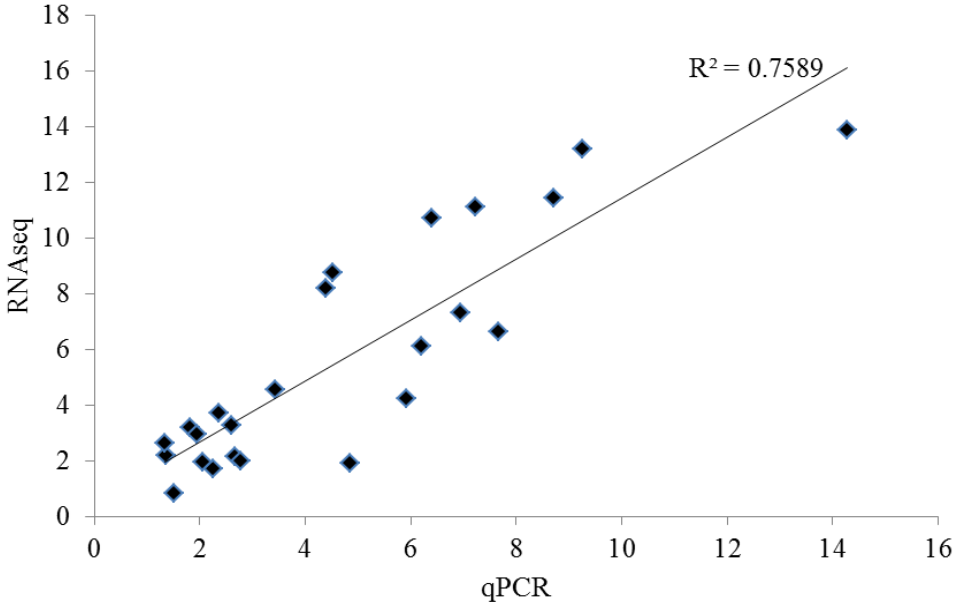


Figure S1.1 Correlation plot between RNAseq and qPCR. R^2 indicates a moderately strong correlation between the two transcriptomic measurements for c-fos, c-jun, and TIS21.

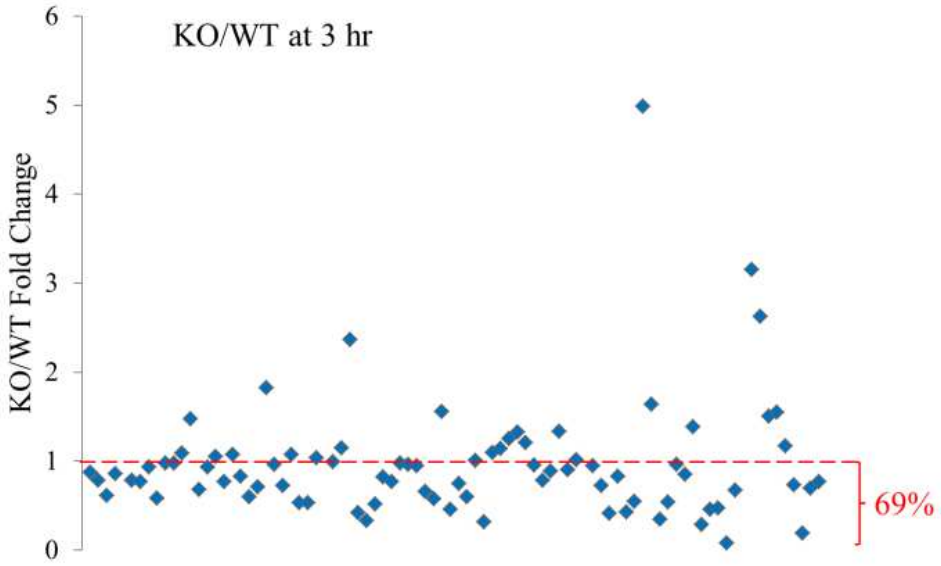


Figure S1.2 Metabolic fold changes at 3 hours. The KO/WT fold change at 3 hours is shown for all metabolites not related to cholesterol metabolism. Each blue dot represents the fold change of a single metabolite.

Table S1.6 List of differentially regulated genes under the FDR cutoff of 0.1 at 0 hours. EntrezID for each gene and the fold change are shown for the differentially regulated genes under the FDR cutoff of 0.1 at 0 hours.

Gene(EntrezID)	Fold Change	Direction
11865	0.204724	Down
12053	0.393643	Down
12266	0.008949	Down
12401	2.856884	Up
12575	0.198582	Down
12660	0.439356	Down
12686	0.322174	Down
12702	0.444323	Down
13087	3.729393	Up
13097	2.623188	Up
13117	2.051496	Up
13119	3.36579	Up
13170	2.932679	Up
13653	0.237975	Down
14245	0.435913	Down
14377	0.419184	Down
14825	0.12529	Down
15446	2.368735	Up
15481	0.451166	Down
15511	0.159204	Down
16006	0.396552	Down
16625	14.23932	Up
16803	0.461486	Down
16819	0.070452	Down
16924	30.6	Up
17167	0.294118	Down
17748	0.496478	Down
18030	0.394558	Down
18113	0.486696	Down
18406	0.177083	Down
18628	2.733906	Up
18950	0.468158	Down
19116	2.138584	Up
20202	0.166667	Down
20208	0.321902	Down

Table S1.6 List of differentially regulated genes under the FDR cutoff of 0.1 at 0 hours. (Continued)

Gene(EntrezID)	Fold Change	Direction
20209	0.224004	Down
20210	0.038726	Down
20211	0.442336	Down
20238	0.246377	Down
20249	0.492884	Down
20503	2.555118	Up
21835	2.276443	Up
22390	3.942857	Up
23893	2.582353	Up
26874	8.777778	Up
27528	4.136842	Up
50770	0.336876	Down
53376	3.360294	Up
53945	2.090191	Up
54608	3.097458	Up
57435	0.276471	Down
57752	0.360386	Down
59012	0.093245	Down
59027	2.307958	Up
64136	0.373161	Down
68616	14.16667	Up
68695	5.304348	Up
70377	0.176367	Down
71145	0.363004	Down
74107	0.15	Down
74126	0.393675	Down
74424	5.684211	Up
76654	3.24484	Up
76737	0.386276	Down
78894	2.561776	Up
79362	2.978495	Up
94071	3.189394	Up
100559	0.011848	Down
103142	2.430442	Up
104158	2.050178	Up
117167	0.397681	Down
171281	4.72	Up

Table S1.6 List of differentially regulated genes under the FDR cutoff of 0.1 at 0 hours. (Continued)

Gene(EntrezID)	Fold Change	Direction
208665	2.11855	Up
209186	0.324786	Down
226016	2.450216	Up
226781	0.421649	Down
231510	3.63	Up
235320	0.161479	Down
240638	2.356452	Up
244416	0.325497	Down
266645	0.336842	Down
333182	37.4	Up
384198	0.198992	Down
432720	2.809764	Up
634650	5.025641	Up
1E+08	3.061523	Up

Table S1.7 List of differentially regulated genes under the FDR cutoff of 0.1 at 0.5 hours. EntrezID for each gene and the fold change are shown for the differentially regulated genes under the FDR cutoff of 0.1 at 0.5 hours.

Gene (EntrezID)	Fold Change	Direction
11910	0.554132	Down
12323	1.83378	Up
13074	1.676251	Up
13082	0.608223	Down
13119	0.417183	Down
14282	0.537545	Down
14377	0.633711	Down
15129	2.51992	Up
15370	0.380304	Down
15505	2.226926	Up
15511	6.128631	Up
15519	1.891496	Up
16007	0.492656	Down
17748	2.467111	Up
18578	0.540451	Down
19734	0.345677	Down
20249	0.447269	Down

Table S1.7 List of differentially regulated genes under the FDR cutoff of 0.1 at 0.5 hours. (Continued)

Gene (EntrezID)	Fold Change	Direction
22057	0.656363	Down
22321	1.917094	Up
23886	0.619285	Down
26874	0.225254	Down
56554	0.377799	Down
57429	2.354608	Up
59012	112.0452	Up
64136	2.211642	Up
66266	1.754417	Up
66438	0.384012	Down
67305	2.26666	Up
67664	0.367302	Down
70377	3.207995	Up
72542	5.951497	Up
74107	2.069946	Up
76737	2.972423	Up
84112	0.522166	Down
84506	0.646868	Down
100559	1.633143	Up
209186	3.231299	Up
211924	0.200474	Down
227627	3.913081	Up
234138	3.221578	Up
244152	1.980171	Up
381530	1.553985	Up
381531	1.675302	Up
384783	0.578006	Down
1E+08	2.62287	Up
1E+08	1.61251	Up
1E+08	2.628045	Up
1E+08	2.225933	Up
1E+08	3.694191	Up
12051	3.103936	Up
12266	1.777198	Up
12475	1.963598	Up
12700	1.510451	Up
12702	3.069246	Up

Table S1.7 List of differentially regulated genes under the FDR cutoff of 0.1 at 0.5 hours. (Continued)

Gene (EntrezID)	Fold Change	Direction
12962	3.143086	Up
13086	0.004984	Down
14654	20.29941	Up
14825	8.690363	Up
15439	2.785179	Up
15458	2.018313	Up
16426	1.79119	Up
16803	2.031897	Up
16819	12.80196	Up
17750	3.52774	Up
18405	2.318999	Up
18406	10.26627	Up
18407	5.094339	Up
18712	1.531466	Up
19152	6.516362	Up
20201	2.436041	Up
20202	2.26456	Up
20208	13.91932	Up
20209	24.77261	Up
20210	7.520327	Up
20211	2.559406	Up
20219	2.357316	Up
21664	2.283476	Up
22041	2.096797	Up
22138	0.244941	Down
27528	0.327731	Down
54608	0.571296	Down
56489	2.188455	Up
67302	2.10902	Up
69553	6.884684	Up
71481	4.697352	Up
71760	0.419508	Down
71780	3.913073	Up
76681	7.79879	Up
76905	3.285607	Up
96875	2.096501	Up
112417	0.505733	Down

Table S1.7 List of differentially regulated genes under the FDR cutoff of 0.1 at 0.5 hours. **(Continued)**

Gene (EntrezID)	Fold Change	Direction
114332	0.642106	Down
114644	2.146963	Up
208292	0.434107	Down
209387	7.081298	Up
211550	3.730184	Up
224674	2.825246	Up
237831	5.88938	Up

Table S1.8 List of differentially regulated genes under the FDR cutoff of 0.1 at 1 hour. EntrezID for each gene and the fold change are shown for the differentially regulated genes under the FDR cutoff of 0.1 at 1 hour.

Gene (EntrezID)	Fold Change	Direction
11465	0.321568	Down
11910	0.344709	Down
12125	0.411548	Down
12475	0.576544	Down
12575	0.376502	Down
12660	0.287547	Down
14281	0.575509	Down
14282	0.143075	Down
14313	0.482736	Down
14377	0.549259	Down
14825	32.54164	Up
15242	1.538072	Up
16326	1.706918	Up
17873	0.268536	Down
19017	0.429856	Down
19041	0.22721	Down
19252	0.399451	Down
20210	0.437674	Down
20515	0.563856	Down
21825	0.583796	Down
21929	2.065802	Up
22138	0.17165	Down
22151	0.502543	Down

Table S1.8 List of differentially regulated genes under the FDR cutoff of 0.1 at 1 hour. (Continued)

Gene (EntrezID)	Fold Change	Direction
54648	0.357824	Down
60599	0.572333	Down
69068	0.428232	Down
69573	0.37528	Down
74211	0.590583	Down
76487	0.229007	Down
80981	0.288017	Down
94226	2.912372	Up
107765	0.600473	Down
16326	1.706918	Up
17873	0.268536	Down
19017	0.429856	Down
19041	0.22721	Down
19252	0.399451	Down
20210	0.437674	Down
20515	0.563856	Down
21825	0.583796	Down
21929	2.065802	Up
22138	0.17165	Down
22151	0.502543	Down
54648	0.357824	Down
60599	0.572333	Down
69068	0.428232	Down
69573	0.37528	Down
74211	0.590583	Down
76487	0.229007	Down
80981	0.288017	Down
94226	2.912372	Up
107765	0.600473	Down
232288	0.547331	Down
240672	0.515911	Down
245195	3.829191	Up
624219	0.349681	Down
11504	1.943473	Up
11622	1.684055	Up
12051	2.550601	Up
12416	0.328352	Down

Table S1.8 List of differentially regulated genes under the FDR cutoff of 0.1 at 1 hour. (Continued)

Gene (EntrezID)	Fold Change	Direction
12606	0.52806	Down
12702	3.135433	Up
12977	2.534045	Up
12986	1.730962	Up
13082	3.420105	Up
13086	106.7133	Up
13170	0.591633	Down
13197	2.052667	Up
13636	1.628049	Up
14284	1.878469	Up
14605	0.51728	Down
14786	0.655433	Down
15220	0.511917	Down
15439	2.051582	Up
15894	2.911318	Up
16175	1.596795	Up
16324	1.70802	Up
16600	0.660816	Down
17750	1.90291	Up
17937	0.381132	Down
18003	1.869924	Up
18030	1.765512	Up
19360	0.230011	Down
19698	1.666539	Up
19734	0.193677	Down
20208	2.606567	Up
20209	3.878866	Up
20296	2.286793	Up
21743	0.364856	Down
24088	1.513674	Up
26877	2.904129	Up
52040	2.517513	Up
53376	0.571905	Down
53412	0.302651	Down
54720	1.608565	Up
55994	0.397286	Down
57247	0.501452	Down

Table S1.8 List of differentially regulated genes under the FDR cutoff of 0.1 at 1 hour. (Continued)

Gene (EntrezID)	Fold Change	Direction
67302	3.203418	Up
69861	2.187984	Up
70377	5.200258	Up
70807	0.264198	Down
72287	0.229269	Down
74194	1.74232	Up
74645	0.339586	Down
78779	0.215229	Down
80859	2.356173	Up
80885	4.713164	Up
96875	4.53107	Up
100090	0.437193	Down
114774	2.239632	Up
117167	2.344133	Up
140887	0.652454	Down
209387	5.49976	Up
211550	2.559212	Up
211770	2.06344	Up
214855	5.149304	Up
229599	0.392694	Down
237831	2.071095	Up
238393	2.611115	Up
242785	0.489417	Down
319520	0.535007	Down
338365	1.568327	Up

Table S1.9 List of differentially regulated genes under the FDR cutoff of 0.1 at 3 hours. EntrezID for each gene and the fold change are shown for the differentially regulated genes under the FDR cutoff of 0.1 at 3 hours.

Gene (EntrezID)	Fold change	Direction
11465	0.505796	Down
11647	0.461494	Down
11865	0.210324	Down
11931	0.427931	Down
11997	0.032968	Down
12014	0.533072	Down

Table S1.9 List of differentially regulated genes under the FDR cutoff of 0.1 at 3 hours. (Continued)

Gene (EntrezID)	Fold change	Direction
12192	0.632444	Down
12227	0.605119	Down
12352	1.510522	Up
12475	0.18859	Down
12575	0.561858	Down
12660	0.403715	Down
12816	0.618893	Down
13803	0.210981	Down
14281	0.294187	Down
14311	1.714423	Up
14313	0.175936	Down
14451	1.747106	Up
14462	37.30401	Up
14612	0.579348	Down
15357	2.276283	Up
15368	0.530595	Down
15370	0.195279	Down
15496	2.560002	Up
16006	0.58128	Down
16181	0.447841	Down
16324	0.62592	Down
16326	5.356849	Up
16625	0.364691	Down
16668	0.405896	Down
16691	0.510542	Down
16763	3.327038	Up
16819	0.474985	Down
16847	0.286724	Down
16918	2.237529	Up
17064	0.606958	Down
17068	0.171466	Down
17691	0.214889	Down
17748	0.48858	Down
17750	0.595455	Down
17864	0.326451	Down
17873	0.376993	Down
18073	0.496603	Down
18405	0.633843	Down

Table S1.9 List of differentially regulated genes under the FDR cutoff of 0.1 at 3 hours. (Continued)

Gene (EntrezID)	Fold change	Direction
18406	0.220316	Down
18787	0.235256	Down
19017	0.426279	Down
19041	0.402875	Down
19730	0.367744	Down
20208	0.424146	Down
20209	0.361094	Down
20210	0.075173	Down
20219	0.664888	Down
20305	0.632098	Down
20494	0.525166	Down
20682	0.612048	Down
21676	0.37672	Down
21817	0.63459	Down
21825	0.295617	Down
23872	0.550438	Down
23886	0.315701	Down
30956	0.549888	Down
50778	0.399779	Down
53608	0.326858	Down
54396	2.043918	Up
54648	0.410667	Down
55927	2.00491	Up
56173	0.663863	Down
56706	0.53191	Down
57266	0.291107	Down
57435	0.597089	Down
57776	0.109855	Down
58804	0.239286	Down
66871	0.543563	Down
67102	0.635502	Down
67302	0.600137	Down
67603	0.527356	Down
68058	0.534805	Down
69573	0.183337	Down
70061	1.867151	Up
70377	3.665629	Up
70620	0.374399	Down

Table S1.9 List of differentially regulated genes under the FDR cutoff of 0.1 at 3 hours. (Continued)

Gene (EntrezID)	Fold change	Direction
71198	0.501514	Down
71481	0.522221	Down
71687	1.769008	Up
71751	0.507663	Down
71904	3.540335	Up
72074	1.70573	Up
72999	0.612395	Down
73205	0.642491	Down
74211	0.53165	Down
74747	0.585503	Down
75739	0.651108	Down
76889	0.46921	Down
76954	0.50736	Down
77037	0.583262	Down
78887	0.320354	Down
80885	0.490015	Down
84112	2.302962	Up
93694	1.929633	Up
99929	0.607388	Down
103784	0.38491	Down
103988	1.537296	Up
104175	1.965131	Up
107869	0.501628	Down
108089	0.159261	Down
108958	2.149326	Up
210808	0.561361	Down
216227	9.308136	Up
217082	1.739717	Up
224440	0.468602	Down
227929	0.605903	Down
231510	0.463964	Down
232288	0.615062	Down
232345	0.30959	Down
233011	0.422011	Down
235493	1.869746	Up
237831	0.288119	Down
242721	3.339599	Up
244416	1.874705	Up

Table S1.9 List of differentially regulated genes under the FDR cutoff of 0.1 at 3 hours. (Continued)

Gene (EntrezID)	Fold change	Direction
246746	0.551615	Down
338365	0.663426	Down
384059	7.121395	Up
384783	0.231037	Down
433022	1.859645	Up
11856	0.335202	Down
12879	0.400181	Down
13074	0.619451	Down
13082	2.352481	Up
13086	0.013076	Down
13836	1.650854	Up
14104	1.587242	Up
14373	3.070845	Up
15490	1.701754	Up
15945	14.76109	Up
16885	1.585945	Up
18207	0.552215	Down
20342	5.445114	Up
20454	1.536252	Up
20519	2.098459	Up
20775	1.638597	Up
26874	0.662171	Down
26877	1.946864	Up
27400	0.346344	Down
54720	2.040431	Up
55994	0.621645	Down
56459	0.510155	Down
57738	4.295452	Up
59012	25.21056	Up
60599	2.22959	Up
66234	1.977918	Up
67379	0.566115	Down
68054	8.993582	Up
69049	2.66007	Up
69847	0.27604	Down
70080	0.078743	Down
80289	1.53375	Up
94071	0.127049	Down

Table S1.9 List of differentially regulated genes under the FDR cutoff of 0.1 at 3 hours. (Continued)

Gene (EntrezID)	Fold change	Direction
109222	3.481015	Up
170439	2.332741	Up
171543	0.64398	Down
209760	1.790693	Up
242484	0.549392	Down
245038	0.530994	Down
338364	0.387189	Down
381531	5.114109	Up
1E+08	1.540822	Up
1E+08	3.508939	Up
1E+08	2.729114	Up
1E+08	2.636495	Up

Table S1.10 Metabolite measurements of phospholipids for KO conditions. Phospholipids with different saturation ratios were measured.

PL	C3-/-	C3-/-	C3-/-	C3-/-	C3-/-	C3-/-	C3-/-
	0	0.5	0.5	1	1	3	3
	-	Sham	Phx	Sham	Phx	Sham	Phx
14:0	0.21	0.11	0.11	0.61	0.18	0.06	0.18
14:1	0.02	0.01	0.00	0.08	0.02	0.00	0.02
16:0	11.44	11.70	10.06	12.71	10.87	10.67	12.50
16:1	0.47	0.54	0.59	0.60	0.46	0.52	0.80
18:0	8.55	8.28	7.43	9.79	8.91	7.91	9.05
18:1 (n-9)	3.09	3.33	2.97	3.48	3.02	3.00	3.76
18:1 (n-7)	0.64	0.79	0.72	0.65	0.59	0.80	0.81
18:2	9.44	9.76	9.62	8.57	8.52	8.41	10.44
18:3 (n-6)	0.10	0.13	0.13	0.17	0.08	0.07	0.10
20:0	0.15	0.15	0.17	0.19	0.17	0.15	0.19
18:3 (n-3)	0.13	0.15	0.19	0.15	0.16	0.14	0.23
20:1	0.11	0.12	0.12	0.13	0.12	0.09	0.07
20:2	0.15	0.18	0.17	0.15	0.12	0.16	0.13
20:3	0.64	0.57	0.69	0.61	0.53	0.60	0.63
22:0	0.25	0.18	0.16	0.51	0.22	0.15	0.23
20:4	6.09	6.12	5.64	5.18	4.86	5.45	5.75
22:1	0.14	0.06	0.05	0.24	0.12	0.06	0.09
20:5	0.43	0.41	0.48	0.31	0.34	0.38	0.48
24:0	0.12	0.14	0.11	0.18	0.15	0.08	0.14
22:4	0.14	0.13	0.16	0.22	0.14	0.06	0.10
24:1	0.19	0.20	0.23	0.17	0.08	0.14	0.17
22:5	0.28	0.31	0.34	0.28	0.27	0.27	0.37
22:6	4.72	4.35	4.23	3.77	3.99	4.02	4.92
Total	47.51	47.72	44.37	48.76	43.91	43.20	51.17

Table S1.11 Metabolite measurements of phospholipids for WT conditions. Phospholipids with different saturation ratios were measured.

PL	WT	WT	WT	WT	WT	WT	WT
	0	0.5	0.5	1	1	3	3
	-	Sham	Phx	Sham	Phx	Sham	Phx
14:0	0.10	0.46	0.13	0.19	0.78	0.07	0.37
14:1	0.00	0.05	0.01	0.02	0.11	0.00	0.04
16:0	9.68	13.63	8.91	15.90	15.87	11.67	13.02
16:1	0.46	0.60	0.45	0.69	0.81	0.50	0.62
18:0	6.54	9.52	5.97	10.83	12.53	6.86	9.73
18:1 (n-9)	2.61	4.05	2.74	4.26	5.09	3.00	4.32
18:1 (n-7)	0.60	0.59	0.46	0.87	0.82	0.59	0.74
18:2	8.45	9.25	7.17	12.11	9.70	11.12	10.61
18:3 (n-6)	0.09	0.16	0.12	0.18	0.24	0.15	0.17
20:0	0.13	0.20	0.15	0.19	0.23	0.15	0.18
18:3 (n-3)	0.15	0.21	0.16	0.20	0.27	0.22	0.26
20:1	0.09	0.13	0.10	0.14	0.18	0.11	0.14
20:2	0.12	0.18	0.13	0.18	0.23	0.17	0.22
20:3	0.55	0.65	0.49	0.73	0.73	0.60	0.64
22:0	0.17	0.35	0.20	0.32	0.53	0.16	0.51
20:4	4.79	4.99	3.91	7.02	5.11	5.72	5.23
22:1	0.05	0.15	0.09	0.03	0.20	0.08	0.35
20:5	0.41	0.51	0.37	0.55	0.37	0.51	0.48
24:0	0.13	0.17	0.12	0.19	0.29	0.14	0.22
22:4	0.13	0.15	0.15	0.13	0.29	0.29	0.21
24:1	0.16	0.15	0.14	0.19	0.28	0.27	0.22
22:5	0.30	0.27	0.24	0.42	0.34	0.40	0.35
22:6	4.36	3.88	3.03	6.18	4.02	5.26	4.48
Total	40.04	50.28	35.24	61.51	59.00	48.04	53.13

Table S1.12 Metabolite measurements of cholesterol esters for KO conditions. Cholesterol esters with different saturation ratios were measured.

CE	C3-/-	C3-/-	C3-/-	C3-/-	C3-/-	C3-/-	C3-/-
	0	0.5	0.5	1	1	3	3
	-	Sham	Phx	Sham	Phx	Sham	Phx
14:0	0.28	0.01	0.49	0.56	0.01	0.28	0.05
14:1	0.02	0.00	0.04	0.05	0.00	0.02	0.01
16:0	4.65	1.00	5.97	5.88	0.86	4.03	1.44
16:1	0.12	0.01	0.20	0.25	0.01	0.15	0.04
18:0	4.92	0.40	5.65	7.08	0.33	3.48	1.02
18:1 (n-9)	1.07	0.13	2.09	3.15	0.11	1.37	0.43
18:1 (n-7)	0.07	0.01	0.16	0.22	0.01	0.09	0.03
18:2	0.22	0.08	0.19	0.19	0.12	0.15	0.14
18:3 (n-6)	0.00	0.00	0.00	0.00	0.00	0.00	0.00
20:0	0.04	0.00	0.06	0.07	0.00	0.03	0.01
18:3 (n-3)	0.00	0.00	0.01	0.00	0.00	0.01	0.00
20:1	0.01	0.00	0.04	0.03	0.00	0.01	0.00
20:2	0.07	0.00	0.08	0.17	0.00	0.10	0.02
20:3	0.01	0.00	0.01	0.00	0.00	0.01	0.00
22:0	0.03	0.00	0.04	0.05	0.00	0.01	0.01
20:4	0.04	0.01	0.03	0.01	0.01	0.05	0.03
22:1	0.00	0.00	0.01	0.00	0.01	0.02	0.00
20:5	0.02	0.00	0.01	0.05	0.00	0.02	0.00
24:0	0.01	0.00	0.02	0.07	0.00	0.03	0.01
22:4	0.00	0.00	0.00	0.00	0.00	0.00	0.00
24:1	0.00	0.00	0.00	0.00	0.00	0.00	0.00
22:5	0.00	0.00	0.00	0.00	0.00	0.00	0.00
22:6	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Total	11.56	1.65	15.09	17.82	1.49	9.88	3.22

Table S1.13 Metabolite measurements of cholesterol esters for WT conditions. Cholesterol esters with different saturation ratios were measured.

CE	WT	WT	WT	WT	WT	WT	WT
	0	0.5	0.5	1	1	3	3
	-	Sham	Phx	Sham	Phx	Sham	Phx
14:0	0.00	0.44	0.00	0.12	0.36	0.04	0.13
14:1	0.00	0.04	0.00	0.01	0.03	0.00	0.01
16:0	0.61	5.76	0.35	2.64	4.84	1.20	1.72
16:1	0.04	0.22	0.00	0.11	0.18	0.03	0.09
18:0	0.18	5.88	0.18	2.08	4.79	0.74	1.53
18:1 (n-9)	0.25	2.25	0.04	0.79	1.81	0.36	0.96
18:1 (n-7)	0.01	0.17	0.00	0.06	0.14	0.02	0.08
18:2	0.16	0.16	0.02	0.34	0.23	0.12	0.16
18:3 (n-6)	0.00	0.00	0.00	0.00	0.00	0.00	0.00
20:0	0.00	0.06	0.00	0.02	0.04	0.01	0.02
18:3 (n-3)	0.00	0.00	0.00	0.01	0.00	0.00	0.00
20:1	0.00	0.00	0.00	0.01	0.01	0.00	0.01
20:2	0.00	0.05	0.00	0.03	0.04	0.01	0.03
20:3	0.00	0.00	0.00	0.01	0.01	0.00	0.00
22:0	0.00	0.03	0.00	0.02	0.03	0.01	0.01
20:4	0.02	0.03	0.01	0.06	0.05	0.02	0.03
22:1	0.00	0.00	0.00	0.00	0.01	0.00	0.00
20:5	0.00	0.00	0.00	0.01	0.02	0.00	0.01
24:0	0.00	0.05	0.00	0.02	0.04	0.00	0.02
22:4	0.00	0.00	0.00	0.00	0.01	0.00	0.00
24:1	0.00	0.00	0.00	0.00	0.00	0.00	0.00
22:5	0.00	0.00	0.00	0.00	0.00	0.00	0.00
22:6	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Total	1.27	15.13	0.60	6.33	12.63	2.58	4.81

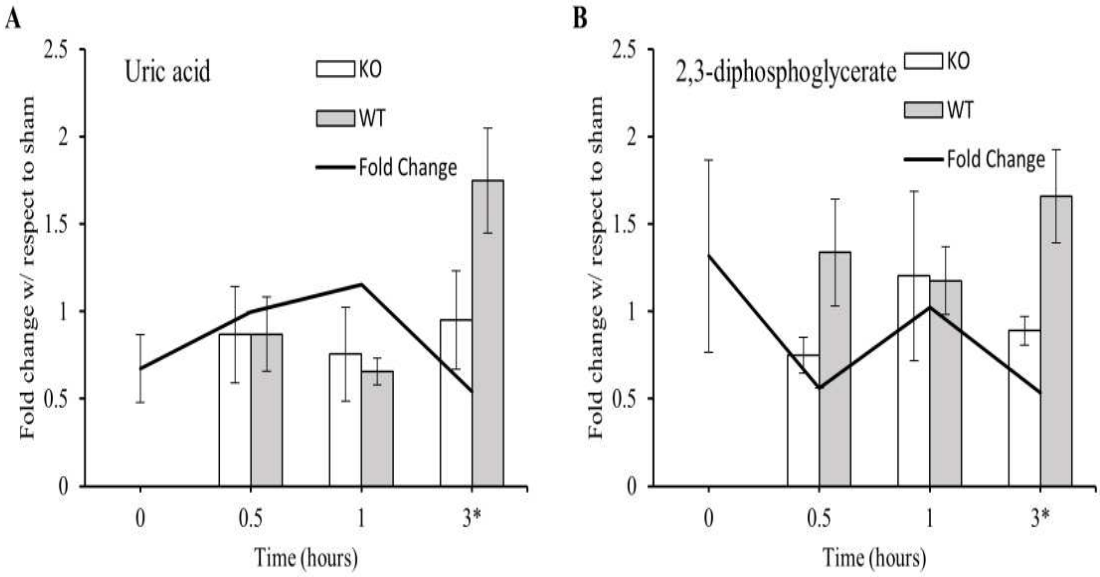


Figure S1.3 Metabolic profiles of uric acid and 2,3-diphosphoglycerate. Metabolic profiles of uric acid and 2,3-diphosphoglycerate are shown. The white bar graph represents the KO data, while the grey bar graph represents the WT data. The black line indicates the fold change difference between the KO and the WT. * next to hour points in the x-axis shows statistical significance of $p<0.05$.

Acknowledgements

Chapter 1, in full, is a re-editing of materials currently being prepared for submission for publication in Jun Min, Robert DeAngelis, Charles Evans, Mano Maurya, Shakti Gupta, Charles Burant, John Lambris, and Shankar Subramaniam. Systems analysis of complement-induced priming phase of liver regeneration, **in preparation**. The dissertation author is the primary investigator and author of this paper. The dissertation author is responsible for designing the overall experiments, performing all analytic methods, and writing of the paper.

CHAPTER 2: TRANSCRIPTOMIC AND INTEGRATIVE ANALYSES OF BILIARY ATRESIA

Introduction

Biliary atresia (BA), or absence of ducts, which drain bile outside the liver, is a rare condition whose pathogenesis must be understood in order to limit its significant public health impact. Despite the rarity, BA is the most common cause of liver failure at birth and of liver transplantation in children, worldwide (82-84). Disease pathogenesis is likely complex, based on circumstantial associations of BA with viral infections (85), a toxin-associated BA-like condition in animal model (86), an interleukin 8 (IL8)-predominant inflammatory signature in affected human liver tissues (87), and multiple susceptibility genes identified in genome-wide association studies from independent investigators (88).

