Spring 2017

Network-based approaches to studying healthy and disease development

Long Gao
University of Iowa

Network-based Approaches to Studying Healthy and Disease Development

by

Long Gao

A thesis submitted in partial fulfillment of the
requirements for the Doctor of
Philosophy degree in Biomedical Engineering in
the Graduate College of
The University of Iowa

May 2017

Thesis Supervisor: Associate Professor Kai Tan

Graduate College
The University of Iowa
Iowa City, Iowa

CERTIFICATE OF APPROVAL

PH.D. THESIS

This is to certify that the Ph.D. thesis of

Long Gao

has been approved by the Examining Committee for the thesis requirement for the Doctor of Philosophy degree in Biomedical Engineering at the May 2017 graduation.

Thesis Committee: Kai Tan, Thesis Supervisor

Terry Braun

Michael Mackey

Kai Wang

Xiaodong Wu

To my family and friends, for their support

ACKNOWLEDGMENTS

First, I would like to thank my thesis advisor, Dr. Kai Tan, for his mentorship and support during the past six years. Kai has provided me with excellent guidance and encouragement throughout my graduate training. He has invested a large amount of time in my individual success and I will always be grateful for his patience and dedication to my development. Under his advice and support I was able to develop my skills as an independent computational biologist.

I would like to thank my committee members, Drs. Terry Braun, Michael Mackey, Kai Wang and Xiaodong Wu, who have willingly set aside time to provide guidance, encouragement and inspiration during my committee meetings. They clearly consider mentoring to be a priority and I am fortunate to have benefited from their professional insights. I also thank our collaborators Drs. Rong Tian, Chi Fung Lee, Nancy Speck, Joanna Tober, and Yan Li, who provided valuable experimental data and thoughtful suggestions.

I am grateful for the help and support from the Department of Biomedical Engineering, which provided many opportunities for professional advancement during my time as a graduate student. In particular, many thanks to Lobb Joshua, who works hard to ensure the administrative needs of graduate students are met in a timely manner. He is a pleasure to work with and I greatly appreciate all of his help.

I would like to express my gratitude to the past and current Tan lab members for providing me with technical expertise and insightful advice during past six years. I especially thank Drs. Jongkwang Kim, Teng Li, Xiaoke Ma, Peng Gao and Yasin Uzun,

for their contribution to my thesis projects. Special thanks go to Lucas Van Tol, who helped me solve countless technical problems in the past six years.

Lastly, I thank my family and friends. Without their support, I would not have accomplished my graduate school goals.

ABSTRACT

Network biology has proven to be a powerful tool for representing and analyzing complex bio-molecular networks. However, there are currently a number of challenges. First, network construction is a prerequisite of network analysis. When the number of samples is limited, state-of-the-art computational methods for network construction are not robust. Second, our current knowledge about the dynamics of molecular networks during disease progression is still limited. Finally, molecular networks have been extensively used to improve the inference accuracy of causal coding variants, but this potential has not been investigated to the same extent for noncoding variants.

To address those challenges, I first developed the inference of Multiple Differential Modules (*i*MDM) algorithm to study network dynamics. This method is able to identify both unique and shared modules from multiple gene co-expression networks, each of which denotes a different perturbation condition. Using *i*MDM algorithm, I identified different types of modules to understand heart failure progression and disease dynamics.

Next, I developed a computational framework to construct condition-specific transcriptional regulatory networks. I also developed a computational method to rank transcription factors in the transcriptional regulatory network using differential gene expression. Applying this framework to RNA-Seq data for hematopoietic stem cell development, I successfully constructed transcriptional regulatory networks and identified key transcriptional factors that play important roles during embryonic hematopoiesis.

Finally, I developed the Annotation of Regulatory Variants using Integrated Networks (ARVIN) algorithm, to identify causal noncoding variants for diseases. By applying ARVIN to seven autoimmune diseases, I obtained a systems understanding of the gene circuitry that is affected by all enhancer mutations in a given disease.

PUBLIC ABSTRACT

Genes and proteins often work together in an intricate network rather than acting in isolation. These biological networks contain abundant information revealing the overall physical and functional landscape of a biological system. Network analysis has been demonstrated as a powerful approach to studying biological phenomena because it provides a global picture of molecular interactions in different cell types and disease states. Existing network analysis methods mostly rely on mining protein-protein interaction networks, transcriptional regulatory networks (TRNs) or gene co-expression networks (Aittokallio and Schwikowski, 2006; Bebek and Yang, 2007; Gitter et al., 2011; Huang and Fraenkel, 2009; Ourfali et al., 2007). Both healthy development and disease progression are driven by dynamic changes in both the activity and connectivity of gene pathways, and network biology provides powerful tools for studying such dynamic changes (Cho et al., 2012).

Currently, there is a lack of computational methods that enable analysis of multiple gene networks to understand the dynamic events. In addition, many computational methods require a large number of gene expression profiles to construct network models. Unfortunately, the number of samples for particular conditions is usually not enough for these methods. Network analysis has also been applied to infer causal genetic variants for diseases (Lee et al., 2009; Zhang et al., 2013a). Although molecular networks can improve the inference accuracy of causal coding variants, their utility has not been examined for causal non-coding variants (Jia et al., 2011; Lee et al., 2011; Linghu et al., 2009; Moreau and Tranchevent, 2012). To address those problems, I developed 3 network-based methods to: **1)** identify dynamic events across multiple gene

networks during healthy development and disease progression; **2)** construct condition-specific gene networks with a limited number of samples; and **3)** infer causal non-coding variants for human diseases.

CONTRIBUTIONS

CHAPTER 1: Long Gao wrote Chapter 1.

CHAPTER 2: Xiaoke Ma, Long Gao and Kai Tan designed the *i*MDM algorithm. Chi Fung Lee performed the experiments and generated the experimental data. Long Gao, Xiaoke Ma and Kai Tan analyzed the data. Long Gao wrote the content in Chapter 2. Long Gao is the second author of the paper "Revealing pathway dynamics in heart diseases by analyzing multiple different networks." (Published June, 2015 in PLOS computational biology)

CHAPTER 3: Long Gao, Kai Tan and Nancy Speck designed the research. Joanna Tober and Yan Li performed the experiments and generated the experimental data. Long Gao and Kai Tan developed the novel algorithm for TRN construction. Long Gao performed data analysis and wrote Chapter 3. Long Gao is the co-first author of the paper "Transcriptional regulatory networks during the endothelial-to-hematopoietic transition in the mouse embryo." (In preparation)

CHAPTER 4: Long Gao, Yasin Uzun and Kai Tan designed the research. Long Gao, Yasin Uzun and Kai Tan developed the ARVIN algorithm. Long Gao and Yasin Uzun performed data analysis. Peng Gao did the experimental validation. Long Gao wrote the content in Chapter 4. Long Gao is the co-first author of the paper "Prioritizing Risk Genetic Variants in Regulatory DNA Sequences Using Disease-relevant Gene Regulatory Networks." (In preparation)

CHAPTER 5: Long Gao wrote Chapter 5.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

3C Chromosome conformation capture

4C Circular chromosome conformation capture

5C Chromosome conformation capture carbon copy

AGM Aorta gonad mesonephros

BM Bone marrow

ChIA-PET chromatin interaction analysis by paired-end tag sequencing

ChIP-Seq Chromatin immunoprecipitation followed by high-throughput sequencing

ENCODE Encyclopedia of DNA Elements

EP Enhancer-Promoter

ES Embryonic stem

eQTL Expression quantitative trait loci

FDR False Discovery Rate

FL Fetal liver

GTEx Genotype-tissue expression project

GWAS Genome wide association studies

HE Hemogenic endothelium

HGMD Human Gene Mutation Database

HP Hematopoietic progenitors

HSC Hematopoietic stem cell

LD Linkage Disequilibrium

NGS Next-generation sequencing

PCST Prize-Collecting Steiner Tree algorithm

PIN	Protein interaction network
PPI	Protein-protein interactions
qPCR	Quantitative polymerase chain reaction
RF	Random Forest
RFE	Recursive Feature Elimination
ROC	Receiver Operating Characteristic
SNP	Single-nucleotide polymorphism
TAC	Traverse aortic constriction
TF	Transcription factor
TRNs	Transcriptional Regulatory Networks
TSS	Transcription start site
WGS	Whole-genome sequencing
YS	Yolk sac

CHAPTER 1: INTRODUCTION

1.1 Overview of network biology

Network approaches have been utilized to model interactions between entities of interest in various areas. At a highly abstract level, most complex systems such as the cell, society and the Internet can be modeled as networks. The components of these systems are represented as a collection of nodes that are connected to each other by edges, with each edge representing the interactions between two nodes (Figure 1A).