While the complex pathogenesis of BA requires critical evaluation, we face some clinical and experimental challenges. Because the disease is often congenital, investigating the role of various factors would require sequential evaluation of fetal and perinatal liver tissues, a task that poses ethical and technical challenges in humans. Furthermore, modeling the interaction between different potential etiologic factors in an animal model is also difficult in the presence of combinatorial effects of multiple susceptibility genes and unknown extraneous triggers required for disease onset and progression (88).

There are multiple forms of BA because of phenotypic heterogeneity. The rarer congenital or the embryonic form of the disease is often associated with other congenital anomalies such as polyspenia and situs invernus (89). The embryonic form of BA occurs in fewer affected infants who are often born with jaundice or bile duct injury which suggests a complication during fetal development (90). The perinatal form, also referred to as the “acquired” form, is the most common form of BA that is characterized by the development of inflammatory and fibrotic symptoms after jaundice-free birth (91). Around 80% of the cases for biliary atresia are acquired form that entails a progressive inflammatory injury of the bile ducts, leading to fibrosis and ablation of the bile ducts (92). There is also the rare cystic variant of BA that is characterized by the presence of a cystic malformation near the obstructed common bile duct (93).

With the emergence of new observations that contradict the established understanding of BA, a more comprehensive view is necessary to design therapeutic or preventative interventions. Our previous understanding of BA attributed the most common ‘isolated’ form of BA to inflammatory or acquired perinatal factors and the less common 'syndromic' form of BA with laterality-type extrahepatic defects to defective genetic programs (87-89). However, we have observed that both forms of BA are associated with poorly developed cilia in cholangiocytes (94), whose proper function is essential for the correct left-right placement of unpaired cardiovascular and gastrointestinal organs. This suggests that both forms of BA may not be as different as previously

thought. We have also observed a group of BA patients with multiple malformations without laterality defects – symptoms that are different from the common forms of BA (95). Finally, despite the association of the 'isolated' BA with inflammation, bile duct patency after surgical reconstruction is not improved with adjunctive steroids that should limit inflammatory responses (96).

In Chapter 2, the first chapter in analyzing the BA data, I focused on performing RNAseq analysis and devising a novel and effective method of integrating both RNAseq and GWAS data. For the RNAseq analysis, I first identified differentially regulated genes and performed enrichment analysis to better understand the BA transcriptome. For the integrative analysis, I devised two different methods of analyzing RNAseq and GWAS data simultaneously and identified the pairs of differentially regulated genes and nearby genomic variants that are associated with BA. I also analyzed the list of genomic variants that did not lead to differential regulation of their nearby genes.

Methods

RNAseq analysis

The human libraries for mRNA-Seq were prepared using Illumina's mRNA Sequencing Sample Preparation kit (#RS-930-1001). Two sample pools of 6 patients each for BA and normal liver allografts prior to implantation were used as RNA libraries. Briefly, poly-A containing mRNA was purified from

total RNA pools using oligo-dT magnetic beads and then subjected to fragmentation. The fragmented RNA was reverse transcribed using random hexamers followed by second strand cDNA synthesis. cDNA fragments were subjected to end-repair, with addition of a single ‘A’ base and then ligation of Illumina adapters. Ligated products were purified on a 2% Agarose gel for size selection and then amplified by PCR to create the final cDNA library. 36 bp, paired-end sequencing was performed on a Genome Analyzer II from Illumina.

CASAVA software was used to generate a table of exon, gene, and splice counts for all chromosomes in both BA and normal groups. DEseq (37), a popular R package for RNAseq data, was used for differential analysis. The genes among the 40% lower quantile of the total read counts were removed to maximize the statistical power while maintaining the majority of differentially regulated genes when the statistical analysis was performed without the removal. For proper variance estimation, the DEseq options of “blind” and “fit-only” were used. The resulting dispersion plot was checked for quality.

The enrichment analysis of differentially regulated genes under the statistical threshold of the false discovery rate (FDR) of 0.1 was performed using DAVID (40). More specifically, KEGG, Panther (97), and BIOCARTA (98) pathways and Gene Ontology: biological process, molecular function, and cellular components were checked for enrichment. Another enrichment analysis was performed with the differentially regulated genes under the p-value of 0.05 which include more genes than using the FDR cutoff. For

specific pathways and biological functions of interest, the numbers of up- and downregulated genes were counted to show the general direction of regulation.

Integrative analysis using RNAseq and GWAS data

An integrative pipeline was designed to identify differentially regulated genes that are accompanied by BA-associated SNPs. From the list of differentially regulated genes under the p-value of 0.05, SNPs that are within 20kb upstream and downstream of each gene were selected as potential BA-associated SNPs. A set-based test from PLINK (99) was performed to identify representative SNPs that are not in linkage disequilibrium (LD) with other SNPs for each gene, perform a family-based transmission disequilibrium test (TDT) for each SNP, calculate the average statistics for each gene based on the TDT statistics, and calculate the empirical p-value for each gene after permuting the phenotype labels of the dataset 10,000 times. The maximum number of representative SNPs for each gene was set to the default value of 5. For the GWAS TDT within the set-based test, 36 family trio genotypes on HumHap Infineum 550K Illumina SNP arrays were used; each family trio consisted of two unaffected parents and one affected child. TDT was preferred over the conventional chi-squared test with principal component analysis (PCA) because the evidence of both linkage and association from TDT can increase the confidence of the biomarkers identified by the integrative results. Furthermore, for quality control of the GWAS data, SNPs with (i) more than 10% missing genotype across all individuals, (ii) minor allele frequency of less

than 0.05, (iii) Hardy-Weinberg equilibrium of greater than 0.001, or (iv) Mendelian error rate of greater than 0.1 across all individuals were excluded from the test. In addition, Mendelian error rate of 0.05 was used to exclude families that did not meet the threshold. The empirical p-value cutoff of 0.05 was used for the set-based test.

The second method of the integrative pipeline was designed to start with the GWAS analysis of SNPs. TDT analysis with the same options as described previously was performed with one change; all 550k SNPs were tested instead of only the potential SNPS derived from RNAseq analysis. To map SNPs to genes, expression quantitative trait loci (eQTL) mapping was used. eQTL is a popular integrative technique to discover a list of SNPs that have been shown to regulate the expression levels of nearby (cis) or distant (trans) genes (100). Unfortunately, eQTL analysis often requires a very large sample size to reduce the number of false-positive mapping. Therefore, instead of de-novo eQTL mapping, I used the previously published eQTL mapping from a study published in 2008 by Dr. Schadt's group who analyzed 39,000 transcripts and 782,476 SNPs in more than 400 human liver samples (101). The eQTL table from the study was used to identify differentially regulated genes that would be targeted by the BA-associated SNPs. The overall methods in the chapter are depicted in the figure below (**Figure 2.1**).

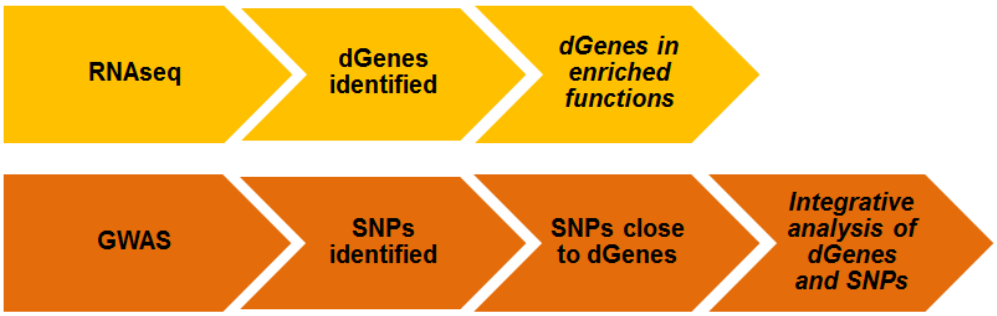


Figure 2.1 Workflows for the transcriptomic and the integrative analyses. For transcriptomic profiling, RNAseq analysis was performed to identify differentially regulated genes (dGenes). Then, enrichment analysis was performed to identify over-represented biological functions and pathways among dGenes. The integrative analysis involved selection of SNPs near dGenes from the GWAS data and application of the set-based test in PLINK to identify pairs of significant genes and SNPs. The integrative analysis also involved eQTL mapping to further complement the results. Italicized results were target sequenced.

For the majority of the SNPs that did not lead to any significant transcriptional changes, Functional SNP prediction tool (FuncPred) from National Institute of Environmental Health Science was used to identify enriched biological functions (102). This particular webtool tested the list of SNPs for transcription factor-binding site activity, splicing regulation, miRNA binding site activity, non-synonymous coding SNPs, and stop codons.

Results

RNAseq preprocessing results

To identify differentially regulated genes, I performed RNAseq analysis on 2 pools of explanted liver tissue samples from 6 BA patients and 6 normal pre-implant liver allografts. However, before performing the differential analysis to identify significant genes in the BA group, I examined the distributions of RNAseq reads in both the BA and the normal group to ensure the majority of the genes had enough sequencing depth for proper statistical comparison (**Figure 2.2**). Although there were 3136 human genes with no reads, the majority of the genes showed at least 1 read count. In addition, the distribution of read counts was not dominated by a few genes that were deeply sequenced; there were less than 5% of genes with more than 1000 read counts. A similar distribution of reads was observed with the normal group.

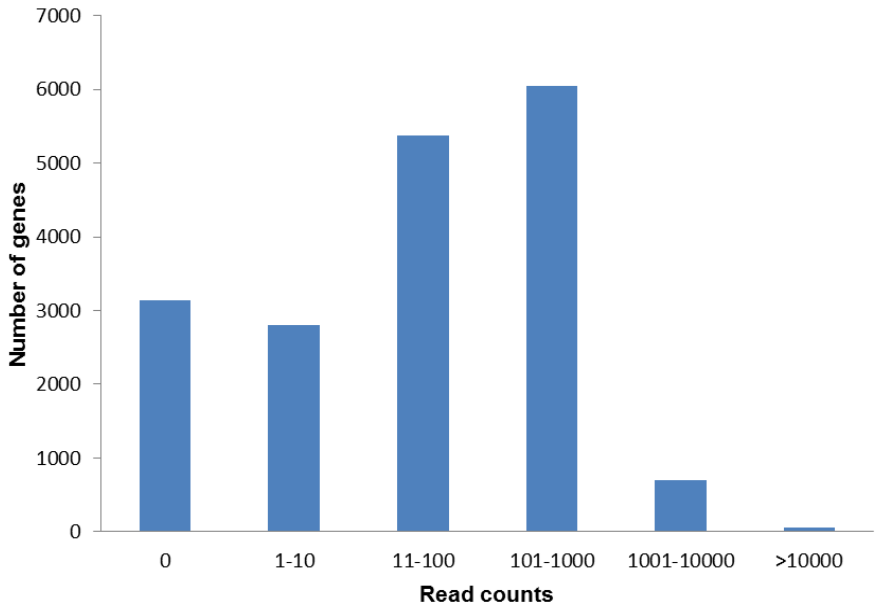


Figure 2.2 Distribution of RNAseq read counts in BA. Bar graphs showing the distribution of RNAseq read counts across all the genes. The number of genes indicates how many genes have certain number of read counts.

In order to estimate the variance for the pooled samples of BA and normal RNAesq data, I tested various combinations of options within DEseq package in R. The final option was chosen based on the dispersion plot (**Figure 2.3**). The black dots indicate each of the variance plotted along the mean of the normalized counts while the red curved line shows the best fitted line. As can be observed from the figure, the distribution of variance cannot be cleanly fit; however, the observed curved line was the best approximation from all of the tried methods.

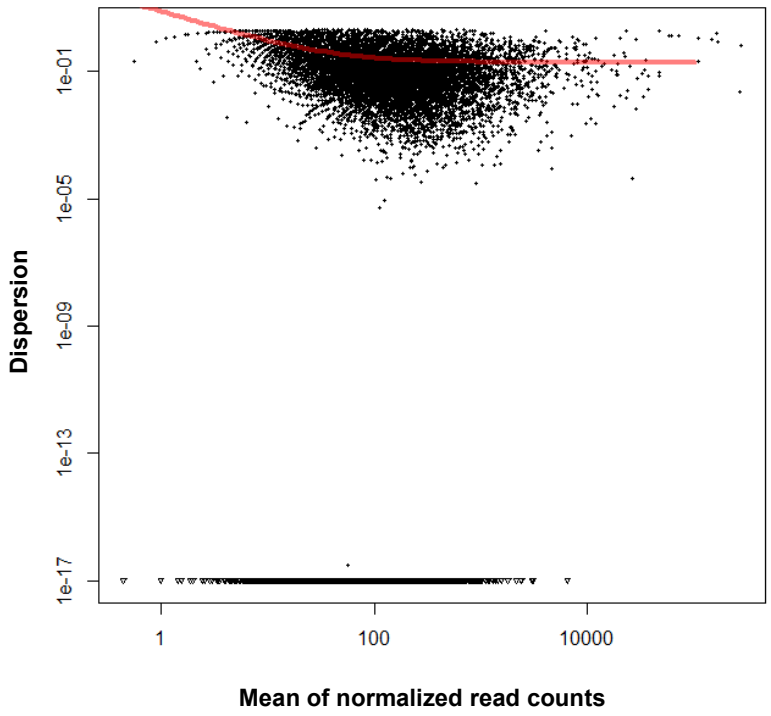


Figure 2.3 RNAseq dispersion plot. Scatterplot of the dispersion for the genes in RNAseq data. The red line indicates the fitted dispersion that is used to estimate the variance for statistical comparison in RNAseq differential analysis.

Identification of differentially regulated genes

After examining the distribution of reads and the estimated variance, I used DEseq to identify differentially regulated genes. A total of 98 significantly differentially regulated genes were identified under the adjusted p-value cutoff of 0.1 using Benjamini-Hochberg method for multiple testing correction. **Table 2.1** shows the list of differentially regulated genes under the FDR cutoff of 0.1 after removing 40% lower quantile of the reads to maximize the statistical power, while maintaining the majority of the differentially regulated genes without the removal. The entire list of 500 differentially regulated genes under the p-value cutoff of 0.05 is available in the Supplementary Material. There are potential BA markers among this list of genes. For example, multiple pro-inflammatory chemokines such as CXCL5, CXCL6, and CXCL10 and acute-phase genes such as SAA1 and SAA2 were identified. Around 65% of differentially regulated genes under the FDR cutoff of 0.1 were upregulated which suggests slight activation of the BA transcriptome. Furthermore, about 20% of genes showed upregulation of 20 fold or higher.

Table 2.1 List of differentially regulated genes. The list of differentially regulated genes under the FDR cutoff of 0.1 after removing 40% lower quantile of the reads from the RNAseq data are shown. “Inf” indicates any number divided by 0.

Gene	Fold change	P-value	FDR
AKR1B10	265.43	1.32E-09	1.17E-05
HBB	116.35	2.32E-09	1.17E-05
MMP7	177.28	3.24E-09	1.17E-05
HBA2	123.41	7.89E-09	2.14E-05
HBA1	193.11	1.01E-08	2.20E-05
SAA1	0.02	3.68E-08	6.47E-05

Table 2.1 List of differentially regulated genes. (Continued)

Gene	Fold change	P-value	FDR
SAA2	0.02	4.18E-08	6.47E-05
KRT17	361.78	8.23E-08	0.000112
LAMC2	205.33	1.13E-07	0.000137
KRT23	101.04	1.71E-07	0.000186
HAO2	0.02	4.29E-07	0.000424
SFRP4	Inf	5.47E-07	0.000494
KRT7	41.53	1.18E-06	0.000987
MUC13	65.07	1.72E-06	0.001335
CNDP1	0.01	2.03E-06	0.001468
TACSTD2	58.87	2.20E-06	0.001488
CLDN4	58.43	3.36E-06	0.002142
STC1	38.30	3.72E-06	0.002244
TESC	89.36	4.55E-06	0.00242
CGA	Inf	4.81E-06	0.00242
FGF23	74.15	4.86E-06	0.00242
PZP	0.03	5.12E-06	0.00242
ANKRD1	195.56	5.13E-06	0.00242
CYP1A2	0.02	6.10E-06	0.002603
CHIT1	Inf	6.15E-06	0.002603
KRT19	34.19	6.24E-06	0.002603
SPINT1	52.15	6.77E-06	0.002718
GREM1	108.37	7.65E-06	0.002963
AVPR1A	0.01	1.10E-05	0.004108
CXCL10	39.11	1.16E-05	0.00419
PRSS22	Inf	1.78E-05	0.006214
RGS4	33.48	2.89E-05	0.009799
IL8	21.21	3.15E-05	0.010198
COMP	Inf	3.20E-05	0.010198
AKR1D1	0.04	3.53E-05	0.010939
GSTM1	0.03	3.79E-05	0.011406
LTBP2	29.77	3.90E-05	0.011445
MOXD1	51.94	4.39E-05	0.012532
CAPG	22.36	5.37E-05	0.014948
TREM2	78.63	5.68E-05	0.015393
FAP	Inf	6.02E-05	0.015879
HP	0.06	6.15E-05	0.015879
CHRNA4	0.02	7.49E-05	0.018598
VTCN1	74.96	7.57E-05	0.018598

Table 2.1 List of differentially regulated genes. (Continued)

Gene	Fold change	P-value	FDR
STC2	47.26	7.72E-05	0.018598
CCL4	30.33	8.43E-05	0.019874
DDX3Y	0.02	8.74E-05	0.020162
BBOX1	0.01	9.33E-05	0.021085
SPP1	15.18	0.00011	0.024409
LUM	16.43	0.000113	0.024409
PDZK1IP1	52.69	0.000118	0.025032
MGP	15.14	0.000122	0.025536
PMEPA1	19.91	0.000128	0.026126
CXCL6	26.81	0.00013	0.026126
UBD	14.66	0.000134	0.026222
DKK1	115.70	0.000135	0.026222
HS3ST2	Inf	0.000147	0.027918
CLIC6	36.83	0.000154	0.028836
CXCL5	111.63	0.000167	0.030673
USH2A	0.01	0.000174	0.031433
DCDC2	23.49	0.000177	0.031433
CFTR	22.38	0.000194	0.033924
ITGBL1	40.13	0.000199	0.03429
FGF19	Inf	0.000225	0.038143
SPINK1	0.08	0.000237	0.039079
HBG2	Inf	0.000238	0.039079
S100A6	13.10	0.000262	0.040567
CLDN7	16.06	0.000263	0.040567
RGS1	14.81	0.000263	0.040567
DBNDD1	24.72	0.000265	0.040567
GPNMB	12.76	0.000266	0.040567
ADCY1	0.05	0.000282	0.042539
OLR1	101.04	0.000295	0.043801
WDR72	0.05	0.00033	0.048391
CCL20	14.85	0.000339	0.049094
IGFALS	0.07	0.000362	0.051684
RAMP1	12.85	0.00037	0.051918
CA12	31.45	0.000373	0.051918
SLCO4C1	0.01	0.000442	0.060488
SGIP1	Inf	0.000446	0.060488
DHODH	0.08	0.000454	0.060776
IGSF9	0.04	0.000474	0.062739

Table 2.1 List of differentially regulated genes. (Continued)

Gene	Fold Change	P-value	FDR
ORM1	0.10	0.000482	0.062936
CRP	0.10	0.000511	0.065509
ACACB	0.09	0.000514	0.065509
PRG4	0.09	0.000519	0.065509
TMEM132A	22.71	0.000566	0.070567
TNFAIP3	11.20	0.000584	0.071973
GNAO1	0.06	0.000608	0.072899
SRD5A2	0.06	0.00061	0.072899
RNASE1	11.41	0.000612	0.072899
GEM	12.81	0.000636	0.074394
GDF15	10.54	0.000638	0.074394
C13orf33	28.03	0.000694	0.080135
PADI1	0.00	0.000716	0.081727
TACSTD1	18.59	0.000751	0.084817
RND2	0.06	0.000783	0.087602
SYT13	Inf	0.000879	0.097239

Enriched inflammation and chemokine signaling pathway in transcriptome

To identify relevant biological pathways associated with these differentially regulated genes, I performed enrichment analysis with Database for Annotation, Visualization and Integrated Discovery (DAVID) online tool. This revealed several over-represented, or enriched, Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways such as the chemokine signaling pathway (p=2.0E-3), the toll-like receptor signaling pathway (p=3.6E-2), and the cytokine-cytokine receptor interaction (p=3.9E-2). It also revealed enriched Gene Ontology (GO) biological processes such as inflammatory response (p=2.9E-5), locomotory behavior (p=4.5E-5), chemotaxis (p=5.5E-5), defense response (p=8.8E-5), and immune response (p=2.7E-4).

Among the enriched biological pathways and processes, the chemokine signaling pathway and inflammatory response were examined due to their biological relevance in the BA pathology. Most of the genes in the chemokine signaling pathway, including chemokine (C-X-C motif) ligand 5 (CXCL5) and IL8, showed significant upregulation (**Figure 2.4a**). By contrast, inflammatory genes showed both up- and downregulation; many pro-inflammatory chemokines including IL8 were upregulated while some acute inflammatory genes, orosomucoid 1 (ORM1), serum amyloid A1 (SAA1), and serum amyloid A2 (SAA2), were downregulated in the BA group (**Figure 2.4b**). Other GO enrichment results are found in **Table 2.2**.

Table 2.2 Enriched Gene Ontology terms. Significantly enriched terms for GO categories are listed.

Gene Ontology categories	Enriched terms
GO: Biological Process	Inflammation, locomotory behavior, chemotaxis, taxis, and defense response
GO: Cellular Components	Extracellular region
GO: Molecular Function	Chemokine activity, chemokine receptor binding, cytokine activity, oxygen binding, polysaccharide binding

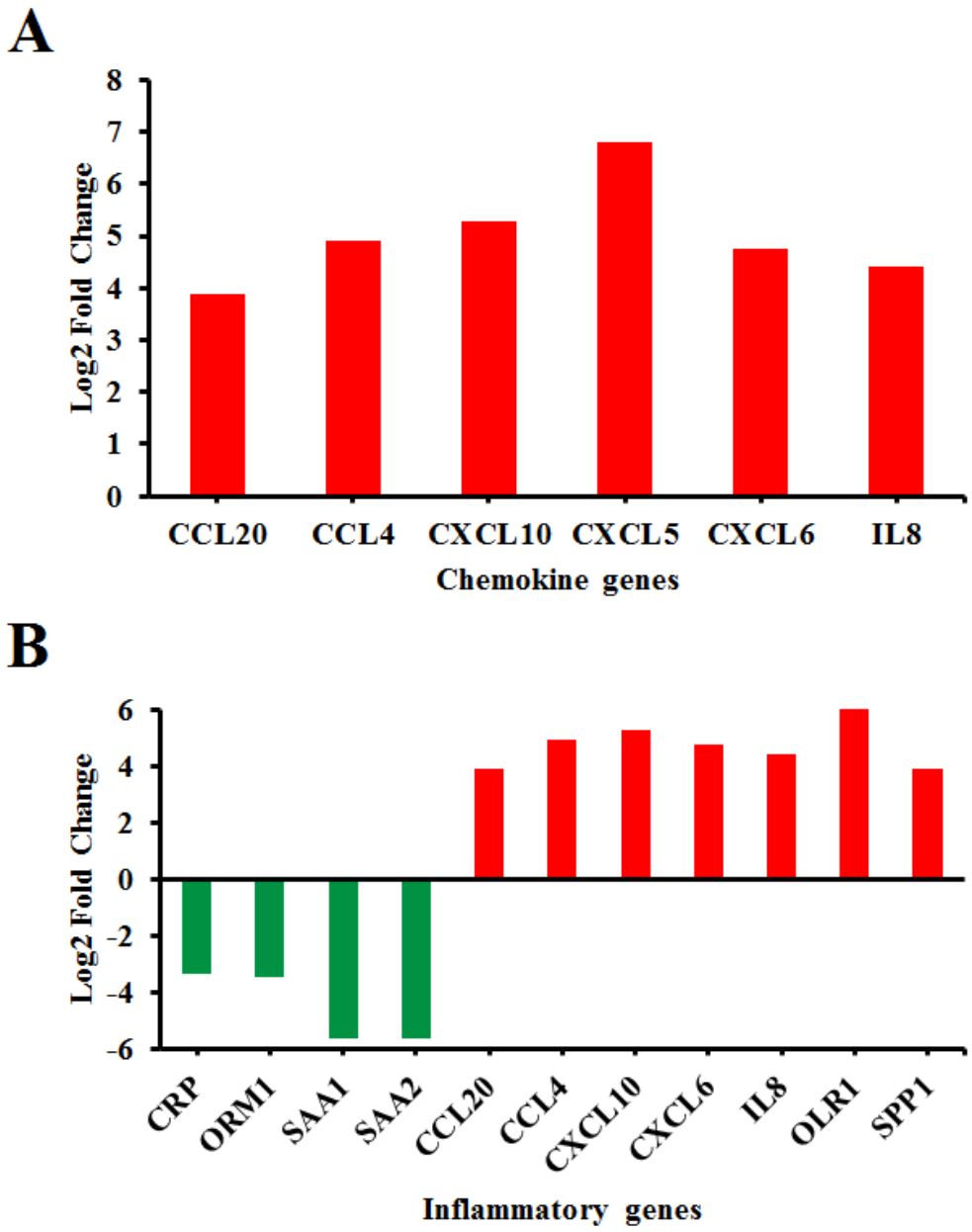


Figure 2.4 Differentially regulated genes in enriched biological categories. (A) Differentially regulated genes in the chemokine signaling pathway (p=2.0E-3) **(B)** Differentially regulated genes in inflammatory response from Gene Ontology: Biological Process terms (p=2.9E-5)
*Red indicates upregulation while green indicates downregulation.
**All genes passed the adjusted p-value cutoff of 0.1 using Benjamini-Hochberg method

Since the lenient p-value cutoff of 0.05 was to be used for the integrative analysis, I also identified enriched biological functions for the larger list of genes. DAVID pathway enrichment results of the 500 genes under the p-value cutoff of 0.05 are shown in **Figure 2.5**. Many metabolic pathways such as drug metabolism, metabolism of xenobiotics, steroid biosynthesis, and retinol metabolism were significantly downregulated in the BA group (**Table 2.3**). From the Panther pathways, cholesterol biosynthesis and glycolysis were the key metabolic pathways. The complement system and coagulation cascades were also enriched with 11 significant genes from the KEGG pathways and 17 from the related Panther pathways (**Table 2.4**). While the complement proteins and fibrinogens were significantly downregulated, plasminogen activators that promote fibrinolysis were upregulated. Overall, differentially regulated genes under the lenient statistical cutoff showed enriched metabolic and complement/coagulation functions in the BA group. Other enriched GO terms can be found in the Supplementary Material.

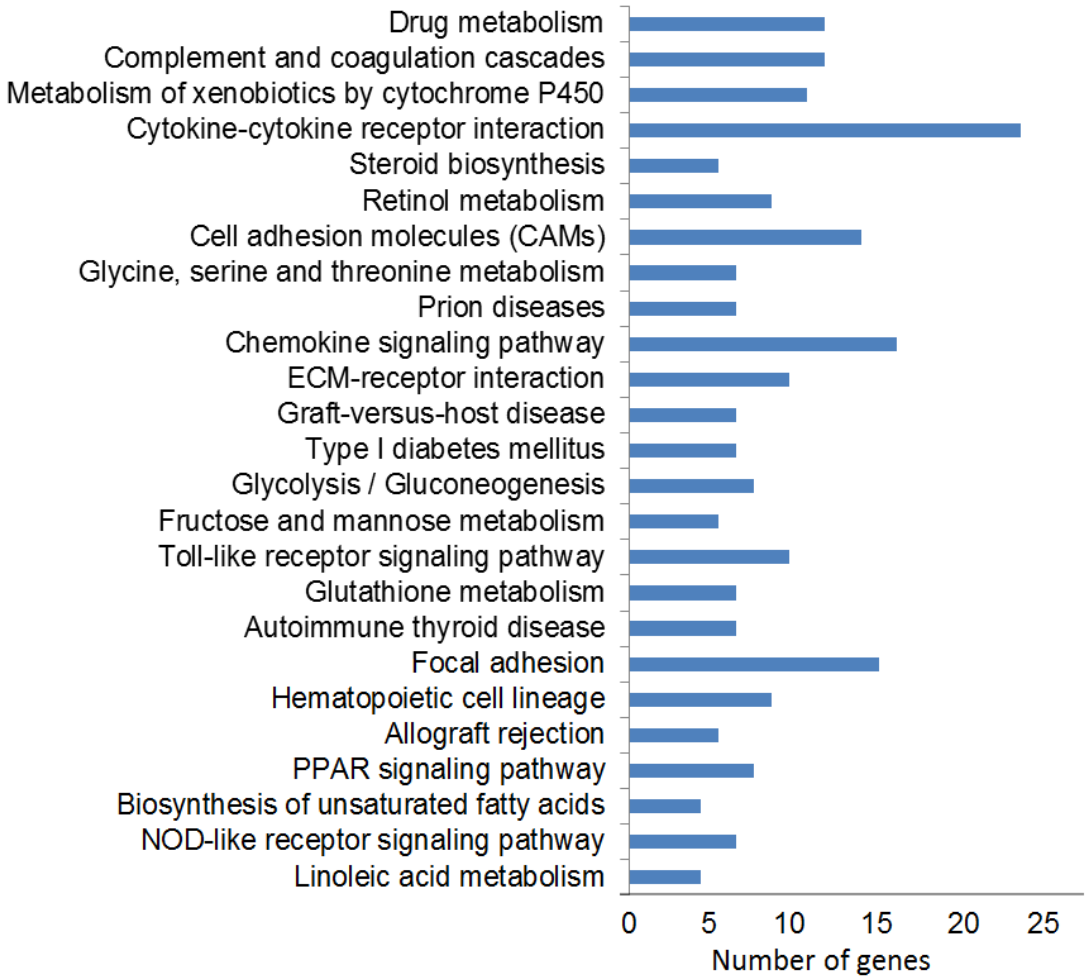


Figure 2.5 Enriched KEGG pathways. Enriched KEGG pathways using the list of differentially regulated genes under the p-value cutoff of 0.05. The enriched pathways are ordered from the highest to the lowest p-value.

Table 2.3 Enriched KEGG pathways. Enriched KEGG pathways with the number of up- and downregulated genes indicating the general direction of regulation for each of the enriched KEGG pathways are shown.

KEGG Pathway	Up/Down
Drug Metabolism	DOWN 2/9
Cytokine-cytokine receptor interaction	UP 21/3
Metabolism of Xenobiotics by cytochrome P450	DOWN 1/9
Chemokine signaling pathway	UP 17/1
Complement and coagulation cascades	DOWN 4/6
Steroid biosynthesis	DOWN 0/5
Retinol metabolism	DOWN 0/8

Table 2.4 Differentially regulated genes in the complement and coagulation cascade. Differentially regulated genes under this enriched KEGG pathway are shown with the fold changes with respect to control.

Gene	Fold change
bradykinin receptor B2	13.51
complement component 4 binding protein, beta	0.24
complement component 6	0.20
complement component 9	0.21
fibrinogen alpha chain	0.20
fibrinogen beta chain	0.16
fibrinogen gamma chain	0.16
plasminogen activator, tissue	9.52
plasminogen activator, urokinase	5.26
plasminogen activator, urokinase receptor	6.41

Differential regulation of MAN1A2

Among many of the differentially regulated genes, MAN1A2 was analyzed in depth because the knockdown of this gene in zebrafish has recently been shown to result in anatomic restriction of bile excretion from the liver, poor intrahepatic bile canaliculi and cilia development, and complete heterotaxy affecting the heart, the liver and the pancreas (manuscript in preparation). Transcriptomic counts for MAN1A2 exons across the BA and the normal group were analyzed in **Table 2.5**. In particular, exon #7 at chr1:118003111-118003234, exon #8 at chr1:118008956-118009049, and exon #10 at chr1:118039385-118039604 showed significant downregulation in the BA group, even after controlling for the total difference of transcript counts between the BA and the normal groups. RTqPCR for these 3 exons confirmed downregulation in the BA transcript count (Supplementary Material). There were also fewer alternatively spliced reads in the BA samples (**Table 2.6**).

Table 2.5 Differential regulation of exons in MAN1A2. Normalized and raw read counts for each of the 13 known exons of MAN1A2 are compared between BA and normal. Highlighted exons in grey indicate significant downregulation in BA.

Exon number	Coordinates (ChrN:start-end)	BA_Normalized Count	Normal_Normalized Count
1	chr1:117910085-117911107	0.01271	0.04692
2	chr1:117944808-117945063	0.01562	0.0625
3	chr1:117948171-117948267	0.04124	0.08247
4	chr1:117957335-117957453	0	0.03361
5	chr1:117963191-117963271	0.02469	0.06173
6	chr1:117984853-117984947	0	0.05263
7	chr1:118003111-118003234	0.01613	0.08871
8	chr1:118008956-118009049	0.01064	0.07447
9	chr1:118035769-118035884	0.01724	0.00862
10	chr1:118039385-118039604	0.00455	0.06818
11	chr1:118042004-118042176	0.02312	0.01734
12	chr1:118045477-118045592	0	0.01724
13	chr1:118065447-118068320	0.01635	0.03097

Table 2.6 Differential alternate splicing in MAN1A2. Normalized and raw read counts for each of the 10 known exons subjected to splicing events are compared between BA and normal.

Coordinates (ChrN:start-end)	BA_Normalized Count	BA_Raw Count	Normal_Normalized Count	Normal_Raw Count
chr1:117911107-117944808	0.01471	1	0.02941	2
chr1:117945063-117948171	0	0	0.01471	1
chr1:117948267-117957335	0.04412	3	0.02941	2
chr1:117957453-117963191	0	0	0.01471	1
chr1:117963271-117984853	0	0	0.01471	1
chr1:117984947-118003111	0	0	0.07353	5
chr1:118009049-118035769	0	0	0.02941	2
chr1:118039604-118042004	0	0	0.02941	2
chr1:118042176-118045477	0	0	0.04412	3
chr1:118045592-118065447	0	0	0.01471	1

Integrative analysis results

I performed integrative analysis using the RNAseq and GWAS data to identify the pairs of differentially regulated genes and their nearby BA-associated SNPs. Of the 500 differentially regulated genes from the RNAseq data under the p-value cutoff of 0.05, 10,144 SNPs from 36 BA family trios that are within ~20kb up- and downstream of each gene were selected as potential BA-associated SNPs. I used the set-based test in PLINK to identify 29 pairs of differentially regulated genes and their associated SNPs that passed the transmission disequilibrium test (TDT) (**Table 2.7**). These SNPs have significantly different allele frequencies in the BA patients and can potentially regulate the expression of nearby genes. Among the 29 genes, complement

component 6 (C6) and lipopolysaccharide binding protein (LBP) are involved in innate immunity while ephrin type-B receptor 2 (EPHB2) and annexin A2 (ANXA2) are associated with hepatic fibrosis.

Table 2.7 Significant pairs of differentially regulated genes and BA-associated SNPs. The pairs of differentially regulated genes and their nearby BA-associated SNPs are identified using the set-based test in PLINK. The p-value cutoff for the genes and the SNPs was 0.05.

Gene	P-value	SNPs
SLCO4C1	0.0015	rs2600834
FGF23	0.0028	rs11063099, rs3812822, rs10437827
LIPG	0.0029	rs11664186
EPHB2*	0.0042	rs6667416, rs4655107, rs10753545, rs4655128, rs12027585
RAMP1	0.0066	rs10185142, rs1584243, rs6729271, rs6738488
GFRA1	0.0092	rs180552, rs180571, rs7087152, rs3901216, rs4751949
GPRC5A	0.0102	rs11055126
PID1	0.0102	rs13034774, rs31276, rs7561470, rs883731, rs6724020
DGAT2*	0.0122	rs1458836
EEF1A2	0.0126	rs11702306
MME	0.0143	rs1025192, rs1816558
ANXA2	0.0144	rs4775260
MGP	0.0158	rs4762785
LBP*	0.0160	rs2232618
AP1M2	0.0214	rs737337
SLC29A4	0.0230	rs6958502
VIM	0.0268	rs243013
C6*	0.0298	rs11743598, rs3805715, rs1801033, rs751138
TESC	0.0328	rs10744888
COL15A1*	0.0330	rs10819542, rs3780622, rs4743322
CD52	0.0332	rs12059495
C3orf25	0.0373	rs3138353
TMEM132A	0.0398	rs3794042
PCSK1N	0.0422	rs4824747, rs2280883
HOPX	0.0441	rs7684910
ANPEP	0.0442	rs1439120
CFTR*	0.0446	rs3808185, rs2237724
SLC28A1	0.0453	rs12438877, rs4247411
PTHR1	0.0483	rs3729704, rs1531136

The previous integrative analysis unfortunately ignores potential BA SNPs that were not associated with any of the differentially expressed genes from the RNAseq data. Therefore, I performed an integrative analysis starting from the TDT of the GWAS data and identified 25,382 BA-associated SNPs associated with BA under the p-value cutoff of 0.05. Although none of the SNPs was found to have the p-value approaching 10^{-7} for 550K SNP due to the low number of available BA family trios, 230 SNPs achieved the p-values of less than 0.001 (Supplementary Material). Annotation on 25,382 genomic variants using Ensembl's Variant Effect Predictor (VEP) (103) tool revealed 53% intronic variants and 15% noncoding transcript variants (**Figure 2.6a**). Among the noncoding transcript variants, 61% were missense mutations (**Figure 2.6b**). Out of 500 differentially regulated genes from the RNAseq data, 190 differentially regulated genes were associated with at least 1 nearby SNP that passed the TDT.

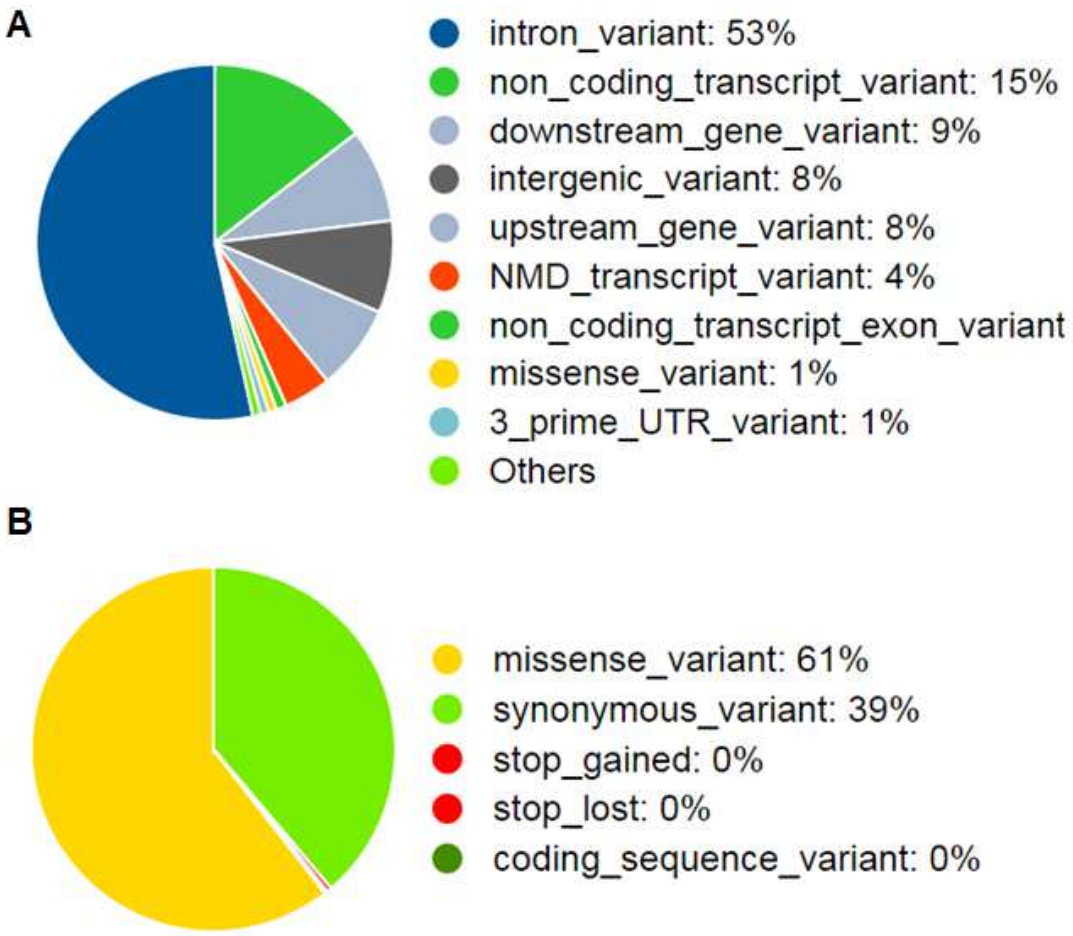


Figure 2.6 Sequence features of significant GWAS variants. (**A**) 25,382 genomic variants using Ensembl's Variant Effect Predictor (VEP) tool are annotated. (**B**) Sequence features of transcript variants in noncoding RNA genes.

The second integrative method used the pre-established eQTL mapping to identify the pairs of significant genes and variants. More specifically, I used eQTL mapping on 25,382 BA-associated SNPs from the TDT analysis to identify their target genes that are also differentially regulated from the RNAseq data (**Table 2.8**). Only 7 BA-associated SNPs identified from the TDT analysis showed potential mapping through eQTL, and the results

showed the skewed distribution for the number of targeted genes. For example, rs1828591 can potentially map to 143 genes, including RAMP1 and SPTBN2 that were differentially regulated, while the rest of the SNPs can map to less than 5 genes.

Table 2.8 eQTL results from the second integrative analysis. Significant variants and their targeted genes are identified.

dSNPs	Number of targeted genes	dGenes
rs1828591	143	RAMP1, SPTBN2
rs198464	4	C11orf9
rs10021418	1	SOD3
rs17567909	3	RXRA
rs3138353	2	C3orf25
rs9462853	1	GNMT
rs11038651	3	SYT13

Functional prediction of unassociated variants

From the second integrative analysis, the vast majority of SNPs discovered from the GWAS data were not “associated” with any of the differentially regulated genes. Therefore, I performed functional prediction on unassociated SNPs using FuncPred to determine which biological functions these SNPs may be involved in. FuncPred result revealed that transcription factor binding site was the most common function (**Table 2.9**).

Table 2.9 Functional prediction of unassociated SNPS. FuncPred results on the majority of SNPs that did not lead to significant transcriptional changes.

Functional category	Number of dSNPs
TFBS	1183
Splicing regulation	300
miRNA binding site	258
nsSNP	263
Stop codon	2

Discussion

The complex transcriptional regulation of inflammatory genes

The transcriptomic data analysis revealed the enriched inflammatory and chemokine signaling pathways in the BA patients. The pro-inflammatory chemokines, IL8 and CXCL5, showed significant transcriptional upregulation, which is consistent with the inflammatory mechanism for the pathogenesis of the 'isolated' form of BA (91). However, a few acute pro-inflammatory genes, SAA1, SAA2 and ORM1, were significantly downregulated. This suggests a more complex regulation of inflammatory genes in the pathogenesis of BA. Recently, the early clinical trial results from administrating corticosteroids, anti-inflammatory agents, to BA patients showed that suppression of inflammatory response may not always improve the condition of a patient (96). This finding, along with the results from this Chapter, demonstrate the need to further investigate the established inflammatory mechanism, which is perhaps

complicated by the intricate balance and timing of its regulation during the various developmental stages of BA.

The roles of fibrosis, immunity, bile acid transfer and lipid metabolism

The integrative analysis using the GWAS and the RNAseq data revealed many pairs of differentially regulated genes and SNPs that are involved in several different biological functions other than inflammation. For example, I identified EPHB2, ANXA2, collagen type XV alpha 1 (COL15A1), and cystic fibrosis transmembrane conductance regulator (CFTR) as fibrotic genes. EPHB2 and ANXA2 are related to hepatic fibrosis that is not only a prominent feature of BA but also a good predictor of outcome in mice following portoenterostomy (104). In addition, COL15A1 is a potential marker for portal fibroblasts with a significant role in biliary fibrosis (105), while CFTR is a well-known gene for cystic fibrosis (106), a disease with similar symptoms as BA.

The integrative analysis revealed a few immune-related genes such as C6 and LBP. C6 is part of the complement system that is known for its involvement in inflammation and immunity against pathogens (107); C6 can help form a membrane attacking complex of the complement system that binds to the surface of bacterial cells and lyse them (108), while LBP helps protect against bacterial infections by facilitating acute-phase immunologic response (109, 110). These immune genes support infection as one of the causes for BA.

Other biological functions, such as bile acid transfer and lipid metabolism, may also play significant roles in BA pathology. Solute carrier organic anion transporter family, member 4C1 (SLCO4C1) is part of the organic anion transporter family that is involved in the membrane transport of bile acid (111). This particular gene was significantly downregulated along with 6 nearby variants. Another significant gene from the transcriptomic profiling, diacylglycerol O-acyltransferase 2 (DGAT2), is involved in the production and accumulation of triglycerides in tissues and has been linked with insulin resistance and diabetes (112). Finally, I discovered a novel missense variant within Niemann-Pick disease, type C1 (NPC1) that is involved in the hedgehog signaling pathway and intracellular cholesterol transfer (113).

The roles of the complement and coagulation cascades

Similar to the results from the liver regeneration study in Chapter 1, the complement system seems to be one of the key regulators of liver-related diseases. For example, multiple complement genes, C4BPB, C6 and C9, and complement-mediated acute phase inflammatory genes, SAA1/2 and ORM1, were downregulated in the BA group. Another important pathway that is closely linked with the complement system is coagulation cascade (114). Coagulation signaling plays a critical role in orchestrating inflammatory response in wound healing and fibrosis (115), which is one of the key features of biliary atresia. From the RNAseq data, fibrinogens were significantly downregulated while plasminogen activators that promote fibrinolysis were

upregulated, which lead to accelerated fibrinolysis and improper blood clotting associated with BA.

The predicted functions of the unassociated SNPs

Since the vast majority of SNPs from the previous analyses were not associated with any differentially regulated genes, I also analyzed the predicted functions of the unassociated SNPs. There are many reasons why significant genomic changes may not lead to transcriptional differences. One, there could be undetected trans eQTL mapping; differentially regulated genes could be regulated by genomic changes that are far away. Another possibility is the activity of transcription factors. In fact, the most commonly predicted function of the unassociated SNPs was transcription factor binding site. It is also possible for SNPs to have non-transcriptional function; for example, SNPs can cause a change protein structure or stability. Last but not least, differential regulation of genes may have occurred during the earlier developmental stage of BA before the RNAseq experiment.

The strengths and the limitations of Chapter 2

Chapter 2 included a thorough investigation of the transcriptomic profiling and the effective integrative analysis to utilize both the RNAseq and the GWAS data. A careful consideration was put into preprocessing of RNAseq analysis as the statistical comparison between the low numbers of pooled samples can result in many false-positives. Through multiple

enrichment analyses including KEGG pathways and GO terms, I identified potential biological functions relevant for BA. The main strength of the integrative method was the identified pairs of significant genes and SNPs that can have a combinatorial effect during the development of BA. Unfortunately, the second integrative analysis using eQTL did not reveal many potential BA markers.

One major limitation of Chapter 2 is the lack of comprehensive understanding of the pathogenesis of BA. Since the focus of this Chapter was on identification of differentially regulated genes and the nearby SNPs that were significantly different in the BA group, I did not analyze the relationships between the significant genes or the SNPs. This issue will be resolved in Chapter 3 where I will reconstruct the comprehensive BA network from multiple data analyses to investigate the relationship between each of the significant genes within the pathogenesis of BA.

Conclusion

In Chapter 2, I performed the RNAseq analysis to identify differentially regulated genes and enriched biological functions and pathways for thorough understanding of the BA transcriptome. The integrative analyses identified the pairs of significant genes and SNPs that can regulate the development of BA together. Furthermore, I analyzed the predicted functions of unassociated SNPs to investigate why certain significant genomic changes did not lead to significant transcriptional regulation. Many of the results from this Chapter will

be integrated into the results of Chapter 3 for derivation of a comprehensive BA network.

Supplementary Materials

Table S2.1 Enriched Panther signaling pathways. Enrichment results on the differentially regulated genes under the FDR cutoff of 0.1. The proportion is the ratio of the number of differentially regulated genes involved in each enriched pathway over the number of total genes in the pathway.

Panther pathways	Count	Proportion	Genes
Plasminogen activating cascade	7	0.151188	PLAT, FGG, FGA, FGB, MMP1, PLAU, PLAUR
Blood coagulation	10	0.215983	PLAT, F2RL3, FGG, SERPINA10, PZP, FGA, FGB, SERPINA1, PLAU, PLAUR
Cholesterol biosynthesis	3	0.064795	HMGCR, HMGCS1, FDPS
Glycolysis	4	0.086393	PKLR, HK2, ENO2, PFKP

Table S2.2 Enriched BIOCARTA signaling pathways. Enrichment results on the differentially regulated genes under the FDR cutoff of 0.1. The proportion is the ratio of the number of differentially regulated genes involved in each enriched pathway over the number of total genes in the pathway.

BIOCARTA pathways	Count	Proportion	Genes
Mechanism of Acetaminophen Activity and Toxicity	3	0.065	PTGS2, CYP2E1, CYP1A2
Regulation of MAP Kinase Pathways Through Dual Specificity Phosphatases	3	0.065	DUSP2, DUSP1, DUSP8
Fibrinolysis Pathway	3	0.065	PLAT, FGA, PLAU
The IGF-1 Receptor and Longevity	3	0.065	IGF1, SOD3, GHR

Table S2.3 Enriched GO:BP. Enriched GO:BP with the number of up- and downregulated genes indicating the general direction of regulation are shown.

GO term	Up/Down
Response to wounding	UP 40/23
Inflammatory response	UP 25/16
Defense response	UP 32/18
Cell-cell signaling	UP 37/9
Chemotaxis	UP 20/2
Response to steroid hormone stimulus	UP 18/6
Lipid biosynthetic process	DOWN 11/20

Table S2.4 Enriched GO:MF. Enriched GO:BP with the number of up- and downregulated genes indicating the general direction of regulation are shown.

GO term	Up/Down
Chemokine activity	UP 14/0
Chemokine receptor binding	UP 14/0
Carbohydrate binding	UP 26/8
Cytokine activity	UP 22/1
Growth factor activity	UP 16/4

Table S2.5 Significant SNPs from the TDT analysis. The list of 230 SNPs under the genomic control-corrected p-values cutoff of 0.001 are shown.

Chromosome	SNP rsID	P-value	GC-corrected P-value
3	rs6777074	9.58E-06	1.41E-05
11	rs10736478	1.95E-05	2.78E-05
4	rs12642830	2.01E-05	2.87E-05
20	rs6078117	2.21E-05	3.15E-05
3	rs2704804	2.21E-05	3.15E-05
1	rs1980	3.61E-05	5.06E-05
2	rs11395	3.96E-05	5.52E-05
11	rs11032492	4.46E-05	6.19E-05
11	rs12789371	4.46E-05	6.19E-05
3	rs4550778	5.31E-05	7.33E-05
7	rs7789165	5.70E-05	7.85E-05
20	rs2327863	7.44E-05	0.000101
3	rs9870687	8.77E-05	0.000119
1	rs7531015	8.77E-05	0.000119
3	rs360460	9.45E-05	0.000128
11	rs7950733	9.62E-05	0.00013
20	rs397020	9.64E-05	0.00013
5	rs6873640	9.64E-05	0.00013
1	rs10900512	0.000101	0.000136
2	rs2367203	0.000101	0.000137
8	rs13279316	0.000124	0.000166
15	rs4965718	0.000124	0.000166
1	rs2065140	0.000124	0.000166
7	rs10268254	0.000124	0.000166
14	rs12895988	0.000124	0.000166

Table S2.5 Significant SNPs from the TDT analysis from 550K array. (Continued)

Chromosome	SNP rsID	P-value	GC-corrected P-value
3	rs360414	0.000156	0.000207
2	rs10175330	0.000156	0.000207
11	rs12789531	0.000157	0.000209
7	rs369982	0.000157	0.000209
13	rs12871141	0.000157	0.000209
13	rs1671966	0.000157	0.000209
7	rs868896	0.000161	0.000214
7	rs757134	0.000161	0.000214
1	rs12751472	0.000161	0.000214
7	rs1639909	0.000161	0.000214
2	rs6433123	0.000162	0.000215
1	rs6702936	0.000162	0.000215
9	rs7032756	0.000162	0.000216
2	rs10932669	0.000162	0.000216
5	rs13155916	0.000162	0.000216
8	rs17464425	0.000183	0.000242
12	rs11574026	0.000183	0.000242
11	rs871346	0.000183	0.000242
7	rs11764590	0.000183	0.000242
15	rs7181587	0.000208	0.000273
15	rs10902583	0.000208	0.000273
13	rs354417	0.000208	0.000273
8	rs10088527	0.000231	0.000302
9	rs2026995	0.000231	0.000302
2	rs12328100	0.000231	0.000302
8	rs4875054	0.000239	0.000312
4	rs10005483	0.000239	0.000312
6	rs4079063	0.000239	0.000312
10	rs2061048	0.000239	0.000312
9	rs1571515	0.000246	0.000321
3	rs1288825	0.000246	0.000321
2	rs6749090	0.000246	0.000321
3	rs4680776	0.000256	0.000334
14	rs1017604	0.000256	0.000334
14	rs1152370	0.000256	0.000334
2	rs13003635	0.000256	0.000334
5	rs2434215	0.000256	0.000334
11	rs498612	0.000256	0.000334

Table S2.5 Significant SNPs from the TDT analysis from 550K array. (Continued)

Chromosome	SNP rsID	P-value	GC-corrected P-value
8	rs10110184	0.000256	0.000334
5	rs11135109	0.000256	0.000334
6	rs6910034	0.000256	0.000334
10	rs10886094	0.000256	0.000334
3	rs2061065	0.000257	0.000335
2	rs12622740	0.000257	0.000335
1	rs12143842	0.000261	0.00034
1	rs12119711	0.000261	0.00034
1	rs2880058	0.000261	0.00034
10	rs180552	0.000261	0.00034
13	rs578196	0.000261	0.00034
6	rs7748185	0.000261	0.00034
20	rs203544	0.000261	0.00034
1	rs1923639	0.000275	0.000358
2	rs7559564	0.000275	0.000358
5	rs11959588	0.000275	0.000358
11	rs1790213	0.000328	0.000425
9	rs6477398	0.000328	0.000425
12	rs224773	0.000347	0.000448
11	rs4757138	0.000347	0.000448
14	rs1187732	0.000347	0.000448
1	rs6657337	0.000347	0.000448
10	rs2025450	0.000359	0.000462
16	rs1816112	0.000386	0.000496
2	rs1124686	0.000386	0.000496
8	rs7002163	0.000386	0.000496
14	rs7144738	0.000393	0.000505
9	rs13295631	0.000393	0.000505
8	rs10503710	0.000393	0.000505
6	rs4946640	0.000407	0.000523
1	rs529989	0.000407	0.000523
17	rs2271921	0.000407	0.000523
8	rs891429	0.000407	0.000523
1	rs6583007	0.000407	0.000523
5	rs13155210	0.000407	0.000523
7	rs701323	0.000407	0.000523
7	rs853052	0.000407	0.000523
20	rs729552	0.000415	0.000533

Table S2.5 Significant SNPs from the TDT analysis from 550K array. (Continued)

Chromosome	SNP rsID	P-value	GC-corrected P-value
3	rs1355760	0.000415	0.000533
8	rs6473190	0.000415	0.000533
17	rs3794730	0.000415	0.000533
1	rs7550692	0.000415	0.000533
20	rs12329577	0.000415	0.000533
1	rs1885644	0.000415	0.000533
18	rs10514034	0.000415	0.000533
2	rs13034774	0.000415	0.000533
14	rs1450688	0.000418	0.000537
3	rs251491	0.000418	0.000537
1	rs6667416	0.000418	0.000537
5	rs10074159	0.000418	0.000537
9	rs10819542	0.000418	0.000537
10	rs1055986	0.000418	0.000537
12	rs4759515	0.000418	0.000537
5	rs325355	0.000418	0.000537
1	rs7540760	0.000418	0.000537
15	rs11858397	0.000465	0.000595
5	rs3097836	0.000465	0.000595
15	rs905436	0.000465	0.000595
3	rs6807750	0.000465	0.000595
2	rs11690506	0.000465	0.000595
11	rs900145	0.000465	0.000595
1	rs4950019	0.000465	0.000595
20	rs1932937	0.000465	0.000595
9	rs170620	0.000465	0.000595
11	rs10832529	0.000465	0.000595
13	rs12021074	0.000465	0.000595
6	rs12202611	0.000465	0.000595
14	rs698334	0.000465	0.000595
1	rs1819548	0.000465	0.000595
4	rs7695738	0.000465	0.000595
12	rs11063099	0.000465	0.000595
7	rs3996329	0.000465	0.000595
14	rs3784115	0.000504	0.000643
8	rs4734507	0.000532	0.000677
10	rs12782822	0.000532	0.000677
7	rs2999574	0.000532	0.000677

Table S2.5 Significant SNPs from the TDT analysis from 550K array. (Continued)

Chromosome	SNP rsID	P-value	GC-corrected P-value
1	rs9436372	0.000532	0.000677
2	rs4670981	0.000532	0.000677
5	rs11950801	0.000556	0.000706
6	rs4711987	0.000556	0.000706
15	rs2413930	0.000556	0.000706
6	rs1668657	0.000556	0.000706
15	rs10519208	0.000556	0.000706
2	rs732278	0.000556	0.000706
17	rs3809790	0.000556	0.000706
2	rs749264	0.000579	0.000734
6	rs551444	0.000579	0.000734
4	rs6448587	0.000579	0.000734
20	rs2235588	0.000579	0.000734
12	rs7307214	0.000579	0.000734
5	rs2883367	0.000579	0.000734
8	rs9642812	0.000604	0.000765
10	rs7085479	0.000604	0.000765
8	rs1033075	0.000604	0.000765
12	rs2470414	0.000604	0.000765
12	rs10746221	0.000604	0.000765
6	rs6457200	0.000604	0.000765
1	rs6659397	0.000604	0.000765
8	rs2347504	0.000604	0.000765
2	rs7422930	0.000604	0.000765
2	rs12714176	0.000604	0.000765
16	rs1861527	0.000644	0.000813
20	rs4810643	0.000644	0.000813
15	rs2041433	0.000644	0.000813
9	rs10761108	0.000644	0.000813
13	rs1890852	0.000644	0.000813
10	rs12771728	0.000644	0.000813
12	rs7980095	0.000644	0.000813
15	rs1033028	0.000644	0.000813
9	rs7025261	0.000647	0.000817
10	rs11192402	0.000647	0.000817
11	rs1618224	0.000647	0.000817
2	rs13024947	0.000647	0.000817
4	rs6854638	0.000647	0.000817

Table S2.5 Significant SNPs from the TDT analysis from 550K array. (Continued)

Chromosome	SNP rsID	P-value	GC-corrected P-value
2	rs10933553	0.000647	0.000817
13	rs9508029	0.000647	0.000817
8	rs4130891	0.000647	0.000817
3	rs6549873	0.000647	0.000817
20	rs11905385	0.000647	0.000817
14	rs698331	0.000647	0.000817
1	rs1632771	0.000647	0.000817
8	rs7003959	0.000647	0.000817
11	rs584368	0.00067	0.000845
2	rs2540975	0.00067	0.000845
1	rs12758848	0.00067	0.000845
12	rs4762298	0.00067	0.000845
13	rs1822970	0.00067	0.000845
4	rs2558133	0.00067	0.000845
7	rs4719495	0.00067	0.000845
20	rs6074272	0.00067	0.000845
3	rs9283639	0.00067	0.000845
1	rs6689318	0.00067	0.000845
5	rs4631137	0.00067	0.000845
1	rs11165761	0.00067	0.000845
13	rs9556365	0.00067	0.000845
9	rs842304	0.00067	0.000845
5	rs2731672	0.00067	0.000845
20	rs1475531	0.00067	0.000845
2	rs6543296	0.00067	0.000845
13	rs626014	0.00067	0.000845
10	rs1773877	0.00067	0.000845
21	rs2831626	0.00067	0.000845
2	rs1601360	0.00067	0.000845
2	rs13031323	0.000674	0.00085
1	rs4655107	0.000674	0.00085
7	rs37089	0.000674	0.00085
6	rs2499663	0.000674	0.00085
2	rs2373423	0.000674	0.00085
9	rs10760289	0.000674	0.00085
16	rs16963728	0.000674	0.00085
2	rs16829095	0.000674	0.00085
21	rs6516823	0.000674	0.00085

Table S2.5 Significant SNPs from the TDT analysis from 550K array. (Continued)

Chromosome	SNP rsID	P-value	GC-corrected P-value
8	rs13260133	0.000772	0.000969
11	rs11215401	0.000772	0.000969
8	rs13282733	0.000772	0.000969
10	rs10508553	0.000789	0.00099
15	rs815093	0.000789	0.00099
13	rs17189299	0.000789	0.00099
1	rs10495334	0.000789	0.00099
2	rs3843330	0.000789	0.00099
3	rs17193050	0.000789	0.00099
8	rs7005442	0.000789	0.00099
15	rs7174078	0.000789	0.00099
2	rs12618749	0.000789	0.00099
9	rs10977530	0.000789	0.00099
9	rs1325116	0.000789	0.00099
12	rs10879474	0.000789	0.00099

Table S2.6 RTqPCR primer sequences for MAN1A2 exons. RTqPCR primer sequences for MAN1A2 exons and GADPH are shown.

Gene (Exon #)	Orientation	Sequence
MAN1A2 Exon6-7	Forward	CACACCTACTGGGATTCCTTGG
	Reverse	GTAGCTGAGGTGGATGAACTCC
MAN1A2 Exon10	Forward	CTCGTGGAGGTCTTACCTTTAT
	Reverse	CTGCTCCTAGTGCAAACATTC
GAPDH	Forward	TCTCCTCTGACTTCAACAGCGACA
	Reverse	CCCTGTTGCTGTAGCCAAATTCGT

Table S2.7 RTqPCR results. Validation results for the two exons of MAN1A2 including their fold changes and the p-values are shown.

Gene	Normal Control (Mean±SE)	BA (Mean±SE)	P-value	Fold change
MAN1A2 exon 6-7	4.8282±0.1597	5.780±0.1873	0.00060	-1.9348
MAN1A2 exon10	4.9679±0.1183	5.9213±0.2016	0.00071	-1.9363

Table S2.8 List of differentially regulated genes from the RNAseq data. The list of differentially regulated genes and their associated raw and adjusted p-values are shown.