Network representations are useful to analyze and visualize complex biological systems. For biological networks, nodes often denote genes/proteins and edges denote interactions between genes/proteins. A key goal of network analysis is to find the structural and functional building blocks of the networks, which are often represented as "modules". Gene modules are groups of genes or gene products that are functionally coordinated, physically interacting or co-regulated. Complex diseases are caused by a combination of genetic factors. Network-based approaches can leverage the idea that complex diseases can be better understood from the perspective of dys-regulated gene modules rather than individual genes (Cho et al., 2012).

1.1.1 Types of biological networks

1.1.1.1 Protein interaction network

In protein interaction networks (PINs), proteins are considered as being physically interacted with each other (Figure 1B). Protein-protein interactions (PPIs) are essential to almost every process in a cell, so understanding PPIs is crucial for understanding cell physiology in normal and disease states. Additionally, a disease is usually the result of complex interactions and perturbations involving many genes rather than a single gene.

There are multiple experimental methods that can be used to detect PPIs including yeast two-hybrid screening and affinity purification coupled to mass spectrometry (Bruckner et al., 2009; Vasilescu et al., 2004). PPIs can also be computationally inferred by techniques such as text mining and machine learning (Quan et al., 2014; Saetre et al., 2010).

Large-scale PPI data sets across different species have been collected and stored in specialized biological databases. Those databases can be classified into three categories: primary databases, meta-databases and prediction databases (De Las Rivas and Fontanillo, 2010). In primary databases, interactions are collected from published work that is proven to exist using experimental methods (Chen et al., 2011). Many primary databases have been built such as Bimolecular Interaction Network Database (BIND) (Bader et al., 2003), Biological General Repository for Interaction Datasets (BioGRID) (Stark et al., 2006), Human Protein Reference Database (HPRD) (Peri et al., 2004), IntAct Molecular Interaction Database (Hermjakob et al., 2004), and Molecular Interactions Database (MINT) (Zanzoni et al., 2002). Meta-databases are usually the integration of primary databases information and some original data such as Agile Protein Interactomes Data Server (Alonso-Lopez et al., 2016) and Interologous Interaction Database (I2D) (Brown and Jurisica, 2005). Prediction databases include many PPIs that are predicted using difference techniques such as Human Protein-Protein Interaction Prediction Database (PIPs) (McDowall et al., 2009) and Known and Predicted Protein–Protein Interactions (STRING) (von Mering et al., 2005).

1.1.1.2 Transcriptional regulatory network

A transcriptional regulatory network (TRN) is a collection of transcription factors (TFs) that interact with their targets in the cell to govern gene expression (Figure 1B).

TFs are proteins that are involved in transcribing DNA to RNA by binding to specific DNA sequences such as enhancers and promoters. The interactions between TFs and their targets can be detected via various technologies including electrophoretic mobility shift assay, DNase footprinting assay, chromatin immunoprecipitation and yeast one-hybrid system (Brenowitz et al., 2001; Ezhkova and Tansey, 2006; Hellman and Fried, 2007; Ouwerkerk and Meijer, 2001). TRNs can also be computationally reconstructed through reverse engineering methods using gene expression profiling data (Marbach et al., 2012) (Figure 2).

TRN serves as a "blueprint" of TF-target interactions, which can be used to generate novel biological hypotheses. An important aspect of this application is that TRNs represent statistically significant predictions of TF-target interactions obtained from high throughput datasets. In this way, TRNs enable us to study regulatory relationship among genes and understand the underlying cellular processes in living cells.

1.1.1.3 Gene functional interaction network

Gene functional interaction network is a conceptual framework that integrates diverse interaction data types including gene co-expression, gene fusion, phylogenetic profiling, co-citation, protein interaction, homologous functional relationship prediction and so on (Figure 1B). Different types of data are usually combined using computational techniques such as naïve Bayes framework. In contrast to physical interaction networks, functional networks calculate the probability how likely two genes functionally interact with each other. Therefore, functional interaction network is considered as a comprehensive combination of physical, genetic and regulatory interactions. So far,

functional interaction networks have been built for a number of species, including yeast, mouse, and human (Guan et al., 2008; Lee et al., 2011; Lee et al., 2004).