Gene	Fold Change	P-value	Adjusted P-value
AKR1B10	265.4259	1.32E-09	1.17E-05
HBB	116.3477	2.32E-09	1.17E-05
MMP7	177.2804	3.24E-09	1.17E-05
HBA2	123.4131	7.89E-09	2.14E-05
HBA1	193.1111	1.01E-08	2.20E-05
SAA1	0.015433	3.68E-08	6.47E-05
SAA2	0.015718	4.18E-08	6.47E-05
KRT17	361.7778	8.23E-08	0.000112
LAMC2	205.3333	1.13E-07	0.000137
KRT23	101.037	1.71E-07	0.000186
HAO2	0.019372	4.29E-07	0.000424
SFRP4	Inf	5.47E-07	0.000494
KRT7	41.53292	1.18E-06	0.000987
MUC13	65.06878	1.72E-06	0.001335
CNDP1	0.009155	2.03E-06	0.001468
TACSTD2	58.87037	2.20E-06	0.001488
CLDN4	58.43386	3.36E-06	0.002142
STC1	38.2963	3.72E-06	0.002244
TESC	89.35802	4.55E-06	0.00242
CGA	Inf	4.81E-06	0.00242
FGF23	74.14815	4.86E-06	0.00242
PZP	0.029777	5.12E-06	0.00242
ANKRD1	195.5556	5.13E-06	0.00242
CYP1A2	0.024493	6.10E-06	0.002603
CHIT1	Inf	6.15E-06	0.002603
KRT19	34.1868	6.24E-06	0.002603
SPINT1	52.14815	6.77E-06	0.002718
GREM1	108.3704	7.65E-06	0.002963
AVPR1A	0.008954	1.10E-05	0.004108
CXCL10	39.11111	1.16E-05	0.00419
PRSS22	Inf	1.78E-05	0.006214
RGS4	33.48148	2.89E-05	0.009799
IL8	21.21281	3.15E-05	0.010198
COMP	Inf	3.20E-05	0.010198
AKR1D1	0.035837	3.53E-05	0.010939
GSTM1	0.030556	3.79E-05	0.011406
LTBP2	29.77208	3.90E-05	0.011445

Table S2.8 List of differentially regulated genes from the RNAseq data. (Continued)

Gene	Fold Change	P-value	Adjusted P-value
MOXD1	51.94444	4.39E-05	0.012532
CAPG	22.36214	5.37E-05	0.014948
TREM2	78.62963	5.68E-05	0.015393
FAP	Inf	6.02E-05	0.015879
HP	0.063091	6.15E-05	0.015879
CHRNA4	0.022142	7.49E-05	0.018598
VTCN1	74.96296	7.57E-05	0.018598
STC2	47.25926	7.72E-05	0.018598
CCL4	30.32922	8.43E-05	0.019874
DDX3Y	0.019874	8.74E-05	0.020162
BBOX1	0.007341	9.33E-05	0.021085
SPP1	15.17662	0.00011	0.024409
LUM	16.43419	0.000113	0.024409
PDZK1IP1	52.69136	0.000118	0.025032
MGP	15.14023	0.000122	0.025536
PMEPA1	19.90982	0.000128	0.026126
CXCL6	26.80741	0.00013	0.026126
UBD	14.65519	0.000134	0.026222
DKK1	115.7037	0.000135	0.026222
HS3ST2	Inf	0.000147	0.027918
CLIC6	36.82963	0.000154	0.028836
CXCL5	111.6296	0.000167	0.030673
USH2A	0.008148	0.000174	0.031433
DCDC2	23.49383	0.000177	0.031433
CFTR	22.37607	0.000194	0.033924
ITGBL1	40.12963	0.000199	0.03429
FGF19	Inf	0.000225	0.038143
SPINK1	0.075957	0.000237	0.039079
HBG2	Inf	0.000238	0.039079
S100A6	13.0963	0.000262	0.040567
CLDN7	16.05974	0.000263	0.040567
RGS1	14.81152	0.000263	0.040567
DBNDD1	24.71605	0.000265	0.040567
GPNMB	12.75607	0.000266	0.040567
ADCY1	0.050007	0.000282	0.042539
OLR1	101.037	0.000295	0.043801
WDR72	0.052569	0.00033	0.048391
CCL20	14.85291	0.000339	0.049094

Table S2.8 List of differentially regulated genes from the RNAseq data. (Continued)

Gene	Fold Change	P-value	Adjusted P-value
IGFALS	0.07047	0.000362	0.051684
RAMP1	12.85342	0.00037	0.051918
CA12	31.45185	0.000373	0.051918
SLCO4C1	0.009586	0.000442	0.060488
SGIP1	Inf	0.000446	0.060488
DHODH	0.084702	0.000454	0.060776
IGSF9	0.044044	0.000474	0.062739
ORM1	0.095941	0.000482	0.062936
CRP	0.097128	0.000511	0.065509
ACACB	0.090268	0.000514	0.065509
PRG4	0.090933	0.000519	0.065509
TMEM132A	22.71296	0.000566	0.070567
TNFAIP3	11.20173	0.000584	0.071973
GNAO1	0.061264	0.000608	0.072899
SRD5A2	0.063342	0.00061	0.072899
RNASE1	11.40741	0.000612	0.072899
GEM	12.80722	0.000636	0.074394
GDF15	10.53987	0.000638	0.074394
C13orf33	28.02963	0.000694	0.080135
PADI1	0	0.000716	0.081727
TACSTD1	18.59259	0.000751	0.084817
RND2	0.059259	0.000783	0.087602
SYT13	Inf	0.000879	0.097239
SERPINA3	0.111911	0.000992	0.107901
DGAT2	0.104755	0.000995	0.107901
ITIH5	19.10288	0.001039	0.11132
CTHRC1	79.85185	0.001047	0.11132
CXCR4	10.70899	0.001084	0.114149
SLC17A2	0.091606	0.001129	0.117677
PTHLH	Inf	0.001139	0.117677
CCL18	25.25926	0.001193	0.121001
COL7A1	0.069941	0.001194	0.121001
F2RL3	28.72222	0.001214	0.121977
HSPA1B	9.506173	0.001269	0.126248
FSTL3	9.833587	0.001336	0.131727
GPRC5A	17.27407	0.001352	0.132071
MMP1	Inf	0.001391	0.134727
NCAM1	17.92593	0.001432	0.137412

Table S2.8 List of differentially regulated genes from the RNAseq data. (Continued)

Gene	Fold Change	P-value	Adjusted P-value
ITGA3	11.97151	0.001462	0.13909
SELE	12.48148	0.001522	0.14352
CCDC80	10.29147	0.001561	0.145939
SEZ6L2	12.57143	0.001579	0.146356
C3orf25	21.45679	0.001626	0.149457
BACE2	10.76035	0.001736	0.158205
PID1	0.107843	0.002012	0.181906
B3GNT5	18.85714	0.002055	0.184006
THY1	9.34127	0.00207	0.184006
OXTR	Inf	0.002103	0.184501
PAPLN	16.5679	0.002109	0.184501
EEF1A2	40.33333	0.00223	0.193187
GSTA2	0.115383	0.002244	0.193187
PDGFA	11.30113	0.002292	0.195796
LRG1	0.135084	0.002417	0.204101
GHR	0.121355	0.002451	0.204101
COL10A1	39.51852	0.002461	0.204101
ANTXR1	9.053498	0.002465	0.204101
FADS1	0.124599	0.002538	0.205976
RXRA	0.133304	0.002541	0.205976
CDH6	18.04233	0.002545	0.205976
HNT	66	0.00268	0.214259
RDH16	0.135031	0.002686	0.214259
SCTR	17.80952	0.002707	0.21436
ADH4	0.137779	0.002759	0.215859
EFEMP1	9.10467	0.002766	0.215859
THBS2	8.062378	0.002857	0.22137
FOXP3	24.03704	0.002919	0.224258
VCAN	8.555556	0.002936	0.224258
THRSP	0.112112	0.002973	0.225477
DEFB1	7.606878	0.003109	0.234205
SLC2A3	7.9702	0.003159	0.236329
SLC30A2	Inf	0.003243	0.240919
LGALS3	7.94709	0.003444	0.254124
CYP2E1	0.147629	0.003489	0.25573
ISLR	9.076132	0.003591	0.261411
NDST1	0.143521	0.003819	0.276137
IHPK3	0.097628	0.003871	0.278086

Table S2.8 List of differentially regulated genes from the RNAseq data. (Continued)

Gene	Fold Change	P-value	Adjusted P-value
KCNE4	15.27778	0.004046	0.286448
FAT	7.071605	0.004051	0.286448
PSAT1	0.129215	0.004074	0.286448
B3GNT3	14.39506	0.004093	0.286448
IL6	9.68	0.004172	0.288367
MME	0.067901	0.004174	0.288367
DDR1	8.698699	0.004223	0.289094
EGLN3	16.17989	0.004238	0.289094
AGXT2	0.138122	0.004327	0.293042
PCSK1N	12.96296	0.00435	0.293042
BICC1	11.35309	0.004451	0.298008
DPEP1	34.62963	0.004548	0.30265
CTSK	9.671498	0.004694	0.310454
FGG	0.1588	0.004736	0.311361
SLC12A2	7.769841	0.004768	0.311566
PDLIM4	18.9037	0.004812	0.312559
BDKRB2	13.76132	0.005021	0.323321
RRAD	12.08642	0.005037	0.323321
ORM2	0.162039	0.005258	0.333983
SDCBP2	9.302469	0.005265	0.333983
KIAA0152	0.15876	0.005462	0.344446
FGB	0.164735	0.005538	0.347201
DTNA	11.74691	0.005719	0.355502
INHBA	8.812071	0.005735	0.355502
CCL3	7.981007	0.005852	0.360693
ADH6	0.159669	0.005925	0.361179
SLC44A3	15.01587	0.005927	0.361179
HAMP	0.167633	0.006128	0.369666
PLAT	9.692008	0.006135	0.369666
ARL4D	0.126627	0.006184	0.369666
PHLDA3	24.44444	0.006203	0.369666
PFKFB3	6.407865	0.006245	0.370176
PAQR5	32.18519	0.006284	0.370436
GATA4	0.148025	0.006362	0.3729
COL16A1	9.777778	0.006411	0.3729
KRT81	Inf	0.006463	0.3729
FA2H	Inf	0.006463	0.3729
MUC20	9.73251	0.006542	0.375457

Table S2.8 List of differentially regulated genes from the RNAseq data. (Continued)

Gene	Fold Change	P-value	Adjusted P-value
INHBC	0.147141	0.006785	0.387365
FLJ45139	0.02859	0.006926	0.391405
IFI6	6.140351	0.006943	0.391405
COL11A1	Inf	0.007	0.391405
WNT10A	Inf	0.007	0.391405
LRRC50	52.96296	0.007159	0.39815
ISG15	6.864198	0.00723	0.39815
HES4	8.888889	0.007231	0.39815
SYT7	0.164011	0.007428	0.406699
ALAS2	17.11111	0.007461	0.406699
LSS	0.165402	0.007717	0.418533
HPX	0.178611	0.007831	0.422596
PPAP2C	10.52991	0.00796	0.426975
TNFRSF12A	6.277092	0.007991	0.426975
BATF	13.73545	0.008711	0.459299
STMN2	50.51852	0.008711	0.459299
ACSS2	0.172115	0.008745	0.459299
EPS8L1	22.54321	0.008765	0.459299
HSPB8	8.613757	0.008817	0.459804
TM4SF1	5.639797	0.009111	0.4717
HTRA3	9.89418	0.009132	0.4717
CFHR3	0.172303	0.00919	0.472315
SPRY1	6.753561	0.009231	0.472315
FADS2	0.177867	0.009442	0.479815
CACNA1H	0.146719	0.009668	0.479815
PFKP	7.71164	0.009716	0.479815
PTGDS	6.773148	0.009735	0.479815
CRYAB	11.24444	0.009745	0.479815
FABP4	11.76955	0.00988	0.479815
IL32	5.387507	0.009944	0.479815
CCDC69	0.164191	0.009947	0.479815
DTX4	0.158804	0.009965	0.479815
HMGCS1	0.182478	0.009984	0.479815
ARRDC2	6.218324	0.010002	0.479815
GPRIN3	13.26984	0.010063	0.479815
IGF2	0.189144	0.010096	0.479815
PLA2G7	7.803419	0.010105	0.479815
PLAUR	6.533333	0.010144	0.479815

Table S2.8 List of differentially regulated genes from the RNAseq data. (Continued)

Gene	Fold Change	P-value	Adjusted P-value
GFPT2	10.66667	0.010166	0.479815
PCOLCE2	0.087019	0.010204	0.479815
SERPINA10	0.186921	0.01026	0.479815
S100A11	5.737189	0.010284	0.479815
AOX1	0.189407	0.010286	0.479815
MMP2	5.956296	0.010307	0.479815
C10orf132	Inf	0.010544	0.488772
LPL	48.07407	0.010642	0.491215
PROM1	8.491228	0.010774	0.495213
HSPA1A	5.626348	0.010846	0.495391
CD52	7.077249	0.01087	0.495391
STARD5	0.135802	0.011132	0.505225
ASPHD1	21.18519	0.011312	0.511234
SERPINE2	7.962963	0.011484	0.516862
SPINT2	6.57284	0.011638	0.519946
IGFBP7	5.15018	0.011648	0.519946
VSTM2L	17.51852	0.011712	0.520668
PRDM1	15.31852	0.011834	0.523918
DKK3	6.330484	0.012055	0.530102
AR	0.126857	0.012071	0.530102
SEMA3C	46.44444	0.01219	0.533164
SOX9	7.604938	0.012398	0.540103
LOC55908	5.55677	0.012534	0.543845
MBNL3	0.137232	0.012874	0.553491
JMJD5	0.166864	0.012916	0.553491
CXCL1	8.854321	0.012918	0.553491
IDH2	0.197401	0.012961	0.553491
STEAP3	0.19625	0.013069	0.553852
KLHL6	11.61111	0.013179	0.553852
SULF1	7.398519	0.013184	0.553852
C15orf52	10	0.013235	0.553852
SLC28A1	0.146813	0.013273	0.553852
MFAP3L	0.159056	0.013276	0.553852
C6	0.20106	0.0134	0.556914
MST150	9.276353	0.013465	0.557466
CFHR4	0.129748	0.013927	0.572611
PDPN	44.81481	0.013989	0.572611
AP1M2	44.81481	0.013989	0.572611

Table S2.8 List of differentially regulated genes from the RNAseq data. (Continued)

Gene	Fold Change	P-value	Adjusted P-value
HIPK2	0.191196	0.014042	0.572614
C13orf15	6.686275	0.014103	0.572923
SLC2A10	0.178987	0.014276	0.577817
PTHR1	0.165093	0.014435	0.578063
FGA	0.207939	0.014447	0.578063
CTGF	5.152137	0.014544	0.578063
CFHR5	0.183258	0.014634	0.578063
HS3ST1	19.82716	0.0147	0.578063
GBA3	0.147785	0.014779	0.578063
ITGAX	5.975309	0.014785	0.578063
SQLE	0.17505	0.014816	0.578063
TMEM156	26.07407	0.014849	0.578063
AQP1	5.28098	0.014871	0.578063
GNMT	0.193873	0.014884	0.578063
HK2	9.703704	0.014922	0.578063
DBN1	6.05291	0.015128	0.582445
ITGA2	10.59259	0.015142	0.582445
C9	0.209931	0.015491	0.593765
STK39	11.87302	0.015746	0.601387
HPR	0.205311	0.015815	0.601915
IGFBP5	4.929485	0.015956	0.605163
ELOVL7	43.18519	0.016085	0.605815
CA9	43.18519	0.016085	0.605815
GFRA1	0.156266	0.016235	0.609354
HSD17B6	0.209708	0.016424	0.612124
VIM	4.708216	0.016495	0.612124
CPN1	0.203704	0.01651	0.612124
JARID1D	0.035427	0.016535	0.612124
CNTFR	0.047009	0.016609	0.612771
CYP2C18	0.181302	0.016675	0.613118
APOD	14.01481	0.016801	0.613609
SMOC2	14.01481	0.016801	0.613609
SLC13A5	0.21356	0.016886	0.614647
KNDC1	0.106996	0.017369	0.630087
GLYATL1	0.204496	0.017448	0.630684
SERINC5	0.158329	0.017501	0.630684
SLCO1B3	0.183054	0.017944	0.644499
S100P	0.174805	0.018102	0.648018

Table S2.8 List of differentially regulated genes from the RNAseq data. (Continued)

Gene	Fold Change	P-value	Adjusted P-value
DUSP5	4.818342	0.018172	0.6484
CXCR7	4.908174	0.01833	0.6519
PRELP	5.317189	0.01848	0.652018
ACP5	4.820031	0.018493	0.652018
DNAJC3	0.176403	0.018536	0.652018
CLDN10	10.04938	0.018574	0.652018
SNTB1	0.202863	0.018675	0.653459
SCD5	10.59259	0.018865	0.657983
PTGS2	6.111111	0.019017	0.661135
ETV4	13.52593	0.019226	0.66465
EPHB2	18.46914	0.01924	0.66465
LSP1	5.229255	0.019579	0.672937
KIAA0746	5.398148	0.019604	0.672937
ANXA13	6.681481	0.019705	0.674268
ABLIM2	40.74074	0.019905	0.674813
ADM	4.8318	0.019922	0.674813
ADRA1A	0.153377	0.020002	0.674813
NEU4	0.17243	0.020029	0.674813
CXCL9	7.022928	0.020065	0.674813
DARC	24.03704	0.020128	0.674813
LIPG	0.19464	0.020157	0.674813
FMOD	5.278583	0.020403	0.680963
MOGAT2	0.124791	0.020534	0.681751
RAB11FIP1	5.137308	0.020552	0.681751
TMEM149	0.187457	0.02081	0.687684
EIF1AY	0.076389	0.020858	0.687684
ATP6V0E2	0.182033	0.020934	0.688108
CLCF1	8.273504	0.02103	0.689148
CRYAA	5.666667	0.021334	0.693382
TXNIP	4.50663	0.02135	0.693382
BHLHB3	39.92593	0.021391	0.693382
C20orf103	23.62963	0.021414	0.693382
CLEC11A	5.764815	0.021643	0.698681
ELOVL6	0.198198	0.021947	0.70641
ANXA3	9.125926	0.022237	0.713612
GALNT2	0.226889	0.022387	0.714634
TBC1D16	0.162037	0.022431	0.714634
DHRS13	0.190811	0.022466	0.714634

Table S2.8 List of differentially regulated genes from the RNAseq data. (Continued)

Gene	Fold Change	P-value	Adjusted P-value
TTC9	6.703704	0.022613	0.717198
TUBB3	11.67901	0.022789	0.718658
ITGB8	23.22222	0.022791	0.718658
LOXL1	9.506173	0.022891	0.719718
CMBL	0.208729	0.023187	0.726909
AEBP1	4.510954	0.023341	0.729631
RGS2	4.760234	0.02363	0.73359
KCNK5	0.148689	0.023677	0.73359
PKLR	0.204651	0.023742	0.73359
KRT80	11.54321	0.023784	0.73359
ULK4	7.550617	0.023957	0.73359
SCRN1	6.140212	0.023977	0.73359
VSIG2	17.38272	0.02399	0.73359
LOC388610	5.748971	0.024009	0.73359
SPTBN2	0.19839	0.02411	0.734604
TTPA	0.188987	0.024716	0.750594
CD7	11.40741	0.024827	0.750594
SOD3	5.65966	0.024842	0.750594
CES1	0.237307	0.025111	0.756615
LYNX1	0.205275	0.025207	0.757394
DACT2	12.54815	0.025322	0.758746
CLGN	0.12908	0.025494	0.760414
CCL21	4.541235	0.025606	0.760414
KLB	0.174908	0.025629	0.760414
PKHD1	5.727669	0.025658	0.760414
SLC1A3	7.834758	0.025751	0.7611
EREG	22.40741	0.025846	0.761814
SH2D3A	12.38519	0.026531	0.779897
GJA1	6.212963	0.027125	0.793789
ASPN	5.139601	0.02715	0.793789
MSC	7.709402	0.027309	0.796295
SOX4	5.104123	0.027488	0.798692
WNT4	22	0.027539	0.798692
HLA-DMB	4.718346	0.027684	0.800756
CD83	5.335097	0.027789	0.801666
TBX15	0.144215	0.028066	0.805879
C8orf80	0.187745	0.028084	0.805879
TAX1BP3	4.809671	0.028293	0.807789

Table S2.8 List of differentially regulated genes from the RNAseq data. (Continued)

Gene	Fold Change	P-value	Adjusted P-value
PLVAP	4.614512	0.028299	0.807789
HLA-DRB1	4.201028	0.028389	0.808228
HBG1	36.66667	0.028679	0.81434
LBP	0.248155	0.029103	0.822358
MRO	0.075446	0.029113	0.822358
C4BPB	0.244622	0.029266	0.823471
RAB25	21.59259	0.029354	0.823471
PMP22	5.01188	0.029402	0.823471
CCL8	10.8642	0.029533	0.823471
GSTP1	4.229599	0.029576	0.823471
CCL5	5.923077	0.029608	0.823471
MOCOS	0.210979	0.029745	0.823905
FAM148B	6.222222	0.029775	0.823905
DUSP8	6.128824	0.0299	0.825261
C15orf48	16.2963	0.030057	0.827487
SLC23A1	0.159144	0.030417	0.834782
CDC42EP5	8.781893	0.030476	0.834782
HLA-DMA	4.547192	0.030587	0.835716
HAS1	9.89418	0.030885	0.841737
FMO2	21.18519	0.0313	0.850908
FLNC	6.824074	0.031672	0.858864
F2RL1	5.354497	0.031995	0.865462
IGF1	0.207462	0.032103	0.866234
ANXA1	4.332745	0.032318	0.869858
MAT1A	0.255882	0.032454	0.870339
CYGB	5.449074	0.032577	0.870339
AKR1C4	0.235867	0.032684	0.870339
HIG2	6.473251	0.032854	0.870339
TSPAN33	0.216786	0.032857	0.870339
LRIG1	0.22335	0.032872	0.870339
CYP51A1	0.235765	0.032897	0.870339
HOPX	13.24074	0.033076	0.872945
PRAGMIN	5.951691	0.033218	0.874436
CPZ	7.537037	0.033294	0.874436
PLIN	0.107843	0.033511	0.878002
GPX2	0.253305	0.033624	0.878845
ST3GAL1	0.227839	0.033978	0.885965
FDPS	0.243228	0.034226	0.886485

Table S2.8 List of differentially regulated genes from the RNAseq data. (Continued)

Gene	Fold Change	P-value	Adjusted P-value
ATF5	0.259578	0.034231	0.886485
YPEL2	0.231041	0.034243	0.886485
DUSP1	3.890112	0.03441	0.888678
C11orf9	4.37648	0.034521	0.889426
C21orf63	5.962963	0.034665	0.891022
GPR125	0.234641	0.034772	0.891656
SLC22A1	0.257278	0.03518	0.900005
EMP1	3.957997	0.03529	0.900679
SLCO2A1	6.111111	0.035454	0.900849
PLAU	5.362007	0.035537	0.900849
CYP26A1	0.058201	0.035604	0.900849
C6orf142	0.141707	0.035629	0.900849
SNAI1	6.925926	0.035949	0.906844
COL6A3	3.997892	0.036214	0.911391
NSDHL	0.228681	0.036601	0.917138
NQO1	7.333333	0.036611	0.917138
AGPAT9	4.835749	0.036992	0.924549
ANXA2	3.875876	0.037905	0.938873
SELENBP1	0.260495	0.037947	0.938873
LARP6	9.312169	0.038001	0.938873
C20orf77	0.24245	0.038077	0.938873
SC5DL	0.249701	0.038108	0.938873
MT1M	0.255982	0.038111	0.938873
PDGFD	6.62716	0.038181	0.938873
CD24	4.139785	0.038328	0.938873
ARL4C	4.15624	0.038344	0.938873
COL15A1	7.822222	0.038464	0.939674
MMP19	4.159098	0.038732	0.944093
ARG2	5.104575	0.038896	0.945979
SEC11C	0.256226	0.039134	0.949638
PHGDH	0.249148	0.039321	0.952053
TMSB10	3.735298	0.039524	0.954821
FCN3	0.26079	0.039643	0.955565
7A5	14.93827	0.040116	0.963123
FOXA1	0.214616	0.040218	0.963123
SULT1C4	9.91358	0.040311	0.963123
GZMB	9.91358	0.040311	0.963123
CHST4	5.703704	0.040471	0.964808

Table S2.8 List of differentially regulated genes from the RNAseq data. (Continued)

Gene	Fold Change	P-value	Adjusted P-value
CYBRD1	4.306878	0.040715	0.968507
NR1D1	4.265203	0.041101	0.974244
HBEGF	4.395062	0.041136	0.974244
HSPA6	9.079365	0.041337	0.976862
MDK	4.242999	0.041619	0.981397
COBLL1	0.251469	0.041711	0.981437
SORD	0.258479	0.041984	0.985252
BHMT	0.271434	0.042058	0.985252
HLA-A	3.686936	0.042158	0.985252
PEG3	0.167532	0.042307	0.985252
UGT1A6	0.132132	0.042353	0.985252
TCEA2	4.089448	0.042418	0.985252
ANPEP	0.273536	0.042662	0.9888
MXRA8	4.306878	0.042772	0.989222
GABRP	19.14815	0.043399	1
DUSP2	6.087146	0.043571	1
NUCB2	0.25697	0.043707	1
ENO2	7.876543	0.044126	1
SLC46A1	0.199761	0.044202	1
AGL	0.233618	0.044286	1
OASL	4.690154	0.044831	1
SLC29A4	0.167756	0.044878	1
CES3	0.189919	0.044922	1
NFKBIE	4.656085	0.045145	1
CCND2	4.820988	0.045179	1
KLF5	5.789474	0.045275	1
ST14	3.998628	0.045568	1
HMGCR	0.252628	0.045691	1
MAMDC4	0.206219	0.046182	1
TMEM97	0.258025	0.046299	1
LMAN1	0.265585	0.046443	1
CD34	4.585702	0.046748	1
PCDH24	0.253669	0.047591	1
SERPINA1	0.286266	0.047934	1
TSPAN8	5.214815	0.048063	1
SEMA3B	5.581481	0.04831	1
SLC16A3	4.155556	0.048421	1
RAB27A	0.24496	0.048449	1

Table S2.8 List of differentially regulated genes from the RNAseq data. (Continued)

Gene	Fold Change	P-value	Adjusted P-value
ETNK2	0.278408	0.048632	1
IGJ	4.653021	0.048823	1
PSKH1	0.236038	0.048955	1
MICAL1	4.617284	0.04912	1
CLDN15	5.748971	0.049287	1
ACHE	8.046296	0.04934	1
LOC388503	0.280709	0.049826	1

Acknowledgements

Chapter 2, in part, is a re-editing of materials published in Mylarappa Ningappa, Jun Min, Brandon W. Higgs, Chethan Ashokkumar, Sarangarajan Ranganathan, and Rakesh Sindhi, Genome-wide association studies in biliary atresia. Wiley interdisciplinary reviews. Systems biology and medicine. 2015.7: 267-273. Chapter 2, in part, is also a re-editing of materials currently being prepared for submission for publication in Juhoon So, Ningappa Mylarappa, Jun Min, Brandon Higgs, Qing Sun, Hakon Hakonarson, Shankar Subramaniam, Donghun Shin and Rakesh Sindhi. The role of MAN1A2 in biliary atresia, **in preparation**. The dissertation author is responsible for analyzing transcriptomic data and performing systems biology methods.

CHAPTER 3: TARGET SEQUENCING, EXOME SEQUENCING, AND NETWORK ANALYSES OF BILIARY ATRESIA

Introduction

The state of a biological system can be studied by examining its essential biopolymers such as DNA, RNA, and proteins. While thorough investigation on any one of these parts, such as RNA high-throughput quantification, can provide functional insights into the system, integration with knowledge on other essential parts is often required to derive the mechanism of the complex system. For example, integration with the GWAS data can identify pairs of differentially regulated genes and nearby BA-associated SNPs, i.e. the method used in Chapter 1. Furthermore, if the system has heterogeneous phenotypes with unknown etiology such as BA, integration with multiple data types could be crucial in unveiling the pathogenesis.

One of the challenges in deriving the mechanism of a rare disease like BA is the statistical power of association resulting from the lack of available patient data. Conventional approaches to GWAS SNP arrays often necessitate p-values of less than 10^{-7} for 500,000 SNPs or 10^{-8} for 1-2 million SNPs from well-established journals (116, 117). This stringent cutoff due to multiple testing corrections requires large numbers of samples to achieve adequate statistical power to detect an association which is often difficult to achieve for a rare disease especially occurring in the pediatric population. For example, **Table 3.1** shows the number of cases and the controls used for

BA GWAS studies (88). While some of the studies have managed to identify a few SNPs with significant p-values, the authors could not identify more statistically significant SNPs for systems biology analyses such as enrichment, pathway, or network analyses.

Table 3.1 Recent biliary atresia GWAS studies. Recent biliary atresia GWAS studies and the number of samples used in each of the study are listed.

Author	Year	GWAS Type	# Cases	# Controls
Levya-Vega	2011	CNV	35	2026
Cui	2013	CNV	61	5088
Garcia-Barcelo	2010	SNP	305	571
Kaewkiattiyot	2011	SNP	124	114
Cheng	2013	SNP	339	401
Tsai	2014	SNP	171	1630
Ningappa	2015	SNP	77	1907

Although multiple GWAS studies analyzed the pathogenesis and etiology of BA, their results were focused on functional analysis of individual variant or its targeted genes of interest. In order to truly understand the complex pathogenesis of BA, we need to improve our systems-level understanding of the disease through integrative analysis. Unfortunately, despite several BA studies with multiple types of data such as RNAseq or GWAS data, no integrative analysis has been performed. Furthermore, no significant progress has been made to derive the comprehensive mechanism of BA.

The clinical and experimental challenges and the existence of multiple forms of BA with different etiologies motivate us to create a comprehensive view of the disease through systems biology. While numerous systems biology approaches to model a disease are available, no conventional network

reconstruction method was appropriate for integrating different high-throughput data to produce highly interpretable results. Therefore, I devised a novel network reconstruction method that can benefit from an effective integration of multiple high-throughput analytic results (**Figure 3.1**).

The goal of Chapter 3 was to understand the comprehensive mechanism of the complex pathogenesis of BA. To this end, I first performed target sequencing analysis on the selected differentially regulated genes from Chapter 2 to discover nearby novel variants. Then, I performed unbiased whole exome sequencing to discover new variants in the coding regions of human genome. Finally, I reconstructed a comprehensive BA network from the significant genes and variants discovered from several analyses in Chapters 2 and 3 to highlight the relevant results and to provide valuable insights into the pathogenesis of BA.