1.1.2 Topological features

1.1.2.1 Degree

Node degree, k is the most elemental characteristic of a node, which tells us how many links connecting this node to other nodes. In directed networks, each node has an incoming degree k_{in} and an outgoing degree k_{out} that depend on the directionality of connected edges. Nodes with a large degree are often called hubs. In the protein interaction networks of various organisms, hub nodes tend to be essential regulators or mediators of cellular functions (Barabasi and Oltvai, 2004). Weighted degree is a variety of node degree that accounts for the weight of connected edges. In this way, connected edges with large weight can contribute to the importance of a given node.

1.1.2.2 Centrality

In graph theory, centrality is a measure which indicates the topological importance of nodes in the network. Different types of centrality measures have been developed to evaluate the node importance, including betweenness centrality, closeness centrality, PageRank centrality, etc.

Betweenness centrality quantifies how many times a node acts as a bridge along the shortest path between two other nodes (Barrat et al., 2004). Specifically, it counts the number of shortest paths that go through a given node for all pairs of other nodes. Exclusion of nodes with large betweenness centrality values can potentially disconnect the entire network, so those nodes are usually considered as "bottlenecks" of the network.

Closeness centrality of a node is defined as the sum of distances from the given node to all other nodes (Bavelas, 1950). Thus, the more centrally located a node is in the network the smaller its closeness centrality.

PageRank is a metric developed by Google to rank the importance of websites in their search engine results (Page, 1998). The basic idea of PageRank is to count the number and quality of linked pages to determine a rough estimate of the website importance. The underlying assumption is that more important websites are likely to receive more links from other important websites.

1.1.2.3 Clustering coefficient

In many networks, if node *i* is connected to node *j*, and node *j* is connected to node *k*, then it is likely that *i* also has a direct connection with *k*. This feature can be quantified using the clustering coefficient which is defined as $C_i = \frac{2n_i}{k_i(k_i-1)}$, where n_i is the number of edges connecting the k_i neighbors of node *i* to each other. This feature measures the tendency of a graph to be organized into clusters. Biological networks show a significant higher average clustering coefficient compared to random networks, which indicates their modular nature (Hao et al., 2012). The clustering coefficient has also been successfully used to summarize important features of unweighted, undirected networks across wide range of applications (Costantini and Perugini, 2014; Nacher et al., 2004; Zhu et al., 2007).

1.1.3 Network based algorithms

1.1.3.1 Distance based methods

A common approach to discover novel gene pathways is to test if there is a path connecting putative causal genes to target genes (Figure 3A). Such a shortest path linking

a causal gene and its target is often utilized to describe their causal relationship (Shih and Parthasarathy, 2012; Zhang et al., 2013b). The intermediate nodes on a shortest path are likely members of the affected pathway. If multiple target genes exist, the Steiner tree algorithm can be applied to identify a set of shortest paths (Sadeghi and Frohlich, 2013). The basic idea behind the Steiner tree algorithm is to find a tree connecting all candidate nodes using the minimum number of edges. In a Steiner tree, the size of the entire tree is minimized instead of minimizing every individual path in this tree (Figure 4). The Steiner tree algorithm has been widely used to discover hidden or new members of a pathway and associations between them.

1.1.3.2 Network flow based methods

Unlike identifying single paths connecting causal gene and target genes, information flow-based approaches aim to find the fraction of information flow going through intermediate nodes and edges (Figure 3B). The idea of flow-based methods is to mimic the current flow in an electronic circuit, where each edge has a certain amount of resistance (R. Ahlswede, 2000). Current flow network provides a useful framework equivalent to a random walk, which is also commonly used in modeling information flow. A key advantage of flow-based approach is their ability to integrate additional data, so it can identify information propagation pathways with increased confidence. Information flow based approach has been utilized to uncover protein functions as well as prioritize causal disease genes (Chen et al., 2011; Hamaneh and Yu, 2014; Wu et al., 2015). In contrast to approaches such as shortest path or random walk, flow based methods produce subgraphs that are more complicated structures rather than linear paths

(Pearson, 1905). In this way, flow based approaches are able to consider the complex interactions between identified subset of nodes.