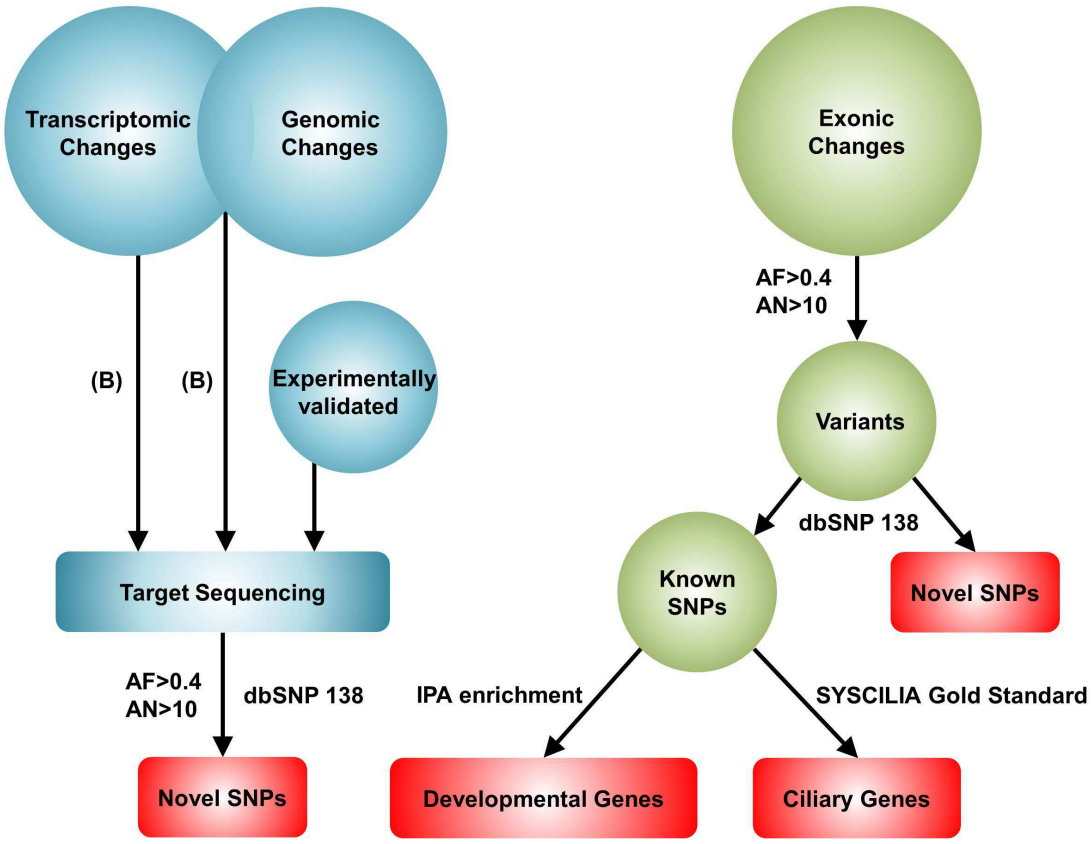


Figure 3.1. Novel systems biology approach for the reconstruction of the BA network. The novel systems biology approach is visualized. First, transcriptomic changes were analyzed from the RNAseq data to identify differentially regulated genes in the BA group. The list of differentially regulated genes were then used together with the list of significant genomic changes from the GWAS analysis to identify pairs of significant genes and SNPs. These results, along with experimentally validated genes, were further explored with target sequencing to identify highly common (AF>0.4 and AN>10) novel SNPs. Exomic changes were also examined with whole exome sequencing. From the whole exome data, the dbSNP 138 database was used to identify novel and known SNPs. Developmental genes mapped from the highly common known SNPs were identified using Ingenuity Pathway Analysis (IPA) enrichment. BA-related ciliary genes were identified with the SYSCILIA gold standard list. Red represents the genes and variants considered for the reconstruction of the BA network.

Methods

Target sequencing

The DNA libraries were prepared from the blood of 43 BA patients using SureSelect kits from Agilent Technologies. A total of 24 genes and their 20kb upstream and downstream sequences were captured using Haloplex custom probes from Agilent Technologies and sequenced in Illumina Hiseq2500. The majority of the genes that were target sequenced were potential BA-related genes derived from the results of RNAseq analysis and integrative analysis and the experimentally validated genes. From the RNAseq results, the genes with the highest fold changes in the enriched chemokine signaling pathway, CXCL5 and CXCL10, and the genes showing significant downregulation instead of upregulation, SAA1, SAA2, and ORM1, from the enriched inflammatory response were selected. IL8 from the RNAseq data was also added to the list for target sequencing because of the recent investigation. From the integrative analysis using the set-based test, the following genes were selected based on their biological annotation and potential contribution to the pathogenesis of BA: EPHB2, vimentin (VIM), GDNF family receptor alpha 1 (GFRA1), DGAT2, FGF23, G protein-coupled receptor, class C, group 5, member A (GPRC5A), ANXA2, alanyl aminopeptidase (ANPEP), lipase, endothelial (LIPG), receptor (G protein-coupled) activity modifying protein 1 (RAMP1), LBP, C6, SLCO4C1, CFTR, and COL15A1. Lastly, ARF6, mannosidase, alpha, class 1A, member 2

(MAN1A2), hypoxia inducible factor 1, alpha subunit (HIF1A), and hypoxia inducible factor 1, alpha subunit inhibitor (HIF1AN) were selected based on the published literature and the previous GWAS results (118).

The alignment for the target sequencing data was performed using Burrows-Wheeler Aligner (BWA) against the hg19 human reference genome (119). Genome Analysis Toolkit (GATK) v3.3-0 was used for local realignment and variant calling with HaplotypeCaller (120). The hard filters, QualByDepth (QD) < 2.0, FisherStrand (FS) > 60, RMSMappingQuality < 40.0, HaplotypeScore > 13.0, MappingQualityRankSumTest < -12.5, and ReadPosRankSumTest < -8, were used to remove potential false-positive SNPs. For indels, similar filtering was applied. The filtered variants were then annotated using the dbSNP 138 database and the VCF annotation v1.0 from SnpEff (121). The variants were considered to be novel if they were not found in the dbSNP 138 database. The additional filters of allele frequency (AF) >0.4 and total number of called alleles (AN) >10 were applied to the annotated variants. The quality and the coverage of the sequencing data were checked with FastQC (122), Integrative Genomics Viewer (IGV) (123), and CalculateHsMetrics from Picard tools (124).

Linkage disequilibrium analysis

Linkage disequilibrium (LD) analysis was performed using Haploview (125) by comparing the LD blocks for the exons in MAN1A2 between the target sequencing data of BA patients and the Phase 1, 1000 genome data of

CEU population as control (126). The total of 99 genotypes were used from the CEU population of 1000 genome data for calculating LD blocks. The default coloring scheme of Haploview was used to display the LD blocks in MAN1A2. Log of the likelihood odds ratio, known as LOD, and D′ were calculated for each pairs of SNPs in the exons of MAN1A2.

Whole exome sequencing analysis

The exome libraries were prepared from the blood DNA of 54 BA patients using SureSelect XT Human All Exon V5 Library from Agilent Technologies and then sequenced in Illumina Hiseq2500. Similar to the target sequencing data analysis, BWA was used for alignment and GATK for variant calling and filtration. Potential PCR duplicates were identified and removed with Picard. Base quality score recalibration from GATK was applied to the whole exome data. The quality and the coverage of the exome sequencing data were checked with FastQC, IGV, and CalculateHsMetrics from Picard. The highly common variants passing the filters of AF>0.4, AN>10, as well as having “moderate” or “high” variant effect from the VCF annotation were selected for further analyses. The “moderate” and “high” variant effects include variant functional categories such as missense, frameshift, splice acceptor, and stop-gained. The novel variants from the filtered list were annotated with the dbNSFP database for functional prediction of missense variants (127). In particular, SIFT, Polyphen2, and LRT from the dbNSFP database were used.

The internal list of the previously identified BA-associated SNPs was derived to help identify SNPs from the whole exome data that show significantly different genomic changes with respect to control. The list was derived from the results of several GWAS studies performed at Children's Hospital of Philadelphia and Children's Hospital of Pittsburgh. The total numbers of BA patients and normal patients across the multiple GWAS datasets were 74 and 1617, respectively. The common SNPs in both the internal list of GWAS SNPs and the filtered variants from the whole exome data were selected for enrichment analysis using Ingenuity Pathway Analysis (IPA) (128) and identification of genes involved in ciliary development and function using SYSCILIA gold standard (SCGSv1) (129).

Whole exome network reconstruction

A custom human interaction network was created in Cytoscape (44), a network visualization tool, by integrating protein-protein interaction data from BIOGRID (45), transcription factor interaction from TRANSFAC (48), and pathway information from KEGG pathways. The topological features of this custom human interaction network, nodes and edges, were used to create the whole exome network for BA, along with the first neighbors of the genes mapped from the highly common variants from the whole exome data. The first neighbors are genes that interact directly with the target genes in a network with no intermediate genes between them. Then, BINGO Cytoscape plugin (130) with default parameters was used to identify over-represented

biological processes to create a GO network. The relevant biological processes that passed Benjamini-Hochberg-corrected p-value of 0.05 were included in the network.

Biliary atresia network reconstruction

The list of significant genes was derived from the combined results of high-throughput analyses: (i) all of the genes from the RNAseq and integrative analyses that were target sequenced, (ii) the genes mapped from the highly common novel SNPs (AF>0.4 and AN>10) from the whole exome data, (iii) BA-related ciliary genes mapped from the highly common known SNPs from the whole exome data, (iv) and the genes in the IPA-enriched biological categories from the whole exome data. The final list of significant genes was as follows: USP6, CXCL5, TNC, C6, VIM, KLRK1, ARF6, ANPEP, GPRC5A, CXCL10, EPHB2, CD44, HIF1AN, SAA1, SAA2, LBP, RAMP1, SLCO4C1, COL15A1, FGF23, MUC6, ANXA2, ORM1, NPC1, HIF1A, DGAT2, LIPG, NEUROD1, HHLA2, GFRA1, IGFBP1, IL8, MAN1A2, HTT, INVS, FSHD region gene 1 family member B (FRG1B), and T cell receptor beta constant 2 (TRBC2).

The initial BA network was derived to include the first and the second neighbors of these significant genes within the custom human interaction network. The second neighbors are genes that interact indirectly with the target genes with only one intermediate gene between them. Then, a smaller network was reconstructed to include as many significant genes that are within

the second neighbors of each other as possible. In other words, the interaction among the significant genes was the main criteria for inclusion into the BA network.

A minimalistic approach was used to condense the BA network by excluding as many neighbor genes and their interactions as possible, while maintaining all of the significant genes as essential information. First, only one neighbor gene out of many potential neighbor genes was selected to connect the significant genes. Second, certain neighbor genes that were present in the initial BA network multiple times were prioritized over those that were present fewer times. Lastly, the interactions between the neighbor genes were ignored. However, certain exceptions were made from this minimalistic approach; MAN1A2 was connected in the network with the third neighbor interaction while a few genes, HTT and VIM, were ignored due to having too many protein-protein interactions in the custom human interaction network.

After the reconstruction of the BA network, each of the significant genes was annotated to identify common biological functions in the network. Furthermore, the common transcription factors that could potentially regulate the genes in the network were identified using DAVID's UCSC_TFBS enrichment (40). EASE score, a modified Fisher exact p-value for DAVID enrichment, of 0.05 was used for the statistical cutoff.

Results

Sequencing metrics

Table 3.2 shows the average metrics for the target and whole exome sequencing data. While the sequencing depth was enough to retain all of the samples for target sequencing experiment, 5 whole exome samples were removed from further analysis after using the cutoff of >80% of bases showing >30x coverage.

Table 3.2 Average alignment metrics for target and whole exome sequencing. The average alignment metrics for target and whole exome sequencing are shown.

Metrics	Target	Whole exome
Genome size	3,137,161,264	3,137,161,264
Total reads	17,169,148.42	94,439,470.59
Unique reads aligned	14,124,285.56	81,657,949.07
Unique reads percent mapping	82%	96%
Mean target coverage	460.60	118.19
Percentage of target bases with 30X coverage	85%	89%
Percentage of target bases with 40X coverage	83%	84%
Percentage of target bases with 50X coverage	81%	78%
Percentage of target bases with 100X coverage	71%	49%

Novel SNPs from the target sequencing data

To identify novel variants around significant genes from the previous analyses, I sequenced the target regions of 24 selected genes from 43 BA patients. Most of these genes were selected based on the results from the RNAseq and integrative analyses. The final metrics for the target sequenced samples were 83.1% of bases showing at least 40x coverage, which is enough to detect a significant number of variants. Most of the targeted genes were

associated with at least one novel SNP passing the stringent allele frequency (AF) cutoff of 0.4 and the total number of alleles in called genotypes (AN) of greater than 10 (**Table 3.3**). This stringent AF cutoff was applied to increase the confidence on potential SNP biomarkers and their target genes. All of the highly common novel SNPs from the target sequencing data were noncoding that may have indirect effects on the transcriptome. There was also one nonsense mutation with lower allele frequency than 0.4 in ANXA2 that can be associated with hepatic fibrosis (131).

Table 3.3 Novel SNPs from target sequencing. Novel SNPs were identified from target sequencing. *SNPs were in close proximity to both SAA1 and SAA2.

Gene	Novel SNPs (AF>0.4)		
	Genomic location	**Allele frequency**	**Allele change**
HIF1AN	chr10:102371717	0.558	A > G
	chr10:102381938:	0.430	G > C
	chr10:102335587	0.419	C > A
	chr10:102335588	0.419	C > T
	chr10:102371719	0.452	A > G
	chr10:102288649	0.417	T > A
FGF23	chr12:4445620	0.465	A > G
	chr12:4445618	0.547	A > G
EPHB2	chr01:23245302	0.488	G > A
	chr01:23224633	0.500	G > A
COL15A1	chr09:101791486	0.477	A > G
	chr09:101838582	0.477	T > A
	chr09:101802438	0.474	G > A
SLCO4C1	chr05:101598264	0.453	A >C
	chr05:101605510	0.442	T > C
	chr05:101605514	0.442	G > A
	chr05:101605515	0.442	T > G
	chr05:101627933	0.462	T > C
DGAT2	chr11:75511683	0.476	A > G
	chr11:75510671	0.405	A > T
MAN1A2	chr01:117991089	0.419	T > C
	chr01:117991086	0.571	T > G
ARF6	chr14:50369743	0.462	G > T
	chr14:50369741	0.480	C > T
SAA1/2*	chr11:18306706	0.479	G > A
	chr11:18275399	0.465	G > A
	chr11:18244418	0.659	T > G

Table 3.3 Novel SNPs from target sequencing (Continued)

Gene	Novel SNPs (AF>0.4)		
	Genomic location	**Allele frequency**	**Allele change**
CXCL5	chr04:74837722	0.500	A > G
	chr04:74837726	0.500	C > G
	chr04:74853963	0.453	A > C
CXCL9	chr04:76928543	0.452	C > A
IL8	chr04:74617686	0.541	T > G
	chr04:74612590	0.444	T > A
	chr04:74612589	0.522	C > G
RAMP1	chr02:238760567	0.429	G > T
ANPEP	chr15:90311139	0.460	T > A
GFRA1	chr10:118013852	0.429	G > A
	chr10:118013856	0.429	G > A
	chr10:118013848	0.500	G > A
	chr10:118013877	0.500	G > A
LIPG	chr18:47090001	0.682	T > A
LBP	chr20:36950014	0.574	T > A

Linkage disequilibrium analysis of SNPs in MAN1A2 exons

To complement the results from the differential analysis of exons in MAN1A2 from Chapter 2, I have performed linkage disequilibrium (LD) analysis on the known exons of MAN1A2. While most of the SNPs within the exons displayed the same LD pattern in the BA and the 1000G data, one particular SNP, rs7514323, showed different LD relationships with the other SNPs within exon #7 (**Figure 3.2**).

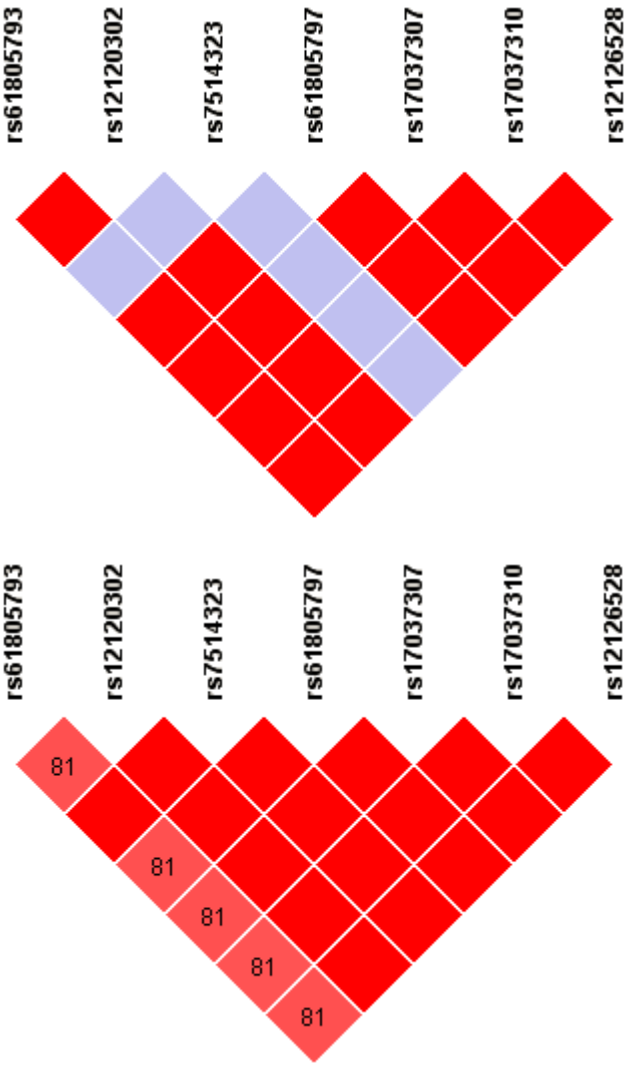


Figure 3.2 Linkage disequilibrium analysis of exon #7 of MAN1A2. LD measurements were calculated for the SNPs in the exon #7 located at chr1:118003111-118003234. Red blocks have D′ (normalized linkage disequilibrium measure or D) of 1.0 and logarithm of odds (LOD) score of greater than or equal to 2.0. Blue blocks have D′ of 1.0 but LOD score of less than 2.0, not passing the LD cutoff.

Novel missense variants from the whole exome data

In addition to the targeted sequencing approach to discover new variants for the selected genes, I utilized the whole exome sequencing approach to discover novel variants from the unbiased exome of 54 BA patients. The whole exome sequencing analysis was also necessary because

of the lack of coding variants discovered from the target sequencing. The final metrics cutoff for the whole exome sequencing was at least 80% of bases showing 30 times or higher coverage, which removed 5 BA samples from further analyses. 51 novel variants were highly common (AF>0.4 and AN>10) among the BA patients and can have potentially significant effect on their target genes according to the variant call format (VCF) annotation; 8 of these variants were missense mutations for 6 different genes (**Table 3.4**). Among the proteins encoded by these genes, ubiquitin carboxyl-terminal hydrolase 6 (USP6) can regulate plasma membrane localization of ADP-ribosylation factor 6 (ARF6), whose role in BA has been recently investigated (132), while mucin 6, oligomeric mucus/gel-forming (MUC6) is involved in reactive biliary epithelium in viral hepatitis (133).

Table 3.4 Novel missense SNPs from whole exome. Novel missense SNPs were identified from the whole exome data. *Genes to be included in the final BA network. **USP6 (Damaging from SIFT and LRT) and MUC6 (Damaging from Polyphen2)

Gene	Novel missense SNP (AF>0.4)			
	Genomic location	**Allele frequency**	**Allele change**	**dbNSFP results**
NPC1*	chr18:21124945	0.653	C > G	Tolerated
TRBC2	chr7:142498833	0.500	A > T	NA
FRG1B	chr20:29632662	0.490	G > T	Tolerated
TRBV6-7	chr7:142143906	0.459	G > C	NA
	chr7:142143907	0.459	T > C	NA
	chr7:142144059	0.449	A > C	NA
USP6*	chr17:5036210	0.459	T > G	Damaging**
MUC6	chr11:1017135	0.449	G > A	Damaging**

Significant developmental genes from the whole exome data

I analyzed highly common known variants from the whole exome data to discover significant developmental genes and pathways. Since the whole

exome of the normal group was not sequenced, and using other published whole exome data as control can significantly increase the error rate due to differences in experimental and bioinformatic methods, I derived the internal list of known BA-associated SNPs from the set of published GWAS data. This allowed me to identify common exonic variants that have also shown statistical difference with respect to control from the GWAS data. After mapping these common variants into genes, I performed enrichment analysis using Ingenuity Pathway Analysis (IPA) to identify over-represented biological processes.

The IPA enrichment analysis on 262 genes mapped from the common variants in the internal list and the whole exome data revealed significantly enriched embryonic development (p=1.94E-2) and hematological system development and function (p=3.10E-2). The genes in embryonic development were neuronal differentiation 1 (NEUROD1), huntingtin (HTT), and insulin-like growth factor binding protein 1 (IGFBP1), while the genes in hematological system development and function were HERV-H LTR-associating 2 (HHLA2), CD44, tenascin C (TNC), and killer cell lectin-like receptor subfamily K, member 1 (KLRK1). A similar result was found when the enrichment was performed on the first neighbors, genes directly interacting with 262 genes, in the custom human interaction network derived from protein-protein and transcription factor interactions.

Another biological process that may be relevant for BA is ciliary development and function. 11 of 262 common genes were involved in ciliary development and function according to the SYSCILIA gold standard list; these

genes were doublecortin domain containing 2 (DCDC2), dynein, axonemal, heavy chain 11 (DNAH11), HTT, intraflagellar transport 88 (IFT88), inversin (INVS), MDM1, pericentriolar material 1 (PCM1), SAS-6 centriolar assembly protein (SASS6), SCL/TAL1 interrupting locus (STIL), spectrin repeat containing, nuclear envelope 2 (SYNE2), and WD repeat domain 35 (WDR35).

Whole exome network

The whole exome data from the BA patients revealed a large number of variants that were excluded when the stringent AF cutoff of 0.4 was used. Therefore, I also used the AF cutoff of 0.2 to analyze the larger list of variants and to identify over-represented GO terms among the first neighbors of the mapped genes in the custom human interaction network (**Figure 3.3**). I also used BINGO Cytoscape plugin to create a network of enriched GO terms that may be relevant to BA pathology. Such identified GO terms were acute inflammatory response (p=2.12E-03), vasculogenesis (p=6.84E-03), embryonic development (p=2.59E-05), liver development (p=4.19E-07) regulation of immune response (p=1.61E-03), epidermal growth factor receptor (EGFR) signaling pathway (p=4.11E-03), and positive regulation of transforming growth factor beta (TGFβ) receptor signaling pathway (p=5.92E-03). Among the enriched pathways, EGFR signaling can affect the development of BA with ARF6 (118), while TGFβ receptor signaling can regulate fibrosis (134).

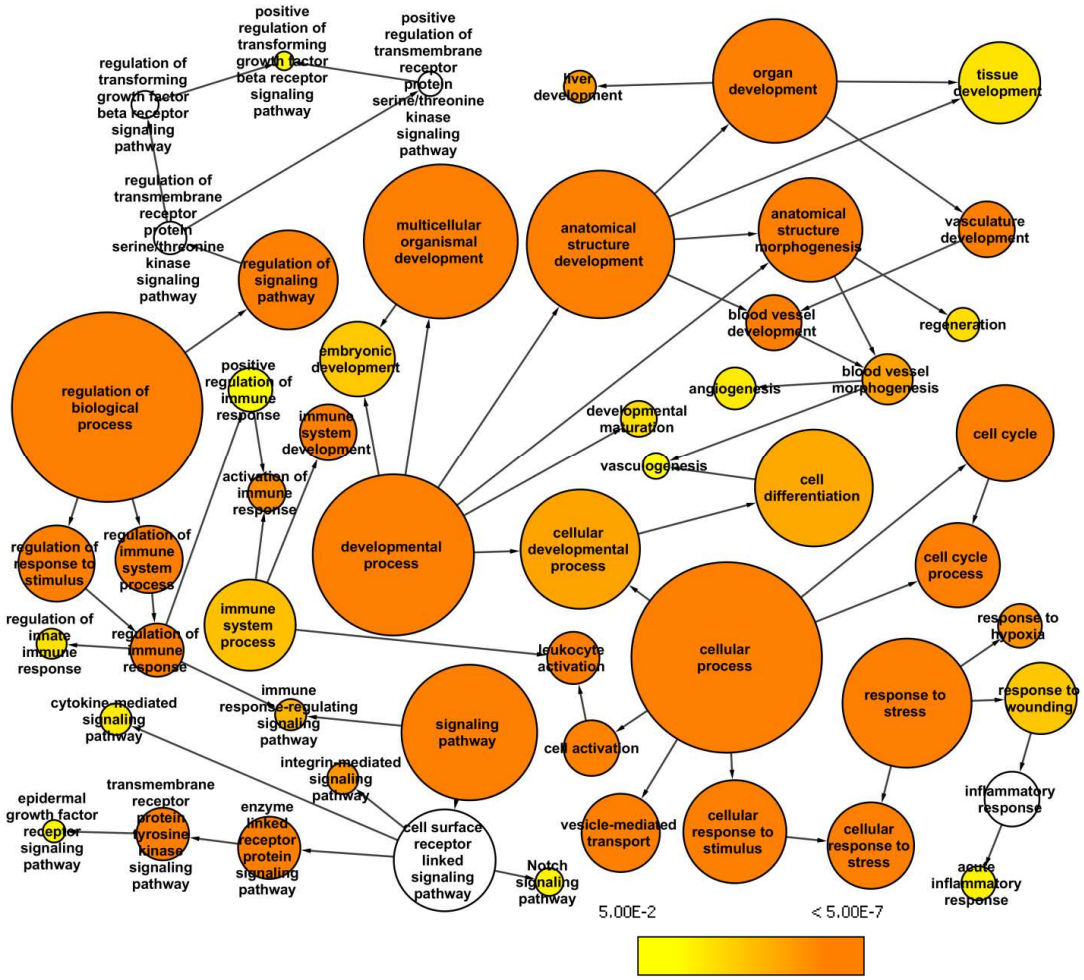


Figure 3.3 Whole exome network. This network was created in Cytoscape with the BINGO plugin to visualize over-represented Gene Ontology: Biological Processes among the genes, which included those mapped from the common variants from the whole exome data (AF>0.2 and AN>10) and their first neighbor genes within the custom human interaction network. The size of a node represents the number of genes annotated with the biological process while the color indicates different p-values for the significance of enrichment.

The comprehensive biliary atresia network

To effectively integrate the key results from different high-throughput data analyses, I utilized a novel systems biology approach to reconstruct a comprehensive BA network. Instead of deriving a large, complex network directly from the integrated genomic and transcriptomic data and applying a set of computational algorithms, I started with the list of significant BA genes

and variants identified from different high-throughput data analyses (**Figure 1**). Using the custom human interaction network, I then derived the initial BA network by integrating the first and the second neighbors (0 or 1 intermediate genes) of the significant genes. These neighbor genes show a close interaction with the significant genes at the protein level and would likely be affected during the pathogenesis of BA. Next, I selected the significant genes that have first or second interaction with other significant genes in the network. As such, the interaction among the significant genes was the main criteria for inclusion into the BA network under the assumption that the main components of a disease network would have close interaction with one another. After reconstructing the initial BA network, I used the minimalistic approach to reduce the number of neighbor genes to provide a highly condensed and interpretable network.

The proposed BA network presents a comprehensive mechanistic view of the complex pathogenesis of BA (**Figure 3.4**). It highlights the sources and the interactions between the significant genes. The network has genes related to key biological functions of BA such as fibrosis, inflammation, immunity, and development (**Figure 3.5**). Through this network, we can understand the interaction between the signiciant biomarkers and also between different known biological functions for BA. Since the size of a node is proportional to the number of interacting genes, the bigger it appears in the network, the more significant impact the gene can have.

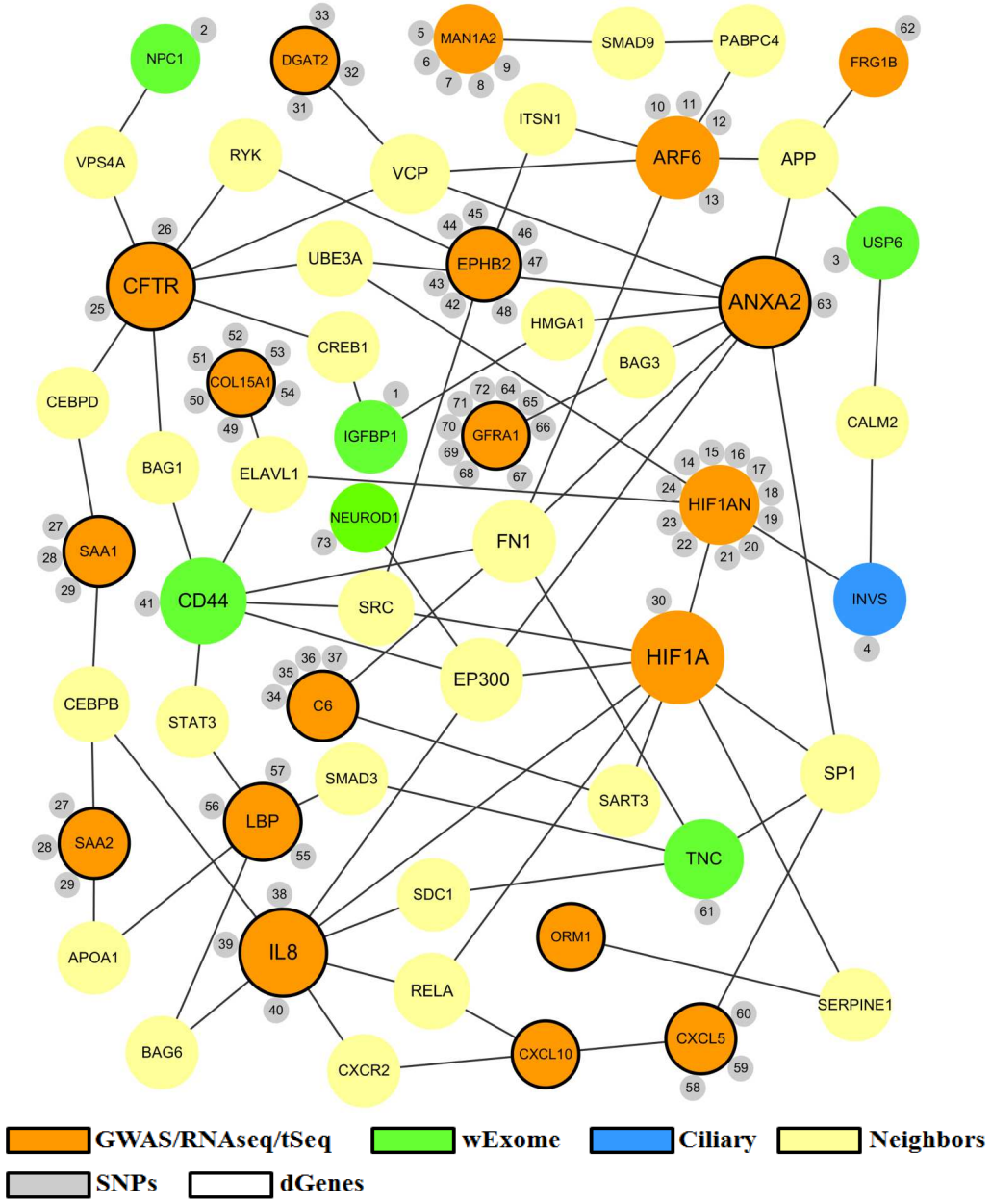


Figure 3.4 Proposed biliary atresia network. This network was created in Cytoscape by using the second neighbor interactions of the significant genes within the custom human interaction network. Interactions with other signficant genes was the key criteria for selection into the final BA network. The size of a node depends on the connectivity within the network; the larger the node, the more connected it is to other genes. The orange nodes represent the significant genes derived from the GWAS, RNAseq, or target sequencing data while the light green nodes represent the significant genes from the whole exome data. One blue ciliary gene is also from the whole exome data. The yellow nodes represent the neighbor genes that link different significant genes through protein-protein interaction. The small grey nodes represent the SNPs that are associated with their attached genes. The entire list of SNPs, both novel and known, can be found in the Supplementary Material. The black circular edges around the nodes represent differentially regulated genes ($p<0.05$).

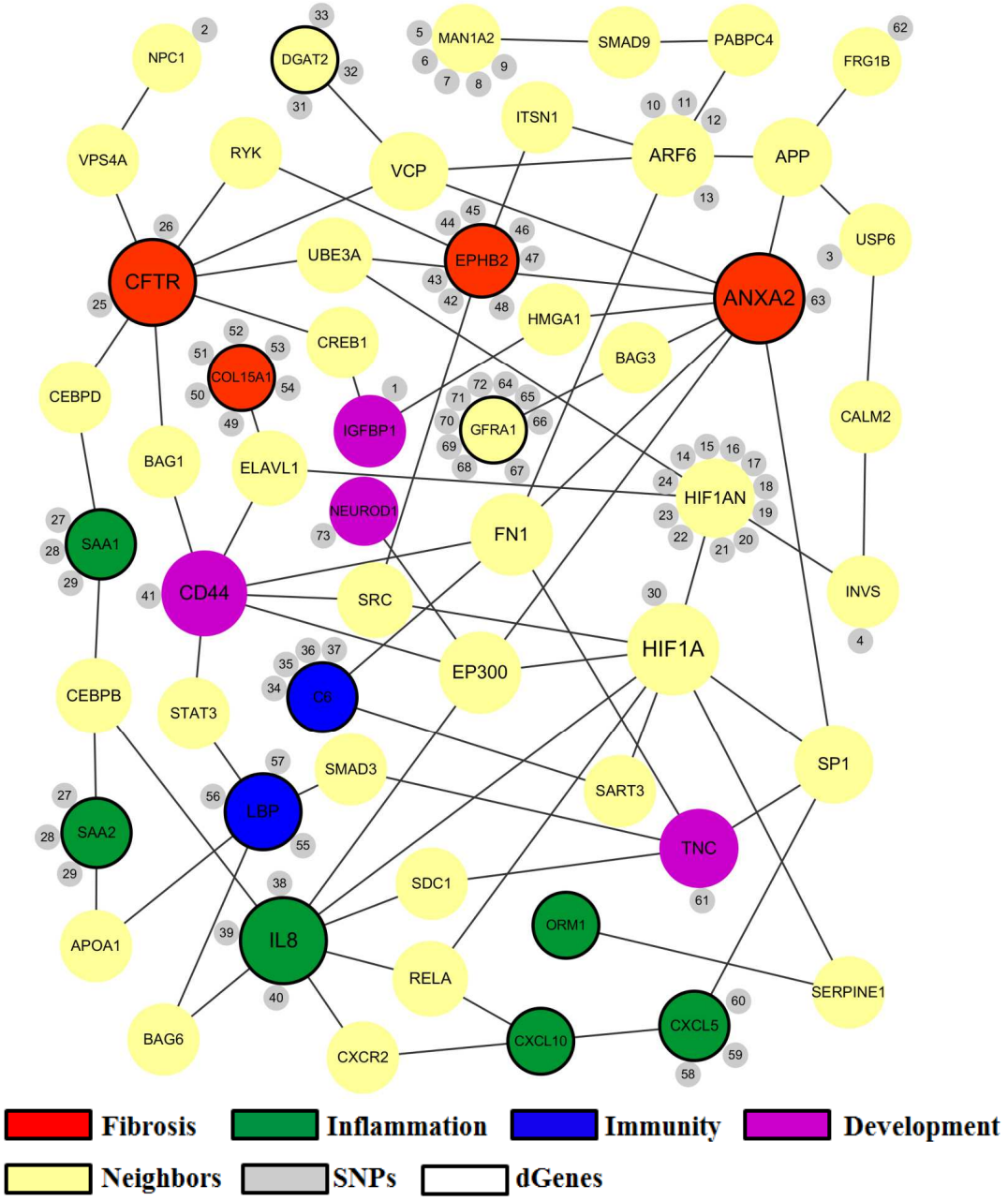


Figure 3.5 Common biological functions in the proposed biliary atresia network. All of the significant genes in the proposed BA network were annotated to identify common biological functions within the network. The red nodes represent the genes related to fibrosis, green related to inflammation, blue related to immune response, and purple related to development. The size of a node depends on the connectivity within the network; the larger the node, the more connected it is to other genes. The yellow nodes represent the neighbor genes that link different significant genes through protein-protein interaction. The small grey nodes represent the SNPs that are associated with their attached genes. The entire list of SNPs, both novel and known, can be found in the Supplementary Material. The black circular edges around the nodes represent differentially regulated genes ($p<0.05$).

The common transcription factors

To identify common transcriptional factors that can regulate the majority of the genes in the proposed BA network, and therefore significantly regulate the development of BA, I have analyzed the enrichment of transcription factor binding sites. **Table 3.5** shows the top 5 common transcription factors sorted by the p-value. The rest of the common transcription factors can be found in the Supplementary Material.

Table 3.5 Top 5 common transcription factors from the BA network. The common transcription factors for the genes in the proposed BA network were identified. *The number of submitted genes that can be regulated by each TF over the total number of genes that can be regulated by each TF is the calculated percentage.

TF	Count	%*	PValue	Genes (EntrezID)
FOXD3	25	49.0196	0.00229	3725, 805, 6774, 2335, 27183, 960, 729, 6259, 6374, 1994, 1462, 4093, 10905, 2048, 1306, 4864, 1080, 6453, 3091, 5196, 1385, 6667, 7337, 4088, 6382
CREBP1	31	60.7843	0.00269	3725, 805, 6774, 5970, 27183, 2335, 8761, 9098, 960, 3484, 6374, 27130, 1994, 573, 1462, 55662, 2048, 1306, 4864, 6453, 1080, 3091, 5196, 1385, 6667, 7337, 6670, 4088, 6382, 2353, 7917
MYB	30	58.8235	0.00293	3725, 805, 6774, 27183, 7415, 2335, 8761, 382, 960, 729, 6259, 3484, 27130, 1994, 573, 1462, 4093, 10905, 2048, 1306, 6714, 6453, 1080, 3091, 1385, 6667, 7337, 4088, 6382, 2353
AP1	36	70.5882	0.00483	805, 6774, 5970, 27183, 7415, 2335, 8761, 382, 9098, 960, 335, 1051, 6259, 27130, 1462, 4093, 10905, 55662, 2048, 2833, 3576, 1306, 6714, 4864, 1080, 84649, 3091, 1385, 6667, 7337, 6670, 4088, 6382, 2353, 5879, 7917
FAC1	30	58.8235	0.00742	3725, 805, 6774, 5970, 2335, 382, 960, 27130, 1462, 4093, 10905, 55662, 2048, 2833, 1306, 6714, 4864, 6453, 1080, 84649, 3091, 1385, 6667, 7337, 4283, 6670, 4088, 6382, 5879, 7917

Discussion

The main findings of this Chapter are the identified potential biomarkers, the relevant biological functions and pathways, and the proposed BA network that was reconstructed using a novel integrative method from different analyses. Here, I delineate the improved understanding of the complex pathogenesis of BA and the essential “interaction” knowledge that is uniquely obtained reconstruction of the highly condensed and interpretable network.

Biliary atresia as a developmental disease

From the whole exome data, I observed the evidence of BA as a developmental disease. The enrichment result revealed the potential importance of embryonic development and hematological system development and function. This is consistent with the suggested pathogenic mechanism for the ‘syndromic’ form of BA- that multiple genomic changes could cause a problem during the embryonic development of the liver and the bile ducts. This theory is further supported by the presence of situs inversus, which is an indication of dysregulated left-right axis determination during the embryonic development, in many BA patients (91). One of the genes involved in left-right axis determination and ciliary development is INVS, whose role in BA pathology has been studied through experimental mouse models (135). Another gene involved in left-right axis determination is fibronectin 1 (FN1). Previous research has shown that FN1-null embryos had abnormal LR

patterning and the pointed pole of the ventral node facing posteriorly instead of anteriorly (136). Although I did not find conclusive evidence for FN1, it interacts with 4 other significant genes, C6, ARF6, ANXA2, and CD44, in the final BA network. Therefore, FN1's interaction with these proteins may qualify FN1 as a key protein in the BA network.

Cilia are microtubule-based organelles that are intricately involved in the left-right asymmetry process (137). The enrichment analysis on the transcriptomic data revealed locomotory behavior, a biological process that is related to ciliary function. Furthermore, 11 genes from the whole exome data were related to ciliary development and function. Besides INVS, PCM1 is another ciliary gene that can interact with HTT, polo-like kinase 1 (PLK1), and centrosomal protein 290 (CEP290) to facilitate proper ciliary development and disassembly (138). Although not in the SYSCILIA gold standard list of ciliary genes, fibroblast growth factor 23 (FGF23) can also affect ciliay function, because it belongs to the FGF family that can regulate cilia-driven nodal flow during the left-right axis specification (137). FGF23 may play a role in facilitating proper ciliary development by interacting with other known FGF proteins, such as FGF8.

"Interaction" knowledge from the proposed BA network

To evaluate BA as a complex disease with multiple mechanisms, I developed a BA network using the interactions of the significant genes discovered from different analyses. The proposed network provides analytic

insights into how groups of genes with either similar or different biological functions can interact to affect the complex pathogenesis of BA. An example of interaction knowledge can be found in CFTR and its interactions. CFTR is the second biggest node in the network, being connected to 7 other significant genes through potential neighbor genes, and it is also part of the fibrotic hub with EPHB2, ANXA2, and COL15A1. From the potential protein-protein interactions in the network, CFTR may influence acute inflammation represented by SAA1, embryonic development represented by IGFBP1, and hematological system and development represented by CD44.

Another significant gene with numerous interactions in the network is IL8, a well-known inflammatory gene that is a part of the inflammation hub with CXCL10, CXCL5, ORM1, SAA1, and SAA2. This gene is located at the center of the inflammation hub and is well-connected to other inflammatory genes, which suggests that IL8 may be a key regulator of inflammation in BA. Additionally, it is connected to LBP, one of the two genes involved in regulation of immune response that is closely related to inflammation.

A small hub of genes representing the developmental mechanism of BA is present in the network. For example, CD44 is known for its wide variety of functions including but not limited to lymphocyte activation, recirculation, hematopoiesis, and tumor metastasis (139). It is part of the developmental hub from the whole exome data and is connected to another developmental gene, TNC. CD44 is also connected to 3 other biological hubs such as inflammation (IL8), immunity (LBP) and fibrosis (CFTR, COL15A1, ANXA2). CD44 may be

considered a key gene involved in the pathogenesis of BA serving as a mediator between different biological functions. Although not included in the network, HTT is another developmental gene that is well connected to many of the significant genes in the network.

In addition to the groups of genes with related biological functions, some of the individual genes and their connections warrant further analysis. For instance, USP6 is a gene involved in membrane localization of ARF6 and regulation of ARF6-dependent endocytic protein trafficking. This particular gene supports the role of ARF6 in BA. According to the bioinformatics tools that can predict whether a missense variant can affect the protein function, namely SIFT and LRT, the novel variant at chr17:5036210 within USP6 was considered to be “damaging.” Furthermore, USP6 is connected to INVS through calmodulin 2 (CALM2), which suggests a potential interaction between the ARF6-dependent BA pathway and the ciliary development.

Common transcription factors in the BA network

Individual genes in the proposed BA network were analyzed to identify common transcription factors that could interact with many of the genes in the network. One such factor is FOXD3, a transcription factor that participates in liver and lung formation from foregut endoderm (140). A study showed that FOXD3 could activate osteopontin enhancer that is expressed in totipotent embryonic stem cells (140), which suggests that FOXD3 is a key transcriptional factor involved in managing the developmental aspect of the

complex pathogenesis of BA. Another important transcription factor is MYB that facilitates proper development of hematopoiesis during embryonic development (141). Lastly, AP1 can regulate gene expression in response to a variety of stimuli, including cytokines, growth factors, stress, and bacterial and viral infections (142). It may regulate the expression of inflammatory genes for the 'syndromic' form of BA.

Potential BA genes not in the network

Despite the valuable information provided by the network, I omitted some of the identified genes from high-throughput analyses due to their lack of interconnectivity with other significant genes. These genes may still be significant because the proposed BA network represents only one large group of variants and genes that are highly interconnected; other groups of variants and genes, such as MUC6 and ANXA2, likely exist. MUC6, in particular, has been linked with viral hepatitis in biliary epithelium and has a highly common novel missense variant that can significantly damage the protein function (133). ANXA2 is another potentially relevant gene that was significantly upregulated in cholangiocytes in primary biliary cirrhosis from a recent study (143).

The strengths and limitations of the proposed BA network

The proposed BA network has several advantages over the conventional non-systematic approach: (i) the network can be used to study both the individual interactions among specific genes and the complex

relationship among the major hubs of genes involved in fibrosis, immunity, inflammation, and development; (ii) the network is highly trainable; future experiments could refine the genes and their individual interactions from the existing version of the network; (iii) the network contains the variants with high allele frequency, which ensures a high degree of confidence and replicability; (iv) common biological functions, such as inflammation, regulation of immune response, and embryonic development, were shared between the proposed BA network and the whole exome network.

Due to the nature of studying a rare population such as BA, a major limitation of the results is the lack of a large enough sample size for high statistical power. I attempted to alleviate this problem by integrating the results of different datasets as well as applying a stringent allele frequency cutoff to only select the variants that are highly frequent in the BA patients. Furthermore, the lack of confidence from statistical results is mitigated by the consistency of the findings with the published results. Many of the genes and variants identified from the various analyses and the reconstructed BA network were involved in the known biological functions of BA, while a few significant genes in the proposed BA network were validated from the published literature.

Conclusion

Chapter 3 provides a mechanistic view of the complex pathogenesis of BA through a novel network reconstruction method which reveals key insights into how significant genes with different biological functions interact to

contribute to the development of BA. In this Chapter, I have first performed target and whole exome sequencing analyses to discover novel variants as potential BA markers. In addition, by integrating multiple analytical results from Chapters 2 and 3, I then reconstructed a comprehensive BA network that is enriched in biological functions such as inflammation, fibrosis, and development.

Supplementary Materials

Table S3.1 Numbered SNPs in the proposed biliary atresia network. SNPs with proper rsID are known SNPs according to the dbSNP 138 database. SNPs that have chrN:start-end:allele change are novel.

#	SNP	Gene
1	rs4619	IGFBP1
2	chr18:21124945:C->G	NPC1
3	chr17:5036210:T->G	USP6
4	rs3813712	INVS
5	rs12131109	MAN1A2
6	rs7531715	MAN1A2
7	rs6657965	MAN1A2
8	chr1:117991089:T->C	MAN1A2
9	chr1:117991086:T->G	MAN1A2
10	rs3126184	ARF6
11	rs10140366	ARF6
12	chr14:50369743:G->T	ARF6
13	chr14:50369741:C->T	ARF6
14	rs2495725	HIF1AN
15	rs7092999	HIF1AN
16	rs3763696	HIF1AN
17	rs2495718	HIF1AN
18	rs1110286	HIF1AN
19	chr10:102371717:A->G	HIF1AN
20	chr10:102381938:G->C	HIF1AN
21	chr10:102335587:C->A	HIF1AN
22	chr10:102335588:C->T	HIF1AN
23	chr10:102371719:A->G	HIF1AN
24	chr10:102288649:T->A	HIF1AN
25	rs3808185	CFTR
26	rs2237724	CFTR
27	chr11:18275399:G->A	SAA1/2
28	chr11:18306706:G->A	SAA1/2
29	chr11:18244418:T->G	SAA1/2
30	rs1957757	HIF1A
31	rs1458836	DGAT2
32	chr11:75511683:A->G	DGAT2
33	chr11:75510671:A->T	DGAT2

Table S3.1 Numbered SNPs in the proposed biliary atresia network. (Continued)

#	SNP	Gene
34	rs11743598	C6
35	rs3805715	C6
36	rs1801033	C6
37	rs751138	C6
38	chr4:74617686:T->G	IL8
39	chr4:74612590:T->A	IL8
40	chr4:74612589:C->G	IL8
41	rs9666607	CD44
42	rs6667416	EPHB2
43	rs4655107	EPHB2
44	rs10753545	EPHB2
45	rs4655128	EPHB2
46	rs12027585	EPHB2
47	chr1:23245302:G->A	EPHB2
48	chr1:23224633:G->A	EPHB2
49	rs10819542	COL15A1
50	rs3780622	COL15A1
51	rs4743322	COL15A1
52	chr9:101791486:A->G	COL15A1
53	chr9:101838582:T->A	COL15A1
54	chr9:101802438:G->A	COL15A1
55	rs2232618	LBP
56	chr20:36980457:G->A	LBP
57	chr20:36950014:T->A	LBP
58	chr4:74853963:A->C	CXCL5
59	chr4:74837722:A->G	CXCL5
60	chr4:74837726:C->G	CXCL5
61	rs1757095	TNC

Table S3.2 List of BA patient samples used in each data sources. "R" indicates the samples that were removed due to low sequencing depth.

Patient #	GWAS-family trio-TDT	Target sequencing	Whole exome sequencing	RNAseq
1	Y	Y	Y	
2	Y			
3	Y	Y	Y	
4	Y	Y	Y	
5	Y	Y	Y	
6	Y	Y	R	
7	Y	Y	R	
8	Y			
9	Y			
10	Y	Y	Y	
11	Y	Y	Y	
12	Y	Y	Y	
13	Y	Y	Y	
14	Y		R	
15	Y			
16	Y			
17	Y	Y	Y	
18	Y	Y	Y	
19	Y	Y	Y	
20	Y	Y	R	
21	Y			
22	Y	Y	Y	
23	Y			
24	Y	Y	Y	
25	Y	Y	Y	Y
26	Y	Y	Y	Y
27	Y	Y	Y	Y
28	Y	Y	Y	Y
29	Y	Y	R	
30	Y			
31	Y	Y	Y	
32	Y	Y	Y	
33	Y	Y	Y	
34	Y			
35	Y	Y	Y	
36	Y		Y	

Table S3.2 List of BA patient samples used in each data sources. (Continued)

Patient #	GWAS-family trio-TDT	Target sequencing	Whole exome sequencing	RNAseq
37		Y	Y	
38			Y	
39		Y	Y	
40		Y	Y	
41		Y	Y	
42		Y	Y	
43			Y	
44		Y	Y	
45		Y	Y	
46		Y	Y	
47		Y	Y	
48		Y	Y	
49			Y	
50		Y	Y	
51		Y	Y	
52		Y	Y	
53		Y	Y	Y
54			Y	
55		Y	Y	
56			Y	
57			Y	
58			Y	
59			Y	
60		Y	Y	
61			Y	
62		Y	Y	
63		Y	Y	Y
Total	36	43	49	6

Table S3.3 Predicted functions of novel variants. Predicted functions of novel variants, both SNPs and indels, with AF>0.4 from the whole exome data are listed. SnpEff and VCF v1.0 annotation databases were used.

Chromosome	Coordinates	Predicted effect	Gene
chr19	56599437	inframe_deletion	ZNF787
chr21	46924425	disruptive_inframe_deletion	COL18A1
chr11	95825374	inframe_deletion	MAML2
chr8	1.45E+08	inframe_deletion	MAFA
chr7	8196567	frameshift_variant	ICA1
chr13	72440658	inframe_deletion	DACH1
chr4	1.48E+08	disruptive_inframe_insertion	POU4F2
chr6	1.1E+08	splice_acceptor_variant&intron_variant	FIG4
chr12	6777069	inframe_insertion	ZNF384
chr17	7750177	disruptive_inframe_insertion	KDM6B
chr5	1.13E+08	inframe_insertion	MCC
chr7	1.42E+08	missense_variant	TRBV6-7
chr14	1.04E+08	frameshift_variant	EXOC3L4
chr4	1388350	frameshift_variant	CRIPAK
chr10	1.26E+08	frameshift_variant	CHST15
chr15	80736696	splice_acceptor_variant etc	RP11-210M15.1
chr14	23744800	inframe_deletion	HOMEZ
chr3	1.34E+08	frameshift_variant	RYK
chr7	1.51E+08	frameshift_variant	AGAP3
chr7	1.51E+08	frameshift_variant	AGAP3
chr3	1.34E+08	frameshift_variant&splice_region_variant	RYK
chr5	72743299	frameshift_variant	FOXD1
chr17	46115122	splice_acceptor_variant&splice_donor_variant&intron_variant	COPZ2
chr17	26708298	frameshift_variant	TMEM199
chr17	26708302	frameshift_variant	TMEM199
chr17	48227384	frameshift_variant&splice_region_variant	PPP1R9B
chr17	48227403	frameshift_variant&splice_region_variant	PPP1R9B
chr14	1.06E+08	frameshift_variant	IGHJ6
chr14	1.06E+08	frameshift_variant	IGHJ6
chr21	45959556	frameshift_variant	KRTAP10-1
chr21	45959557	frameshift_variant	KRTAP10-1
chr18	21124945	missense_variant	NPC1
chr20	1592150	frameshift_variant	SIRPB1
chr20	1592154	frameshift_variant	SIRPB1
chr3	75790810	frameshift_variant	ZNF717

Table S3.3 Predicted functions of novel variants. (Continued)

chr7	1.42E+08	missense_variant	TRBC2
chr17	21319650	inframe_deletion	KCNJ12
chr2	96616501	frameshift_variant	ANKRD36C
chr20	29632662	missense_variant	FRG1B
chr15	23685604	frameshift_variant	GOLGA6L2
chr7	1.42E+08	missense_variant	TRBV6-7
chr7	1.42E+08	missense_variant	TRBV6-7
chr17	5036210	missense_variant	USP6
chr11	1017135	missense_variant	MUC6
chr9	1.32E+08	frameshift_variant	CCBL1
chr12	51740407	frameshift_variant&splice_region_variant	CELA1
chr12	51740415	frameshift_variant	CELA1
chr12	51740416	frameshift_variant	CELA1
chr1	1850627	disruptive_inframe_deletion	TMEM52
chr11	56143782	frameshift_variant	OR8U1
chr3	10088407	splice_donor_variant&intron_variant	FANCD2

Table S3.4 Common transcription factors from the BA network. The common transcription factors for the genes in the proposed BA network were identified. The transcription factors are ordered by the p-values. *The number of submitted genes that can be regulated by each TF over the total number of genes that can be regulated by each TF is the calculated percentage.

TF	Count	%*	P-value	Genes (EntrezID)
FOXD3	25	49.0196	0.00229	3725, 805, 6774, 2335, 27183, 960, 729, 6259, 6374, 1994, 1462, 4093, 10905, 2048, 1306, 4864, 1080, 6453, 3091, 5196, 1385, 6667, 7337, 4088, 6382
CREBP1	31	60.7843	0.00269	3725, 805, 6774, 5970, 27183, 2335, 8761, 9098, 960, 3484, 6374, 27130, 1994, 573, 1462, 55662, 2048, 1306, 4864, 6453, 1080, 3091, 5196, 1385, 6667, 7337, 6670, 4088, 6382, 2353, 7917
MYB	30	58.8235	0.00293	3725, 805, 6774, 27183, 7415, 2335, 8761, 382, 960, 729, 6259, 3484, 27130, 1994, 573, 1462, 4093, 10905, 2048, 1306, 6714, 6453, 1080, 3091, 1385, 6667, 7337, 4088, 6382, 2353
AP1	36	70.5882	0.00483	805, 6774, 5970, 27183, 7415, 2335, 8761, 382, 9098, 960, 335, 1051, 6259, 27130, 1462, 4093, 10905, 55662, 2048, 2833, 3576, 1306, 6714, 4864, 1080, 84649, 3091, 1385, 6667, 7337, 6670, 4088, 6382, 2353, 5879, 7917

Table S3.4 Predicted functions of novel variants. (Continued)

TF	Count	%*	P-value	Genes (EntrezID)
FAC1	30	58.8235	0.00742	3725, 805, 6774, 5970, 2335, 382, 960, 27130, 1462, 4093, 10905, 55662, 2048, 2833, 1306, 6714, 4864, 6453, 1080, 84649, 3091, 1385, 6667, 7337, 4283, 6670, 4088, 6382, 5879, 7917
RREB1	25	49.0196	0.01299	805, 6774, 5970, 2335, 7415, 27183, 8761, 6259, 1462, 4093, 55662, 2833, 2048, 6714, 4864, 1080, 6453, 84649, 3091, 6667, 7337, 4088, 6382, 2353, 7917
FOXJ2	37	72.5490	0.01466	805, 6774, 5970, 27183, 7415, 2335, 382, 9098, 960, 729, 6374, 6259, 1994, 27130, 3579, 573, 1462, 4093, 10905, 55662, 2048, 2833, 1306, 4864, 1080, 6453, 3091, 5196, 1385, 6667, 7337, 6670, 4088, 6382, 2353, 5879, 7917
FOXO4	32	62.7451	0.01481	3725, 805, 6774, 7415, 2335, 8761, 9098, 960, 3929, 6259, 3484, 27130, 1994, 1462, 10905, 55662, 2048, 1306, 6714, 6453, 1080, 84649, 3091, 1385, 6667, 7337, 6670, 4088, 6382, 2353, 5879, 7917
STAT	23	45.0980	0.01832	3725, 1306, 6453, 6774, 5970, 5196, 27183, 2335, 7415, 1385, 9098, 6667, 960, 7337, 729, 6374, 6259, 4088, 6382, 2353, 1462, 7917, 2048
TBP	23	45.0980	0.01873	3725, 805, 6453, 6774, 3091, 27183, 8761, 382, 1385, 9098, 960, 7337, 6670, 3484, 27130, 4088, 2353, 1462, 4093, 10905, 55662, 7917, 2048
EN1	31	60.7843	0.02413	3725, 805, 6774, 5970, 27183, 7415, 2335, 9098, 960, 1051, 27130, 1994, 573, 1462, 4093, 10905, 55662, 2048, 2833, 1306, 4864, 6453, 1080, 3091, 1385, 6667, 7337, 6670, 4088, 2353, 7917
IRF7	27	52.94118	0.027574	3725, 805, 6774, 2335, 8761, 960, 729, 3484, 6374, 27130, 1994, 573, 1462, 55662, 2048, 1306, 4864, 6453, 3091, 5196, 6667, 7337, 4283, 4088, 6382, 2353, 7917
PAX6	31	60.78431	0.029324	805, 6288, 5970, 7415, 2335, 9098, 960, 1051, 729, 6259, 27130, 1994, 573, 1462, 4093, 10905, 55662, 2048, 2833, 4864, 1080, 6453, 3091, 1385, 6667, 7337, 4088, 6382, 2353, 5879, 7917
FOXO1	25	49.01961	0.032449	1306, 6714, 805, 6453, 1080, 6289, 1385, 9098, 6667, 960, 7337, 729, 6259, 6670, 1994, 27130, 4088, 6382, 573, 2353, 1462, 4093, 10905, 2048, 2833

Table S3.4 Predicted functions of novel variants. (Continued)

TF	Count	%*	P-value	Genes (EntrezID)
RFX1	34	66.66667	0.035076	3725, 805, 6774, 5970, 27183, 7415, 2335, 8761, 382, 9098, 960, 1051, 729, 3929, 6374, 1052, 27130, 573, 1462, 10905, 2048, 2833, 1306, 6714, 4864, 6453, 84649, 3091, 1385, 6667, 4283, 4088, 6382, 5879
HSF2	25	49.01961	0.037073	3725, 6774, 6288, 5970, 27183, 2335, 7415, 382, 960, 729, 3929, 27130, 1462, 573, 10905, 2048, 1306, 1080, 84649, 7337, 4283, 6670, 4088, 6382, 7917
NFKAPPAB	22	43.13725	0.037542	1306, 6714, 1080, 6774, 84649, 5970, 3091, 8761, 1385, 6667, 960, 3929, 6374, 6259, 6670, 573, 1462, 5879, 55662, 2048, 7917, 3576
SRY	25	49.01961	0.038226	3725, 1306, 805, 6453, 1080, 6774, 3091, 2335, 8761, 1385, 6667, 960, 7337, 729, 3929, 6259, 6670, 27130, 1994, 4088, 2353, 1462, 4093, 10905, 2048
TST1	26	50.98039	0.038318	6774, 5970, 27183, 7415, 2335, 8761, 9098, 960, 6259, 27130, 1994, 10905, 2048, 1306, 6714, 4864, 6453, 1080, 84649, 3371, 1385, 6670, 4088, 2353, 5879, 7917
HNF1	31	60.78431	0.043075	805, 6774, 2335, 8761, 382, 9098, 960, 729, 6259, 3484, 27130, 1994, 1462, 10905, 55662, 2048, 6714, 4864, 6453, 1080, 84649, 3091, 1385, 6667, 7337, 6670, 4088, 6382, 2353, 5879, 7917
GATA	26	50.98039	0.045219	3725, 6714, 805, 6289, 6774, 1080, 6453, 5970, 3091, 2335, 1385, 960, 7337, 729, 3929, 3484, 6670, 6259, 4088, 1462, 2353, 5879, 4093, 10905, 55662, 2833
HLF	24	47.05882	0.045863	1306, 805, 6453, 6289, 1080, 6288, 3091, 2335, 9098, 7337, 729, 6670, 3484, 1994, 27130, 4088, 6382, 573, 2353, 1462, 5879, 10905, 55662, 7917

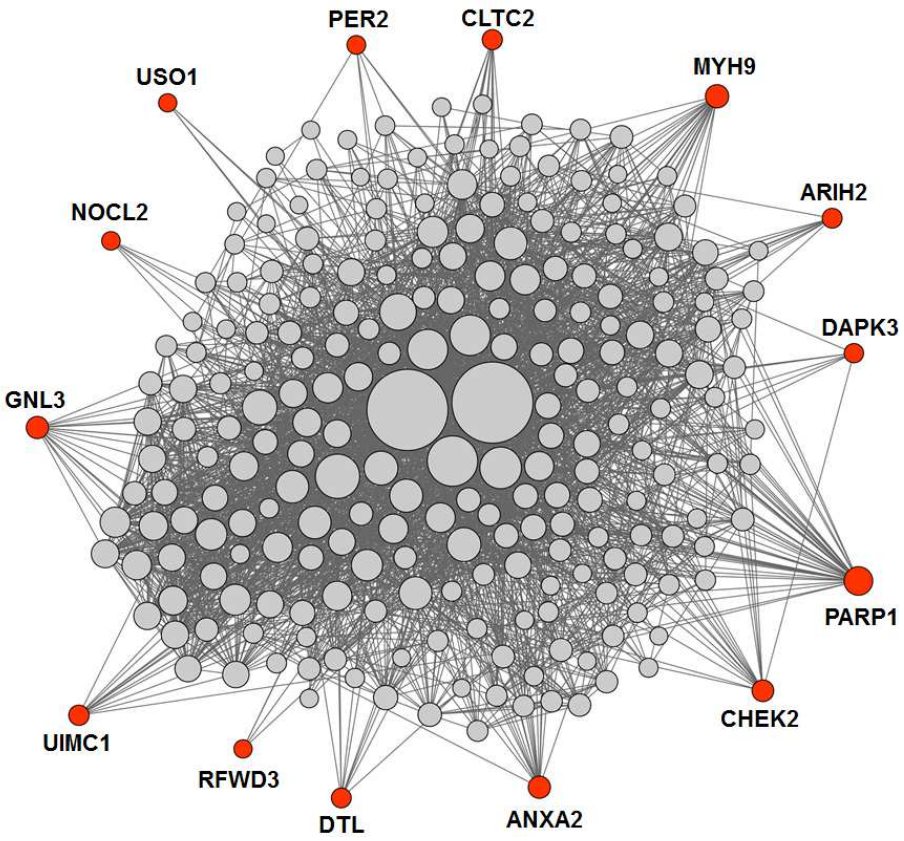


Figure S3.1 The first cluster of the MCODE algorithm on the whole exome network. MCODE clustering algorithm was applied to the whole exome network to identify smaller clusters of genes that are densely connected to each other. The cluster with the best MCODE score is shown. The top 5 GO:BP terms were also identified for the genes in the cluster. The red nodes indicate the original significant genes from the whole exome data.

Table S3.5 List of genes mapped from the common SNPs in the whole exome data and the internal list of GWAS. The common SNPs in the two lists are shown with their gene symbols and descriptions.

EntrezID	Gene Symbol	Gene Description
368	ABCC6	ATP-binding cassette, sub-family C, member 6 pseudogene 2; ATP-binding cassette, sub-family C (CFTR/MRP), member 6
593	BCKDHA	branched chain keto acid dehydrogenase E1, alpha polypeptide
960	CD44	CD44 molecule (Indian blood group)
1072	CFL1	cofilin 1 (non-muscle)
1179	CLCA1	chloride channel accessory 1
1187	CLCNKA	chloride channel Ka
1292	COL6A2	collagen, type VI, alpha 2
1297	COL9A1	collagen, type IX, alpha 1

Table S3.5 List of genes mapped from the common SNPs in the whole exome data and the internal list of GWAS. (Continued)

EntrezID	Gene Symbol	Gene Description
1629	DBT	dihydrolipoamide branched chain transacylase E2
1994	ELAVL1	ELAV (embryonic lethal, abnormal vision, Drosophila)-like 1 (Hu antigen R)
2195	Fat1	FAT tumor suppressor homolog 1 (Drosophila)
2203	Fbp1	fructose-1,6-bisphosphatase 1
2317	FLNB	filamin B, beta (actin binding protein 278)
2444	FRK	fyn-related kinase
2524	FUT2	fucosyltransferase 2 (secretor status included)
2638	gc	group-specific component (vitamin D binding protein)
2868	GRK4	G protein-coupled receptor kinase 4
3046	HBE1	hemoglobin, epsilon 1
3048	HBG2	hemoglobin, gamma G
3064	HTT	huntingtin
3371	TNC	tenascin C
3373	Hyal1	hyaluronoglucosaminidase 1
3484	IGFBP1	insulin-like growth factor binding protein 1
3508	IGHMBP2	immunoglobulin mu binding protein 2
3601	IL15RA	interleukin 15 receptor, alpha
3823	Klrc3	killer cell lectin-like receptor subfamily C, member 3
3882	Krt32	keratin 32
4008	Lmo7	LIM domain 7
4036	LRP2	low density lipoprotein-related protein 2
4240	MFGE8	milk fat globule-EGF factor 8 protein
4259	Mgst3	microsomal glutathione S-transferase 3
4522	mthfd1	methylenetetrahydrofolate dehydrogenase (NADP+ dependent) 1, methenyltetrahydrofolate cyclohydrolase, formyltetrahydrofolate synthetase
4585	MUC4	mucin 4, cell surface associated
4760	NEUROD1	neurogenic differentiation 1
4892	Nrap	nebulin-related anchoring protein
4992	OR1F1	olfactory receptor, family 1, subfamily F, member 1
5002	Slc22a18	solute carrier family 22, member 18
5003	SLC22A18AS	solute carrier family 22 (organic cation transporter), member 18 antisense
5108	pcm1	pericentriolar material 1
5176	Serpinf1	serpin peptidase inhibitor, clade F (alpha-2 antiplasmin, pigment epithelium derived factor), member 1
5275	Serpinb13	serpin peptidase inhibitor, clade B (ovalbumin), member 13

Table S3.5 List of genes mapped from the common SNPs in the whole exome data and the internal list of GWAS. (Continued)

EntrezID	Gene Symbol	Gene Description
5288	PIK3C2G	phosphoinositide-3-kinase, class 2, gamma polypeptide
5625	PRODH	proline dehydrogenase (oxidase) 1
5858	PZP	pregnancy-zone protein
6332	SCN7A	sodium channel, voltage-gated, type VII, alpha
6370	ccl25	chemokine (C-C motif) ligand 25
6491	STIL	SCL/TAL1 interrupting locus
6493	SIM2	single-minded homolog 2 (Drosophila)
6519	Slc3a1	solute carrier family 3 (cystine, dibasic and neutral amino acid transporters, activator of cystine, dibasic and neutral amino acid transport), member 1
6565	Slc15a2	solute carrier family 15 (H+/peptide transporter), member 2
6585	slit1	slit homolog 1 (Drosophila)
6614	SIGLEC1	sialic acid binding Ig-like lectin 1, sialoadhesin
6653	SORL1	sortilin-related receptor, L(DLR class) A repeats-containing
7143	TNR	tenascin R (restrictin, janusin)
7766	ZNF223	zinc finger protein 223
7772	ZNF229	zinc finger protein 229
7866	Ifrd2	interferon-related developmental regulator 2
8029	cubn	cubilin (intrinsic factor-cobalamin receptor)
8100	IFT88	intraflagellar transport 88 homolog (Chlamydomonas)
8214	DGCR6	DiGeorge syndrome critical region gene 6
8302	KLRC4	killer cell lectin-like receptor subfamily C, member 4
8372	HYAL3	hyaluronoglucosaminidase 3
8558	cdk10	cyclin-dependent kinase 10
8602	NOP14	NOP14 nucleolar protein homolog (yeast)
8701	DNAH11	dynein, axonemal, heavy chain 11
8735	MYH13	myosin, heavy chain 13, skeletal muscle
8736	myom1	myomesin 1, 185kDa
8793	TNFRSF10D	tumor necrosis factor receptor superfamily, member 10d, decoy with truncated death domain
8871	SYNJ2	synaptojanin 2
8877	SPHK1	sphingosine kinase 1
8899	prpf4b	similar to hCG1820375; PRP4 pre-mRNA processing factor 4 homolog B (yeast)
9013	TAF1C	TATA box binding protein (TBP)-associated factor, RNA polymerase I, C, 110kDa
9154	Slc28a1	solute carrier family 28 (sodium-coupled nucleoside transporter), member 1
9389	SLC22A14	solute carrier family 22, member 14

Table S3.5 List of genes mapped from the common SNPs in the whole exome data and the internal list of GWAS. (Continued)

EntrezID	Gene Symbol	Gene Description
9510	Adamts1	ADAM metallopeptidase with thrombospondin type 1 motif, 1
9518	Gdf15	growth differentiation factor 15
9581	PREPL	prolyl endopeptidase-like
9609	rab36	RAB36, member RAS oncogene family
9808	KIAA0087	KIAA0087
10160	FARP1	FERM, RhoGEF (ARHGEF) and pleckstrin domain protein 1 (chondrocyte-derived)
10345	TRDN	triadin
10350	Abca9	ATP-binding cassette, sub-family A (ABC1), member 9
10400	pemt	phosphatidylethanolamine N-methyltransferase
10531	PITRM1	pitrilysin metallopeptidase 1
10616	RBCK1	RanBP-type and C3HC4-type zinc finger containing 1
10827	Fam114a2	family with sequence similarity 114, member A2
11066	snrnp35	ATP-binding cassette, sub-family B (MDR/TAP), member 5; small nuclear ribonucleoprotein 35kDa (U11/U12)
11148	HHLA2	HERV-H LTR-associating 2
11196	SEC23IP	SEC23 interacting protein
11201	poli	polymerase (DNA directed) iota
11214	AKAP13	A kinase (PRKA) anchor protein 13
11264	PXMP4	peroxisomal membrane protein 4, 24kDa
22824	Hspa4l	heat shock 70kDa protein 4-like
22838	RNF44	ring finger protein 44
22876	INPP5F	inositol polyphosphate-5-phosphatase F
23013	spen	spen homolog, transcriptional regulator (Drosophila)
23217	ZFR2	zinc finger RNA binding protein 2
23223	Rrp12	ribosomal RNA processing 12 homolog (S. cerevisiae)
23224	SYNE2	spectrin repeat containing, nuclear envelope 2
23279	NUP160	nucleoporin 160kDa
23325	KIAA1033	KIAA1033
23345	SYNE1	spectrin repeat containing, nuclear envelope 1
23351	khnyn	KIAA0323
23362	PSD3	pleckstrin and Sec7 domain containing 3
23460	Abca6	ATP-binding cassette, sub-family A (ABC1), member 6
23627	Prnd	prion protein 2 (dublet)
23767	Flrt3	fibronectin leucine rich transmembrane protein 3
24142	NAT6	N-acetyltransferase 6 (GCN5-related)
25878	MXRA5	matrix-remodelling associated 5

Table S3.5 List of genes mapped from the common SNPs in the whole exome data and the internal list of GWAS. (Continued)

EntrezID	Gene Symbol	Gene Description
25938	heatr5a	HEAT repeat containing 5A
27122	DKK3	dickkopf homolog 3 (Xenopus laevis)
27130	Invs	inversin
27283	TINAG	tubulointerstitial nephritis antigen
28671	TRAV13-1	T cell receptor alpha variable 13-1
29070	CCDC113	coiled-coil domain containing 113
29119	CTNNA3	catenin (cadherin-associated protein), alpha 3
50617	ATP6V0A4	ATPase, H+ transporting, lysosomal V0 subunit a4
50999	Tmed5	transmembrane emp24 protein transport domain containing 5
51222	ZNF219	zinc finger protein 219
51321	Amz2	archaelysin family metallopeptidase 2
51473	DCDC2	doublecortin domain containing 2
51530	zc3hc1	zinc finger, C3HC-type containing 1
51700	cyb5r2	cytochrome b5 reductase 2
53827	FXYD5	FXYD domain containing ion transport regulator 5
53904	MYO3A	myosin IIIA
54465	etaa1	Ewing tumor-associated antigen 1
54502	RBM47	RNA binding motif protein 47
54522	ankrd16	ankyrin repeat domain 16
54596	L1TD1	LINE-1 type transposase domain containing 1
54714	CNGB3	cyclic nucleotide gated channel beta 3
54860	Ms4a12	membrane-spanning 4-domains, subfamily A, member 12
54881	Tex10	testis expressed 10
55062	WIPI1	WD repeat domain, phosphoinositide interacting 1
55101	ATP5SL	ATP5S-like
55106	SLFN12	schlafen family member 12
55132	Larp1b	La ribonucleoprotein domain family, member 1B
55258	Thnsl2	threonine synthase-like 2 (S. cerevisiae)
55584	CHRNA9	cholinergic receptor, nicotinic, alpha 9
55614	KIF16B	kinesin family member 16B
55624	POMGNT1	protein O-linked mannose beta1,2-N-acetylglucosaminyltransferase
55742	PARVA	parvin, alpha
55757	Uggt2	UDP-glucose ceramide glucosyltransferase-like 2
55781	RIOK2	RIO kinase 2 (yeast)
55833	UBAP2	ubiquitin associated protein 2
56547	MMP26	matrix metallopeptidase 26

Table S3.5 List of genes mapped from the common SNPs in the whole exome data and the internal list of GWAS. (Continued)

EntrezID	Gene Symbol	Gene Description
56890	Mdm1	Mdm1 nuclear protein homolog (mouse)
56893	UBQLN4	ubiquilin 4
56916	Smarcad1	SWI/SNF-related, matrix-associated actin-dependent regulator of chromatin, subfamily a, containing DEAD/H box 1
57127	RHBG	Rh family, B glycoprotein (gene/pseudogene)
57188	Adamtsl3	ADAMTS-like 3
57539	WDR35	WD repeat domain 35
57572	Dock6	dedicator of cytokinesis 6
57647	DHX37	DEAH (Asp-Glu-Ala-His) box polypeptide 37
58499	znf462	zinc finger protein 462
60401	EDA2R	ectodysplasin A2 receptor
60681	Fkbp10	FK506 binding protein 10, 65 kDa
63893	UBE2O	ubiquitin-conjugating enzyme E2O
64651	CSRNP1	cysteine-serine-rich nuclear protein 1
79345	OR51B2	olfactory receptor, family 51, subfamily B, member 2
79443	Fyco1	FYVE and coiled-coil domain containing 1
79482	OR5AL1	olfactory receptor, family 5, subfamily AL, member 1 (gene/pseudogene)
79671	NLRX1	NLR family member X1
79677	SMC6	structural maintenance of chromosomes 6
79785	RERGL	RERG/RAS-like
79841	agbl2	ATP/GTP binding protein-like 2
79849	PDZD3	PDZ domain containing 3
80010	Rmi1	RMI1, RecQ mediated genome instability 1, homolog (S. cerevisiae)
80144	FRAS1	Fraser syndrome 1
80198	MUS81	MUS81 endonuclease homolog (S. cerevisiae)
80205	CHD9	chromodomain helicase DNA binding protein 9
80274	Scube1	signal peptide, CUB domain, EGF-like 1
80309	SPHKAP	SPHK1 interactor, AKAP domain containing
81618	ITM2C	integral membrane protein 2C
81704	DOCK8	dedicator of cytokinesis 8
83878	Ushbp1	Usher syndrome 1C binding protein 1
84224	NBPF3	neuroblastoma breakpoint family, member 3
84467	FBN3	fibrillin 3
84639	IL1F10	interleukin 1 family, member 10 (theta)
84700	MYO18B	myosin XVIIIB
84899	Tmtc4	transmembrane and tetratricopeptide repeat containing 4

Table S3.5 List of genes mapped from the common SNPs in the whole exome data and the internal list of GWAS. (Continued)

EntrezID	Gene Symbol	Gene Description
90075	ZNF30	zinc finger protein 30
90313	TP53I13	tumor protein p53 inducible protein 13
90668	LRRC16B	leucine rich repeat containing 16B
91373	Uap1l1	UDP-N-acteylglucosamine pyrophosphorylase 1-like 1
91862	marveld3	MARVEL domain containing 3
91937	TIMD4	T-cell immunoglobulin and mucin domain containing 4
92106	oxnad1	oxidoreductase NAD-binding domain containing 1
92196	DAPL1	death associated protein-like 1
93190	C1orf158	chromosome 1 open reading frame 158
113146	AHNAK2	AHNAK nucleoprotein 2
114780	PKD1L2	polycystic kidney disease 1-like 2
114784	CSMD2	CUB and Sushi multiple domains 2
114826	SMYD4	SET and MYND domain containing 4
116236	ABHD15	abhydrolase domain containing 15
119679	OR52J3	olfactory receptor, family 52, subfamily J, member 3
119692	OR51S1	olfactory receptor, family 51, subfamily S, member 1
122618	PLD4	phospholipase D family, member 4
124044	SPATA2L	spermatogenesis associated 2-like
125965	Cox6b2	cytochrome c oxidase subunit VIb polypeptide 2 (testis)
126364	LRRC25	leucine rich repeat containing 25
126370	OR1I1	olfactory receptor, family 1, subfamily I, member 1
126375	ZNF792	zinc finger protein 792
126549	Ankle1	ankyrin repeat and LEM domain containing 1
126767	AADACL3	arylacetamide deacetylase-like 3
127602	DNAH14	dynein, axonemal, heavy chain 14
128372	OR6N1	olfactory receptor, family 6, subfamily N, member 1
129025	ZNF280A	zinc finger protein 280A
136288	C7orf57	chromosome 7 open reading frame 57
138881	OR1L8	olfactory receptor, family 1, subfamily L, member 8
138882	OR1N2	olfactory receptor, family 1, subfamily N, member 2
140733	macrod2	MACRO domain containing 2
146723	C17orf77	chromosome 17 open reading frame 77
147929	ZNF565	zinc finger protein 565
152137	CCDC50	coiled-coil domain containing 50
159989	CCDC67	coiled-coil domain containing 67
162972	ZNF550	zinc finger protein 550

Table S3.5 List of genes mapped from the common SNPs in the whole exome data and the internal list of GWAS. (Continued)

EntrezID	Gene Symbol	Gene Description
163786	Sass6	spindle assembly 6 homolog (C. elegans)
169611	OLFML2A	olfactomedin-like 2A
199223	ttc21a	tetratricopeptide repeat domain 21A
202243	CCDC125	coiled-coil domain containing 125
202500	TCTE1	t-complex-associated-testis-expressed 1
203427	Slc25a43	solute carrier family 25, member 43
219479	OR5R1	olfactory receptor, family 5, subfamily R, member 1
219790	RTKN2	rhotekin 2
220323	OAF	OAF homolog (Drosophila)
221074	SLC39A12	solute carrier family 39 (zinc transporter), member 12
221806	VWDE	von Willebrand factor D and EGF domains
221935	SDK1	sidekick homolog 1, cell adhesion molecule (chicken); hypothetical LOC730351
245802	MS4A6E	membrane-spanning 4-domains, subfamily A, member 6E
254122	SNX32	sorting nexin 32
255189	PLA2G4F	phospholipase A2, group IVF
255239	ANKK1	ankyrin repeat and kinase domain containing 1
255394	TCP11L2	t-complex 11 (mouse)-like 2
256051	ZNF549	zinc finger protein 549
256297	PTF1A	pancreas specific transcription factor, 1a
266747	RGL4	ral guanine nucleotide dissociation stimulator-like 4
282763	OR51B5	olfactory receptor, family 51, subfamily B, member 5
283152	CCDC153	coiled-coil domain containing 153
284018	C17orf58	chromosome 17 open reading frame 58
284415	VSTM1	V-set and transmembrane domain containing 1
284418	Fam71e2	family with sequence similarity 71, member E2
284958	nt5dc4	5'-nucleotidase domain containing 4
285315	c3orf33	chromosome 3 open reading frame 33
338751	OR52L1	olfactory receptor, family 52, subfamily L, member 1
340273	Abcb5	ATP-binding cassette, sub-family B (MDR/TAP), member 5; small nuclear ribonucleoprotein 35kDa (U11/U12)
341568	OR8S1	olfactory receptor, family 8, subfamily S, member 1
344561	GPR148	G protein-coupled receptor 148
348654	GEN1	Gen homolog 1, endonuclease (Drosophila)
374308	PTCHD3	patched domain containing 3
374907	B3GNT8	UDP-GlcNAc:betaGal beta-1,3-N-acetylglucosaminyltransferase 8
375298	CERKL	ceramide kinase-like

Table S3.5 List of genes mapped from the common SNPs in the whole exome data and the internal list of GWAS. (Continued)

EntrezID	Gene Symbol	Gene Description
390157	OR8K1	olfactory receptor, family 8, subfamily K, member 1
390882	OR7G2	olfactory receptor, family 7, subfamily G, member 2
391196	OR2M7	olfactory receptor, family 2, subfamily M, member 7
399814	C10orf120	chromosome 10 open reading frame 120
403284	OR6C68	olfactory receptor, family 6, subfamily C, member 68
440193	CCDC88C	coiled-coil domain containing 88C
440822	PIWIL3	piwi-like 3 (Drosophila)
441151	TMEM151B	transmembrane protein 151B
643866	Cbln3	cerebellin 3 precursor
1E+08	SNORD121B	small nucleolar RNA, C/D box 121B
1E+08	CD300LD	CD300 molecule-like family member d

Supplementary material: The role of ARF6 in BA

(The following material is re-edited material from the published article "The role of ARF6 in Biliary Atresia" in PLOS ONE 2015. My work for this paper involved performing enrichment and network analyses as well as deriving the comprehensive mechanism of the role of ARF6 in BA based on the findings of the study. This work was important to justify including ARF6 in the proposed network of BA in this Chapter. The following material only highlights my work from multiple sections of this published study)

Method

From the network created in Cytoscape using protein-protein interactions, the first neighbor network using the genes that could be mapped from the top 1000 significant SNPs from the GWAS CHP cohort was created.

For SNP-to-gene mapping, +/- 20kb window was applied to the mapping file provided by the manufacturer for the Infinium HumanHap550K BeadChip. These first neighbor genes from the network were submitted to Ingenuity Pathway Analysis (IPA) for upstream regulator analysis and enrichment of canonical pathways and biological functional categories.

Result

299 unique genes were associated with 419 SNPs out of the 1000 top-ranked SNPs. 2506 genes were mapped as the first neighbor genes from the human protein-protein interaction network created in Cytoscape. Ingenuity's upstream regulator analysis on the first neighbor genes revealed that out of 632 potential upstream regulators, the 35th ranked regulator, sorted by enrichment p-value, was EGF (p=1.26E-7). Other significant EGF-related upstream regulators were TNF (p=3.14E-25) and ERK1/2 (p=3E-14). The significant canonical pathways were ERK/MAPK (p=5.74E-37) and cAMP-mediated signaling (p=1.5E-35) pathways while enriched functional categories were cellular proliferation (p=1.85E-103) and cellular development (p=1.63E-46).

Discussion

The proposed ARF6 mechanism based on the zebrafish phenotypes, GWAS analysis, and enrichment analysis starts with EGF binding to EGFR to activate GEP100-ARF6-ERK pathway. ARF6 is a critical member of EGFR

pathway and is activated by the binding of EGFR to GEP100. EGF is also a well-known upstream regulator of ERK/MAPK pathway. This EGFR-GEP100-ARF6-ERK pathway can be disrupted by AG1468, EGFR inhibitor, or the knockdown of ARF6. This disruption can cause downregulation of ERK/MAPK pathway and its several downstream transcription factors such as CREB and ELK1. CREB signaling, in particular, has been implicated to be important due to a couple of significant SNPs that map to the coding region of CREB-related genes from the BA GWAS data. Enriched cAMP-mediated signaling pathway from the enrichment analysis also supports the role of CREB signaling. With decreased activation of such transcription factors to mediate the proper level of transcription of their target genes, cellular proliferation and development suffer, which then leads to poor bile duct development with fibrosis and/or cirrhosis.

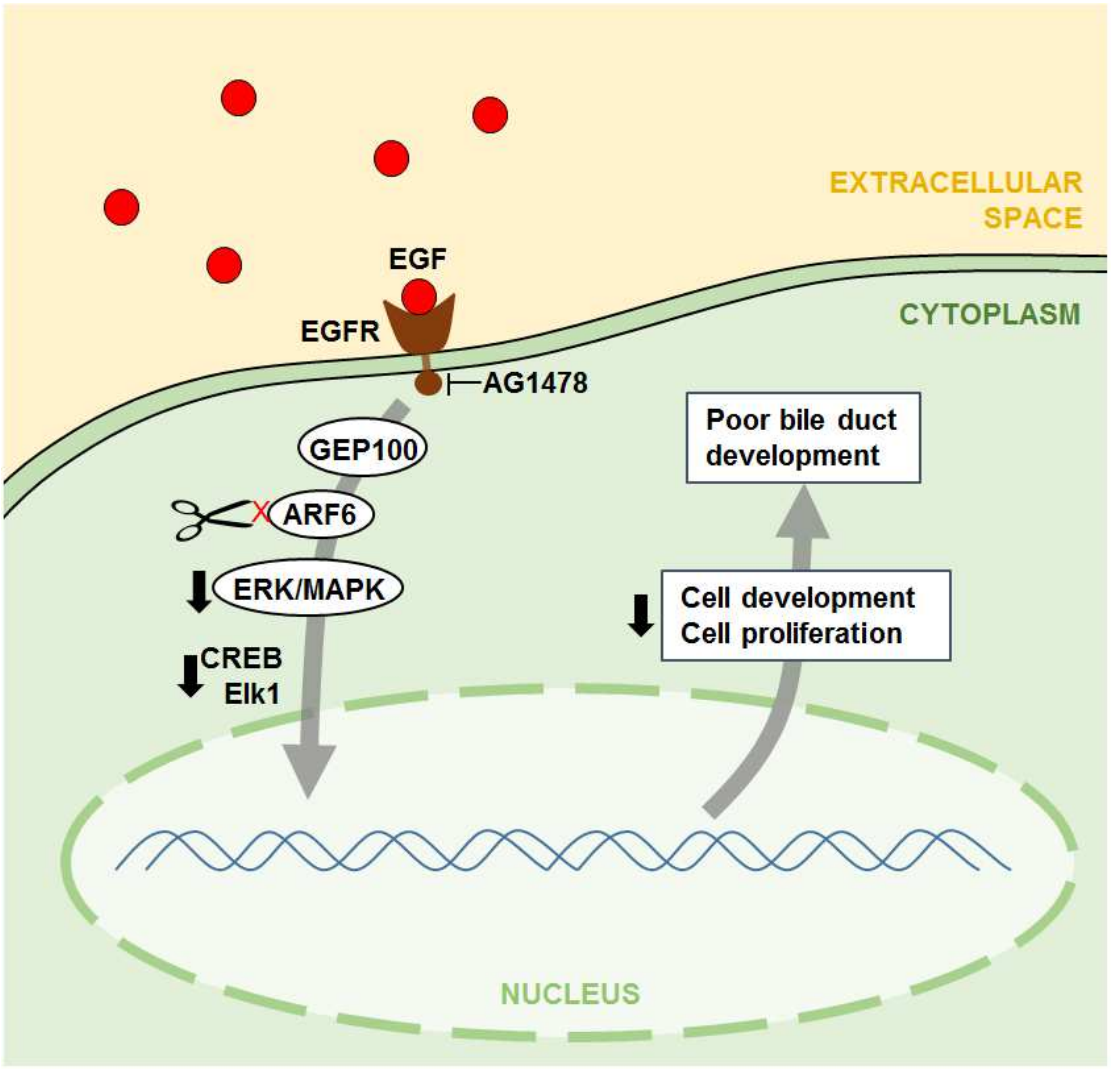


Figure S3.2 Proposed mechanism for the role of ARF6 in BA. EGF binds to EGFR to activate GEP100-ARF6-ERK pathway. EGF is also a well-known upstream regulator of ERK/MAPK pathway. This EGFR-GEP100-ARF6-ERK pathway can be disrupted by AG1468, EGFR inhibitor, or the knockdown of ARF6. This disruption can cause downregulation of ERK/MAPK pathway and its several downstream transcription factors such as CREB and ELK1. With decreased activation of such transcription factors to mediate the proper level of transcription of their target genes, cellular proliferation and development suffer, which then leads to poor bile duct development with fibrosis and/or cirrhosis.

Acknowledgements

Chapter 3, in part, is a re-editing of materials currently being prepared for submission for publication in Ningappa Mylarappa*, Jun Min*, Brandon Higgs, Qing Sun, Hakon Hakonarson, Donghun Shin, Shankar Subramaniam, and Rakesh Sindhi. Systems analysis of biliary atresia through integration of high-throughput biological data, **in preparation**. The dissertation author is the co-first author of this paper and was responsible for designing and performing all analytic methods and writing of the paper. Chapter 3, in part, is also a re-editing of materials published in Mylarappa Ningappa, Juhoon So, Joseph Glessner, Chethan Ashokkumar, Sarangarajan Ranganathan, Jun Min, Brandon W. Higgs, Qing Sun, Kimberly Haberman, Lori Schmitt, Silvia Vilarinho, Pramod K. Mistry, Gerard Vockley, Anil Dhawan, George K. Gittes, Hakon Hakonarson, Ronald Jaffe, Shankar Subramaniam, Donghun Shin, and Rakesh Sindhi. The role of ARF6 in biliary atresia. PLOS ONE. 2015. The dissertation author was responsible for analyzing data and performing systems biology methods to derive a comprehensive mechanism, which is depicted in a diagram.

CONCLUSION

The liver is a unique organ with a variety of functions and the ability to regenerate from a small injury. It has complex anatomy and physiology and is the second largest organ in the human body besides skin. Unfortunately, despite many years of investigation into this organ, we still lack a thorough understanding of different types of liver disease that lead to acute or chronic liver failure. Deriving the pathogenesis of a liver disease with complex phenotypes or molecular mechanisms has been a long-standing challenge and likely warrants a comprehensive systems biology approach. Such approach can integrate multiple analyses of high-throughput molecular datasets such as genomics, transcriptomics, proteomics, and metabolomics to allow unbiased screening and identification of significant biomarkers, pathways, and processes for the systems-level understanding of a disease.

In this dissertation, I used novel systems biology approaches to derive the mechanisms of liver regeneration and pathologies. In Chapter 1, I studied the early phase of liver regeneration and its relationship to the complement system. Liver regeneration is a unique process in the liver that allows mature hepatocytes to re-enter the cell cycle to proliferate to replace lost or damaged cells (15). During the priming phase, in particular, which occurs shortly after damage to the liver, a myriad of cellular signals are induced in hepatocytes to allow for the successful progression into the proliferation and the termination phase of regeneration (16). From the previous studies, the complement

system has been shown to promote liver regeneration but its comprehensive mechanism has not been fully investigated (27). Therefore, I have examined the role of the complement system on transcriptomic and metabolomic regulations during the priming phase of complement-induced liver regeneration by using C3-/- mice across multiple time points. I discovered that the complement system seems to activate c-fos and promote TNFα signaling pathway, which then activates acute phase genes such as SAAs and ORMs. The complement system also regulates cholesterol metabolism, which is important for cell cycle and proliferation and is monitored closely during liver regeneration. Finally, the complement system causes various metabolic changes towards the later priming phase of liver regeneration.

In Chapters 2 and 3, the goal was to understand the comprehensive mechanism for the development of BA. Biliary atresia (BA) is a rare, complex disease of the liver and the bile ducts with unknown etiology (17). To understand this disease, I have analyzed different types of high-throughput data such as GWAS, RNAseq, and target and whole exome sequencing. I have discovered several potential BA markers and enriched biological functions and pathways. By integrating multiple analytical results, I have also reconstructed the BA pathogenic network, which enabled a systems-level understanding of human BA biology that is highlighted by the interaction between key biological functions such as fibrosis, inflammation, immunity, and development. Through this network, I have also analyzed the relationship

between the genes involved in different biological functions within the pathogenic mechanism of BA.

The results from the dissertation, including the proposed mechanisms, significantly improve our understanding of the physiology of the liver. For example, the relationship between the complement system, immune response, and metabolism seems to be the key in allowing successful liver regeneration of a normal liver. This relationship, when explored more in detail, may facilitate therapeutic development to restore the regenerative capacity of the diseased livers. Furthermore, inflammatory, fibrotic, and developmental genes can interact to create a complex phenotype in liver such as BA that leads to difficult differential diagnosis and treatment. Each significant perturbation within the proposed network in Chapter 3 may be responsible for development of heterogeneous phenotypes in the BA patients. The findings from the dissertation suggest that intricate interactions between significant genomic, transcriptomic, and metabolic changes must be analyzed in depth to identify diverse biological functions and pathways that contribute to the physiology of the liver.

REFERENCES

1. Grada, A., and K. Weinbrecht. 2013. Next-generation sequencing: methodology and application. *The Journal of investigative dermatology* 133: e11.

2. Zhu, M., M. Yu, and S. Zhao. 2009. Understanding quantitative genetics in the systems biology era. *International journal of biological sciences* 5: 161-170.

3. Rao, V. S., K. Srinivas, G. N. Sujini, and G. N. Kumar. 2014. Protein-protein interaction detection: methods and analysis. *International journal of proteomics* 2014: 147648.

4. Huang da, W., B. T. Sherman, and R. A. Lempicki. 2009. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic acids research* 37: 1-13.

5. Chuang, H. Y., M. Hofree, and T. Ideker. 2010. A decade of systems biology. *Annual review of cell and developmental biology* 26: 721-744.

6. Mitra, S., S. Das, and J. Chakrabarti. 2013. Systems biology of cancer biomarker detection. *Cancer biomarkers : section A of Disease markers* 13: 201-213.

7. Tucker, T., M. Marra, and J. M. Friedman. 2009. Massively parallel sequencing: the next big thing in genetic medicine. *American journal of human genetics* 85: 142-154.

8. Wang, Z., M. Gerstein, and M. Snyder. 2009. RNA-Seq: a revolutionary tool for transcriptomics. *Nature reviews. Genetics* 10: 57-63.

9. Zhao, S., W. P. Fung-Leung, A. Bittner, K. Ngo, and X. Liu. 2014. Comparison of RNA-Seq and microarray in transcriptome profiling of activated T cells. *PloS one* 9: e78644.

10. Zhang, J., R. Chiodini, A. Badr, and G. Zhang. 2011. The impact of next-generation sequencing on genomics. *Journal of genetics and genomics = Yi chuan xue bao* 38: 95-109.

11. Oliva, J., B. A. French, X. Qing, and S. W. French. 2010. The identification of stem cells in human liver diseases and hepatocellular carcinoma. *Experimental and molecular pathology* 88: 331-340.

12. Zakim, D., and T. D. Boyer. 2003. *Hepatology : a textbook of liver disease*. Saunders, Philadelphia.

13. Czaja, A. J. 2014. Hepatic inflammation and progressive liver fibrosis in chronic liver disease. *World journal of gastroenterology* 20: 2515-2532.

14. Seto, W. K., C. L. Lai, and M. F. Yuen. 2012. Acute-on-chronic liver failure in chronic hepatitis B. *Journal of gastroenterology and hepatology* 27: 662-669.

15. Taub, R. 2004. Liver regeneration: from myth to mechanism. *Nature reviews. Molecular cell biology* 5: 836-847.

16. Su, A. I., L. G. Guidotti, J. P. Pezacki, F. V. Chisari, and P. G. Schultz. 2002. Gene expression during the priming phase of liver regeneration after partial hepatectomy in mice. *Proceedings of the National Academy of Sciences of the United States of America* 99: 11181-11186.

17. Petersen, C., and M. Davenport. 2013. Aetiology of biliary atresia: what is actually known? *Orphanet journal of rare diseases* 8: 128.

18. Clavien, P. A. 2008. Liver regeneration: a spotlight on the novel role of platelets and serotonin. *Swiss medical weekly* 138: 361-370.

19. Clavien, P. A., H. Petrowsky, M. L. DeOliveira, and R. Graf. 2007. Strategies for safer liver surgery and partial liver transplantation. *The New England journal of medicine* 356: 1545-1559.

20. Gotohda, N., H. Iwagaki, M. Ozaki, T. Kinoshita, M. Konishi, T. Nakagohri, S. Takahashi, S. Saito, T. Yagi, and N. Tanaka. 2008. Deficient response of IL-6 impaired liver regeneration after hepatectomy in patients with viral hepatitis. *Hepato-gastroenterology* 55: 1439-1444.

21. Hines, I. N., M. Kremer, F. Isayama, A. W. Perry, R. J. Milton, A. L. Black, C. L. Byrd, and M. D. Wheeler. 2007. Impaired liver regeneration and increased oval cell numbers following T cell-mediated hepatitis. *Hepatology* 46: 229-241.

22. Tanemura, A., S. Mizuno, H. Wada, T. Yamada, T. Nobori, and S. Isaji. 2012. Donor age affects liver regeneration during early period in the graft liver and late period in the remnant liver after living donor liver transplantation. *World journal of surgery* 36: 1102-1111.

23. Torbenson, M., S. Q. Yang, H. Z. Liu, J. Huang, W. Gage, and A. M. Diehl. 2002. STAT-3 overexpression and p21 up-regulation accompany impaired regeneration of fatty livers. *The American journal of pathology* 161: 155-161.

24. Yang, S. Q., H. Z. Lin, A. K. Mandal, J. Huang, and A. M. Diehl. 2001. Disrupted signaling and inhibited regeneration in obese mice with fatty livers: implications for nonalcoholic fatty liver disease pathophysiology. *Hepatology* 34: 694-706.

25. Michalopoulos, G. K. 2010. Liver regeneration after partial hepatectomy: critical analysis of mechanistic dilemmas. *The American journal of pathology* 176: 2-13.

26. Zimmermann, A. 2004. Regulation of liver regeneration. *Nephrology, dialysis, transplantation : official publication of the European Dialysis and Transplant Association - European Renal Association* 19 Suppl 4: iv6-10.

27. DeAngelis, R. A., M. M. Markiewski, and J. D. Lambris. 2006. Liver regeneration: a link to inflammation through complement. *Advances in experimental medicine and biology* 586: 17-34.

28. Grisham, J. W. 1962. A morphologic study of deoxyribonucleic acid synthesis and cell proliferation in regenerating rat liver; autoradiography with thymidine-H3. *Cancer research* 22: 842-849.

29. Nygard, I. E., K. E. Mortensen, J. Hedegaard, L. N. Conley, T. Kalstad, C. Bendixen, and A. Revhaug. 2012. The genetic regulation of the terminating phase of liver regeneration. *Comparative hepatology* 11: 3.

30. Higgins G, A. R. 1932. Experimental pathology of the liver 1. Restoration of the liver of the white rat following partial surgical removal. *Arch Pathol*: 186–202.

31. Slater Tf Fau - Cheeseman, K. H., K. U. Cheeseman Kh Fau - Ingold, and K. U. Ingold. Carbon tetrachloride toxicity as a model for studying free-radical mediated liver injury.

32. DeAngelis, R. A., M. M. Markiewski, I. Kourtzelis, S. Rafail, M. Syriga, A. Sandor, M. R. Maurya, S. Gupta, S. Subramaniam, and J. D. Lambris. 2012. A complement-IL-4 regulatory circuit controls liver regeneration. *Journal of immunology* 188: 641-648.

33. Mastellos, D., J. C. Papadimitriou, S. Franchini, P. A. Tsonis, and J. D. Lambris. 2001. A novel role of complement: mice deficient in the fifth component of complement (C5) exhibit impaired liver regeneration. *Journal of immunology* 166: 2479-2486.

34. Strey, C. W., M. Markiewski, D. Mastellos, R. Tudoran, L. A. Spruce, L. E. Greenbaum, and J. D. Lambris. 2003. The proinflammatory mediators C3a and C5a are essential for liver regeneration. *The Journal of experimental medicine* 198: 913-923.

35. Hu, J., H. Ge, M. Newman, and K. Liu. 2012. OSA: a fast and accurate alignment tool for RNA-Seq. *Bioinformatics* 28: 1933-1934.

36. Anders, S., P. T. Pyl, and W. Huber. 2015. HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics* 31: 166-169.

37. Anders, S., and W. Huber. 2010. Differential expression analysis for sequence count data. *Genome biology* 11: R106.

38. Spandidos, A., X. Wang, H. Wang, and B. Seed. 2010. PrimerBank: a resource of human and mouse PCR primer pairs for gene expression detection and quantification. *Nucleic acids research* 38: D792-799.

39. Wang, X., A. Spandidos, H. Wang, and B. Seed. 2012. PrimerBank: a PCR primer database for quantitative gene expression analysis, 2012 update. *Nucleic acids research* 40: D1144-1149.

40. Huang da, W., B. T. Sherman, and R. A. Lempicki. 2009. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature protocols* 4: 44-57.

41. Ogata, H., S. Goto, K. Sato, W. Fujibuchi, H. Bono, and M. Kanehisa. 1999. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic acids research* 27: 29-34.

42. Ashburner, M., C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature genetics* 25: 25-29.

43. Krzywinski, M., J. Schein, I. Birol, J. Connors, R. Gascoyne, D. Horsman, S. J. Jones, and M. A. Marra. 2009. Circos: an information aesthetic for comparative genomics. *Genome research* 19: 1639-1645.

44. Shannon, P., A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker. 2003. Cytoscape: a software

environment for integrated models of biomolecular interaction networks. *Genome research* 13: 2498-2504.

45. Stark, C., B. J. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz, and M. Tyers. 2006. BioGRID: a general repository for interaction datasets. *Nucleic acids research* 34: D535-539.

46. Szklarczyk, D., A. Franceschini, M. Kuhn, M. Simonovic, A. Roth, P. Minguez, T. Doerks, M. Stark, J. Muller, P. Bork, L. J. Jensen, and C. von Mering. 2011. The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic acids research* 39: D561-568.

47. von Mering, C., M. Huynen, D. Jaeggi, S. Schmidt, P. Bork, and B. Snel. 2003. STRING: a database of predicted functional associations between proteins. *Nucleic acids research* 31: 258-261.

48. Wingender, E., X. Chen, R. Hehl, H. Karas, I. Liebich, V. Matys, T. Meinhardt, M. Pruss, I. Reuter, and F. Schacherer. 2000. TRANSFAC: an integrated system for gene expression regulation. *Nucleic acids research* 28: 316-319.

49. Morris, J. H., L. Apeltsin, A. M. Newman, J. Baumbach, T. Wittkop, G. Su, G. D. Bader, and T. E. Ferrin. 2011. clusterMaker: a multi-algorithm clustering plugin for Cytoscape. *BMC bioinformatics* 12: 436.

50. Newman, A. M., and J. B. Cooper. 2010. AutoSOME: a clustering method for identifying gene expression modules without prior knowledge of cluster number. *BMC bioinformatics* 11: 117.

51. Passarelli, M. K., A. G. Ewing, and N. Winograd. 2013. C(60)-SIMS Studies of Glycerophospholipid in a LIPID MAPS Model System: KDO(2)-Lipid A Stimulated RAW 264.7 Cells. *Surface and interface analysis : SIA* 45: 298-301.

52. Sud, M., E. Fahy, D. Cotter, E. A. Dennis, and S. Subramaniam. 2012. LIPID MAPS-Nature Lipidomics Gateway: An Online Resource for Students and Educators Interested in Lipids. *Journal of chemical education* 89: 291-292.

53. Ingenuity Target Explorer. 2015. QIAGEN. https://targetexplorer.ingenuity.com/

54. Hu, Y. W., L. Zheng, and Q. Wang. 2010. Regulation of cholesterol homeostasis by liver X receptors. *Clinica chimica acta; international journal of clinical chemistry* 411: 617-625.

55. Chen, W., G. Chen, D. L. Head, D. J. Mangelsdorf, and D. W. Russell. 2007. Enzymatic reduction of oxysterols impairs LXR signaling in cultured cells and the livers of mice. *Cell metabolism* 5: 73-79.

56. Repa, J. J., K. E. Berge, C. Pomajzl, J. A. Richardson, H. Hobbs, and D. J. Mangelsdorf. 2002. Regulation of ATP-binding cassette sterol transporters ABCG5 and ABCG8 by the liver X receptors alpha and beta. *The Journal of biological chemistry* 277: 18793-18800.

57. Jo, Y., and R. A. Debose-Boyd. 2010. Control of cholesterol synthesis through regulated ER-associated degradation of HMG CoA reductase. *Critical reviews in biochemistry and molecular biology* 45: 185-198.

58. Phan, L. M., S. C. Yeung, and M. H. Lee. 2014. Cancer metabolic reprogramming: importance, main features, and potentials for precise targeted anti-cancer therapies. *Cancer biology & medicine* 11: 1-19.

59. Liang, H., and W. F. Ward. 2006. PGC-1alpha: a key regulator of energy metabolism. *Advances in physiology education* 30: 145-151.

60. Bartoloni, L., M. Wattenhofer, J. Kudoh, A. Berry, K. Shibuya, K. Kawasaki, J. Wang, S. Asakawa, I. Talior, B. Bonne-Tamir, C. Rossier, J. Michaud, E. R. McCabe, S. Minoshima, N. Shimizu, H. S. Scott, and S. E. Antonarakis. 2000. Cloning and characterization of a putative

human glycerol 3-phosphate permease gene (SLC37A1 or G3PP) on 21q22.3: mutation analysis in two candidate phenotypes, DFNB10 and a glycerol kinase deficiency. *Genomics* 70: 190-200.

61. Moh, A., Y. Iwamoto, G. X. Chai, S. S. Zhang, A. Kano, D. D. Yang, W. Zhang, J. Wang, J. J. Jacoby, B. Gao, R. A. Flavell, and X. Y. Fu. 2007. Role of STAT3 in liver regeneration: survival, DNA synthesis, inflammatory reaction and liver mass recovery. *Laboratory investigation; a journal of technical methods and pathology* 87: 1018-1028.

62. Riehle, K. J., J. S. Campbell, R. S. McMahan, M. M. Johnson, R. P. Beyer, T. K. Bammler, and N. Fausto. 2008. Regulation of liver regeneration and hepatocarcinogenesis by suppressor of cytokine signaling 3. *The Journal of experimental medicine* 205: 91-103.

63. Thorn, C. F., Z. Y. Lu, and A. S. Whitehead. 2004. Regulation of the human acute phase serum amyloid A genes by tumour necrosis factor-alpha, interleukin-6 and glucocorticoids in hepatic and epithelial cell lines. *Scandinavian journal of immunology* 59: 152-158.

64. Westra, J., J. Bijzet, B. Doornbos-van der Meer, M. H. van Rijswijk, and P. C. Limburg. 2006. Differential influence of p38 mitogen activated protein kinase (MAPK) inhibition on acute phase protein synthesis in human hepatoma cell lines. *Annals of the rheumatic diseases* 65: 929-935.

65. Lemmers, A., T. Gustot, A. Durnez, S. Evrard, C. Moreno, E. Quertinmont, V. Vercruysse, P. Demetter, D. Franchimont, O. Le Moine, A. Geerts, and J. Deviere. 2009. An inhibitor of interleukin-6 trans-signalling, sgp130, contributes to impaired acute phase response in human chronic liver disease. *Clinical and experimental immunology* 156: 518-527.

66. Sasaki, M., N. Yoneda, S. Kitamura, Y. Sato, and Y. Nakanuma. 2012. A serum amyloid A-positive hepatocellular neoplasm arising in alcoholic cirrhosis: a previously unrecognized type of inflammatory hepatocellular tumor. *Modern pathology : an official journal of the United States and Canadian Academy of Pathology, Inc* 25: 1584-1593.

67. Ryden, I., P. Pahlsson, and S. Lindgren. 2002. Diagnostic accuracy of alpha(1)-acid glycoprotein fucosylation for liver cirrhosis in patients undergoing hepatic biopsy. *Clinical chemistry* 48: 2195-2201.

68. O'Brien, K. D., and A. Chait. 2006. Serum amyloid A: the "other" inflammatory protein. *Current atherosclerosis reports* 8: 62-68.

69. Coetzee, G. A., A. F. Strachan, D. R. van der Westhuyzen, H. C. Hoppe, M. S. Jeenah, and F. C. de Beer. 1986. Serum amyloid A-containing human high density lipoprotein 3. Density, size, and apolipoprotein composition. *The Journal of biological chemistry* 261: 9644-9651.

70. de Beer, M. C., N. R. Webb, J. M. Wroblewski, V. P. Noffsinger, D. L. Rateri, A. Ji, D. R. van der Westhuyzen, and F. C. de Beer. 2010. Impact of serum amyloid A on high density lipoprotein composition and levels. *Journal of lipid research* 51: 3117-3125.

71. Artl, A., G. Marsche, S. Lestavel, W. Sattler, and E. Malle. 2000. Role of serum amyloid A during metabolism of acute-phase HDL by macrophages. *Arteriosclerosis, thrombosis, and vascular biology* 20: 763-772.

72. Jahangiri, A., M. C. de Beer, V. Noffsinger, L. R. Tannock, C. Ramaiah, N. R. Webb, D. R. van der Westhuyzen, and F. C. de Beer. 2009. HDL remodeling during the acute phase response. *Arteriosclerosis, thrombosis, and vascular biology* 29: 261-267.

73. Kim, K. H., G. Y. Lee, J. I. Kim, M. Ham, J. Won Lee, and J. B. Kim. 2010. Inhibitory effect of LXR activation on cell proliferation and cell cycle progression through lipogenic activity. *Journal of lipid research* 51: 3425-3433.

74. Peraldi, P., G. S. Hotamisligil, W. A. Buurman, M. F. White, and B. M. Spiegelman. 1996. Tumor necrosis factor (TNF)-alpha inhibits insulin signaling through stimulation of the p55 TNF receptor and activation of sphingomyelinase. *The Journal of biological chemistry* 271: 13018-13022.

75. Osborne, A. R., V. V. Pollock, W. R. Lagor, and G. C. Ness. 2004. Identification of insulin-responsive regions in the HMG-CoA reductase promoter. *Biochemical and biophysical research communications* 318: 814-818.

76. Jones, R. G., and C. B. Thompson. 2009. Tumor suppressors and cell metabolism: a recipe for cancer growth. *Genes & development* 23: 537-548.

77. Malarkey, D. E., K. Johnson, L. Ryan, G. Boorman, and R. R. Maronpot. 2005. New insights into functional aspects of liver morphology. *Toxicologic pathology* 33: 27-34.

78. Kmiec, Z. 2001. Cooperation of liver cells in health and disease. *Advances in anatomy, embryology, and cell biology* 161: III-XIII, 1-151.

79. Nipic, D., A. Pirc, B. Banic, D. Suput, and I. Milisav. 2010. Preapoptotic cell stress response of primary hepatocytes. *Hepatology* 51: 2140-2151.

80. Auger, J. L., S. Haasken, and B. A. Binstadt. 2012. Autoantibody-mediated arthritis in the absence of C3 and activating Fcgamma receptors: C5 is activated by the coagulation cascade. *Arthritis research & therapy* 14: R269.

81. Saggu, G., C. Cortes, H. N. Emch, G. Ramirez, R. G. Worth, and V. P. Ferreira. 2013. Identification of a novel mode of complement activation on stimulated platelets mediated by properdin and C3(H2O). *Journal of immunology* 190: 6457-6467.

82. Hsiao, C. H., M. H. Chang, H. L. Chen, H. C. Lee, T. C. Wu, C. C. Lin, Y. J. Yang, A. C. Chen, M. M. Tiao, B. H. Lau, C. H. Chu, M. W. Lai, and G. Taiwan Infant Stool Color Card Study. 2008. Universal screening for biliary atresia using an infant stool color card in Taiwan. *Hepatology* 47: 1233-1240.

83. Engelmann, G., J. Schmidt, J. Oh, H. Lenhartz, D. Wenning, U. Teufel, M. W. Buchler, G. F. Hoffmann, and J. Meyburg. 2007. Indications for pediatric liver transplantation. Data from the Heidelberg pediatric liver transplantation program. *Nephrology, dialysis, transplantation : official publication of the European Dialysis and Transplant Association - European Renal Association* 22 Suppl 8: viii23-viii28.

84. Petersen, C., D. Harder, Z. Abola, D. Alberti, T. Becker, C. Chardot, M. Davenport, A. Deutschmann, K. Khelif, H. Kobayashi, N. Kvist, J. Leonhardt, M. Melter, M. Pakarinen, J. Pawlowska, A. Petersons, E. D. Pfister, M. Rygl, R. Schreiber, R. Sokol, B. Ure, C. Veiga, H. Verkade, B. Wildhaber, B. Yerushalmi, and D. Kelly. 2008. European biliary atresia registries: summary of a symposium. *European journal of pediatric surgery : official journal of Austrian Association of Pediatric Surgery ... [et al] = Zeitschrift fur Kinderchirurgie* 18: 111-116.

85. Mack, C. L. 2007. The pathogenesis of biliary atresia: evidence for a virus-induced autoimmune disease. *Seminars in liver disease* 27: 233-242.

86. Lorent, K., W. Gong, K. A. Koo, O. Waisbourd-Zinman, S. Karjoo, X. Zhao, I. Sealy, R. N. Kettleborough, D. L. Stemple, P. A. Windsor, S. J. Whittaker, J. R. Porter, R. G. Wells, and M. Pack. Identification of a plant isoflavonoid that causes biliary atresia.

87. Honsawek, S., V. Chongsrisawat, P. Vejchapipat, N. Thawornsuk, P. Tangkijvanich, and Y. Poovorawan. 2005. Serum interleukin-8 in children with biliary atresia: relationship with disease stage and biochemical parameters. *Pediatric surgery international* 21: 73-77.

88. Ningappa, M., J. Min, B. W. Higgs, C. Ashokkumar, S. Ranganathan, and R. Sindhi. 2015. Genome-wide association studies in biliary atresia. *Wiley interdisciplinary reviews. Systems biology and medicine* 7: 267-273.

89. Bates, M. D., J. C. Bucuvalas, M. H. Alonso, and F. C. Ryckman. 1998. Biliary atresia: pathogenesis and treatment. *Seminars in liver disease* 18: 281-293.

90. Kelly, D. A., and M. Davenport. 2007. Current management of biliary atresia. *Archives of disease in childhood* 92: 1132-1135.

91. Bessho, K., and J. A. Bezerra. 2011. Biliary atresia: will blocking inflammation tame the disease? *Annual review of medicine* 62: 171-185.

92. Feldman, A. G., and C. L. Mack. 2012. Biliary atresia: cellular dynamics and immune dysregulation. *Seminars in pediatric surgery* 21: 192-200.

93. Caponcelli, E., A. S. Knisely, and M. Davenport. 2008. Cystic biliary atresia: an etiologic and prognostic subgroup. *Journal of pediatric surgery* 43: 1619-1624.

94. Chu, A. S., P. A. Russo, and R. G. Wells. 2012. Cholangiocyte cilia are abnormal in syndromic and non-syndromic biliary atresia. *Modern pathology : an official journal of the United States and Canadian Academy of Pathology, Inc* 25: 751-757.

95. Schwarz, K. B., B. H. Haber, P. Rosenthal, C. L. Mack, J. Moore, K. Bove, J. A. Bezerra, S. J. Karpen, N. Kerkar, B. L. Shneider, Y. P. Turmelle, P. F. Whitington, J. P. Molleston, K. F. Murray, V. L. Ng, R. Romero, K. S. Wang, R. J. Sokol, J. C. Magee, R. Childhood Liver Disease, and N. Education. 2013. Extrahepatic anomalies in infants with biliary atresia: results of a large prospective North American multicenter study. *Hepatology* 58: 1724-1731.

96. Bezerra, J. A., C. Spino, J. C. Magee, B. L. Shneider, P. Rosenthal, K. S. Wang, J. Erlichman, B. Haber, P. M. Hertel, S. J. Karpen, N. Kerkar, K. M. Loomes, J. P. Molleston, K. F. Murray, R. Romero, K. B. Schwarz, R. Shepherd, F. J. Suchy, Y. P. Turmelle, P. F. Whitington, J. Moore, A. H. Sherker, P. R. Robuck, and R. J. Sokol. Use of corticosteroids after hepatoportoenterostomy for bile drainage in infants with biliary atresia: the START randomized clinical trial.

97. Mi, H., B. Lazareva-Ulitsky, R. Loo, A. Kejariwal, J. Vandergriff, S. Rabkin, N. Guo, A. Muruganujan, O. Doremieux, M. J. Campbell, H. Kitano, and P. D. Thomas. 2005. The PANTHER database of protein

families, subfamilies, functions and pathways. *Nucleic acids research* 33: D284-288.

98. D., N. BioCarta. Biotech Softw. Internet Rep. 2001;2:117–120. doi: 10.1089/152791601750294344.

99. Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M. A. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. de Bakker, M. J. Daly, and P. C. Sham. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. *American journal of human genetics* 81: 559-575.

100. Nica, A. C., and E. T. Dermitzakis. 2013. Expression quantitative trait loci: present and future. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* 368: 20120362.

101. Schadt, E. E., C. Molony, E. Chudin, K. Hao, X. Yang, P. Y. Lum, A. Kasarskis, B. Zhang, S. Wang, C. Suver, J. Zhu, J. Millstein, S. Sieberts, J. Lamb, D. GuhaThakurta, J. Derry, J. D. Storey, I. Avila-Campillo, M. J. Kruger, J. M. Johnson, C. A. Rohl, A. van Nas, M. Mehrabian, T. A. Drake, A. J. Lusis, R. C. Smith, F. P. Guengerich, S. C. Strom, E. Schuetz, T. H. Rushmore, and R. Ulrich. 2008. Mapping the genetic architecture of gene expression in human liver. *PLoS biology* 6: e107.

102. Xu, Z., and J. A. Taylor. 2009. SNPinfo: integrating GWAS and candidate gene information into functional SNP selection for genetic association studies. *Nucleic acids research* 37: W600-605.

103. Yourshaw, M., S. P. Taylor, A. R. Rao, M. G. Martin, and S. F. Nelson. 2015. Rich annotation of DNA sequencing variants by leveraging the Ensembl Variant Effect Predictor with plugins. *Briefings in bioinformatics* 16: 255-264.

104. Farrington, C., D. Novak, C. Liu, and A. B. Haafiz. 2010. Immunohistochemical localization of transforming growth factor beta-1 and its relationship with collagen expression in advanced liver fibrosis due to biliary atresia. *Clinical and experimental gastroenterology* 3: 185-191.

105. Wells, R. G. 2014. Portal Fibroblasts in Biliary Fibrosis. *Current pathobiology reports* 2: 185-190.

106. Cant, N., N. Pollock, and R. C. Ford. 2014. CFTR structure and cystic fibrosis. *The international journal of biochemistry & cell biology* 52: 15-25.

107. Ricklin, D., G. Hajishengallis, K. Yang, and J. D. Lambris. 2010. Complement: a key system for immune surveillance and homeostasis. *Nature immunology* 11: 785-797.

108. Nangaku, M., J. Pippin, and W. G. Couser. 1999. Complement membrane attack complex (C5b-9) mediates interstitial disease in experimental nephrotic syndrome. *Journal of the American Society of Nephrology : JASN* 10: 2323-2331.

109. Ngure, R. M., P. D. Eckersall, N. K. Mungatana, J. N. Mburu, F. W. Jennings, J. Burke, and M. Murray. 2009. Lipopolysaccharide binding protein in the acute phase response of experimental murine Trypanosoma brucei brucei infection. *Research in veterinary science* 86: 394-398.

110. Schroedl, W., B. Fuerll, P. Reinhold, M. Krueger, and C. Schuett. 2001. A novel acute phase marker in cattle: lipopolysaccharide binding protein (LBP). *Journal of endotoxin research* 7: 49-52.

111. Mikkaichi, T., T. Suzuki, T. Onogawa, M. Tanemoto, H. Mizutamari, M. Okada, T. Chaki, S. Masuda, T. Tokui, N. Eto, M. Abe, F. Satoh, M. Unno, T. Hishinuma, K. Inui, S. Ito, J. Goto, and T. Abe. 2004. Isolation and characterization of a digoxin transporter and its rat homologue expressed in the kidney. *Proceedings of the National Academy of Sciences of the United States of America* 101: 3569-3574.

112. Levin, M. C., M. Monetti, M. J. Watt, M. P. Sajan, R. D. Stevens, J. R. Bain, C. B. Newgard, R. V. Farese, Sr., and R. V. Farese, Jr. 2007. Increased lipid accumulation and insulin resistance in transgenic mice expressing DGAT2 in glycolytic (type II) muscle. *American journal of physiology. Endocrinology and metabolism* 293: E1772-1781.

113. Eaton, S. 2008. Multiple roles for lipids in the Hedgehog signalling pathway. *Nature reviews. Molecular cell biology* 9: 437-445.

114. Markiewski, M. M., B. Nilsson, K. N. Ekdahl, T. E. Mollnes, and J. D. Lambris. 2007. Complement and coagulation: strangers or partners in crime? *Trends in immunology* 28: 184-192.

115. Mercer, P. F., and R. C. Chambers. 2013. Coagulation and coagulation signalling in fibrosis. *Biochimica et biophysica acta* 1832: 1018-1027.

116. Spencer, C. C., Z. Su, P. Donnelly, and J. Marchini. 2009. Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip. *PLoS genetics* 5: e1000477.

117. Park, A. K., and H. Kim. 2007. [A review of power and sample size estimation in genomewide association studies]. *Journal of preventive medicine and public health = Yebang Uihakhoe chi* 40: 114-121.

118. Ningappa, M., J. So, J. Glessner, C. Ashokkumar, S. Ranganathan, J. Min, B. W. Higgs, Q. Sun, K. Haberman, L. Schmitt, S. Vilarinho, P. K. Mistry, G. Vockley, A. Dhawan, G. K. Gittes, H. Hakonarson, R. Jaffe, S. Subramaniam, D. Shin, and R. Sindhi. 2015. The Role of ARF6 in Biliary Atresia. *PloS one* 10: e0138381.

119. Li, H., and R. Durbin. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25: 1754-1760.

120. McKenna, A., M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, and M. A. DePristo. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research* 20: 1297-1303.

121. Cingolani, P., A. Platts, L. Wang le, M. Coon, T. Nguyen, L. Wang, S. J. Land, X. Lu, and D. M. Ruden. 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs

in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly* 6: 80-92.

122. Andrews, S. Andrews S. 2010. FastQC: a quality control tool for high throughput sequence data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc.

123. Thorvaldsdottir, H., J. T. Robinson, and J. P. Mesirov. 2013. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in bioinformatics* 14: 178-192.

124. PICARD. 2015. Broad Institute. http://broadinstitute.github.io/picard/.

125. Barrett, J. C., B. Fry, J. Maller, and M. J. Daly. 2005. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* 21: 263-265.

126. Genomes Project, C., G. R. Abecasis, A. Auton, L. D. Brooks, M. A. DePristo, R. M. Durbin, R. E. Handsaker, H. M. Kang, G. T. Marth, and G. A. McVean. 2012. An integrated map of genetic variation from 1,092 human genomes. *Nature* 491: 56-65.

127. Liu, X., X. Jian, and E. Boerwinkle. 2011. dbNSFP: a lightweight database of human nonsynonymous SNPs and their functional predictions. *Human mutation* 32: 894-899.

128. IPA. 2015. QIAGEN Redwood City. www.qiagen.com/ingenuity.

129. van Dam, T. J., G. Wheway, G. G. Slaats, S. S. Group, M. A. Huynen, and R. H. Giles. 2013. The SYSCILIA gold standard (SCGSv1) of known ciliary components and its applications within a systems biology consortium. *Cilia* 2: 7.

130. Maere, S., K. Heymans, and M. Kuiper. 2005. BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics* 21: 3448-3449.

131. Zhang, L., X. Peng, Z. Zhang, Y. Feng, X. Jia, Y. Shi, H. Yang, Z. Zhang, X. Zhang, L. Liu, L. Yin, and Z. Yuan. 2010. Subcellular proteome analysis unraveled annexin A2 related to immune liver fibrosis. *Journal of cellular biochemistry* 110: 219-228.

132. Martinu, L., J. M. Masuda-Robens, S. E. Robertson, L. C. Santy, J. E. Casanova, and M. M. Chou. 2004. The TBC (Tre-2/Bub2/Cdc16) domain protein TRE17 regulates plasma membrane-endosomal trafficking through activation of Arf6. *Molecular and cellular biology* 24: 9752-9762.

133. Sasaki, M., S. B. Nakanuma Y Fau - Ho, Y. S. Ho Sb Fau - Kim, and Y. S. Kim. 1998. Increased MUC6 apomucin expression is a characteristic of reactive biliary epithelium in chronic viral hepatitis. *J Pathol* 185(2):191-8.

134. Leask, A., and D. J. Abraham. 2004. TGF-beta signaling and the fibrotic response. *FASEB journal : official publication of the Federation of American Societies for Experimental Biology* 18: 816-827.

135. Shimadera, S., N. Iwai, E. Deguchi, O. Kimura, S. Fumino, and T. Yokoyama. 2007. The inv mouse as an experimental model of biliary atresia. *Journal of pediatric surgery* 42: 1555-1560.

136. Pulina, M., D. Liang, and S. Astrof. 2014. Shape and position of the node and notochord along the bilateral plane of symmetry are regulated by cell-extracellular matrix interactions. *Biology open* 3: 583-590.

137. Hirokawa, N., Y. Tanaka, and Y. Okada. 2009. Left-right determination: involvement of molecular motor KIF3, cilia, and nodal flow. *Cold Spring Harbor perspectives in biology* 1: a000802.

138. Wang, G., Q. Chen, X. Zhang, B. Zhang, X. Zhuo, J. Liu, Q. Jiang, and C. Zhang. 2013. PCM1 recruits Plk1 to the pericentriolar matrix to promote primary cilia disassembly before mitotic entry. *Journal of cell science* 126: 1355-1365.

139. Zhou, J., P. S. Nagarkatti, Y. Zhong, J. Zhang, and M. Nagarkatti. 2011. Implications of single nucleotide polymorphisms in CD44 exon 2 for risk of breast cancer. *European journal of cancer prevention : the official journal of the European Cancer Prevention Organisation* 20: 396-402.

140. Guo, Y., R. Costa, H. Ramsey, T. Starnes, G. Vance, K. Robertson, M. Kelley, R. Reinbold, H. Scholer, and R. Hromas. 2002. The embryonic stem cell transcription factors Oct-4 and FoxD3 interact to regulate endodermal-specific promoter expression. *Proceedings of the National Academy of Sciences of the United States of America* 99: 3663-3667.

141. Sandberg, M. L., S. E. Sutton, M. T. Pletcher, T. Wiltshire, L. M. Tarantino, J. B. Hogenesch, and M. P. Cooke. 2005. c-Myb and p300 regulate hematopoietic stem cell proliferation and differentiation. *Developmental cell* 8: 153-166.

142. Hess, J., P. Angel, and M. Schorpp-Kistner. 2004. AP-1 subunits: quarrel and harmony among siblings. *Journal of cell science* 117: 5965-5973.

143. Kido, O., K. Fukushima, Y. Ueno, J. Inoue, D. M. Jefferson, and T. Shimosegawa. 2009. Compensatory role of inducible annexin A2 for impaired biliary epithelial anion-exchange activity of inflammatory cholangiopathy. *Laboratory investigation; a journal of technical methods and pathology* 89: 1374-1386.