1.2 Heart failure

Heart failure is a complex disease involving multiple genetic and environmental factors. It is still the most devastating cardiovascular disease in terms of morbidity, mortality, quality of life, and health care cost (Guyatt, 1993). Heart failure occurs when the heart cannot pump enough blood to satisfy the body's needs. There are two main types of heart failure: 1) heart failure due to left ventricular dysfunction, 2) heart failure with normal ejection fraction but a contract defect in left ventricular (Mant et al., 2011). The clinical manifestation of heart failure is mainly caused by the primary myocardial disease, common coronary artery disease, hypertension, and inherited cardiomyopathy. Although the etiology highly varies, heart failure is due to a derangement of interplay among the cardiac, renal, and vascular systems (Segovia Cubero et al., 2004). Among clinical causes of heart failure, cardiac hypertrophy and mitochondrial dysfunction are two important factors.

1.2.1 Overview of cardiac hypertrophy

Cardiac hypertrophy is an adaptive response to pressure or volume stress, mutations of sarcomere proteins, loss of contractile mass from prior infarction. Hypertrophic growth accompanies many forms of heart diseases including hypertension, heart failure, and vascular disease (Tardiff, 2006). The mammalian heart is an organ that can grow and change in response to physiological and pathological overload. To accommodate alterations in its workload, the heart undergoes hypertrophic enlargement by increasing the size of individual cardiac myocytes (Waring et al., 2014). During

embryonic development, the heart grows through proliferation and hypertrophy of cardiac myocytes. At around embryonic day 8, heart tube starts to format primitive avascular structure that contains a few layers of cardiac myocytes. To obtain nutrients and oxygen through diffusion, those layers are attached to endocardial surface area. To maximize diffusion, myocardial wall increases thickness by myocyte proliferation. After birth, the heart growth is achieved by hypertrophy of individual cardiac myocytes during postnatal development (Maillet et al., 2013). Furthermore, a significant expansion of the myocardial vessels occurs in heart to satisfy the increased oxygen and metabolic demands (Komuro, 2001).

Cardiac hypertrophy in normal growth or in trained athletes is described as physiological hypertrophy. Such hypertrophy is defined by normal or enhanced contractile function and normal organization of cardiac structure (Krymskii, 1958). However, cardiac hypertrophy in patients with cardiovascular diseases is considered as pathological hypertrophy (Weber and Brilla, 1991). In this type of hypertrophy, contractile dysfunction, interstitial fibrosis, and re-expression of fetal cardiac genes are often observed. It is induced by a number of common disease stimuli such as hypertension and myocardial infarction (Kahan and Bergfeldt, 2005). In response to those stimuli, cardiac hypertrophy is induced via myocytes growth in length and/or width. Therefore, the heart can increase cardiac pump function and decrease ventricular wall tension. As this compensated hypertrophy develops, patients are susceptible to heart failure and arrhythmia even sudden death (Shimizu and Minamino, 2016).

At the molecular level, studies revealed central pathways, effectors and transcriptional regulators that have critical roles in the development of cardiac

hypertrophy. So far, several signaling pathways have been implicated, including mitogen-activated protein kinase (MAPK) pathway, calcineurin–nuclear factor of activated T cells (NFAT) pathway and insulin-like growth factor-I (IGF-I)–phosphatidylinositol 3-kinase (PI3K)– AKT/protein kinase B (PKB)–mammalian target of rapamycin (mTOR) pathway (Hou and Kang, 2012; Hunter and Chien, 1999; Ruwhof and van der Laarse, 2000). Transverse aortic constriction (TAC) in the mouse is a commonly used experimental model for cardiac hypertrophy and heart failure (Kuang et al., 2013; Zhou et al., 2015). TAC first induces compensated cardiac hypertrophy that is related to a temporary enhancement of cardiac contractility. Gradually, the response to the chronic pressure overload results in cardiac dilatation and heart failure. This technique has been extensively used as an efficient way to mimic and study human cardiovascular diseases.

1.2.2 Mitochondrial dysfunction in heart diseases

Mitochondrion is an organelle that is found in all eukaryotic cells, in which the processes of respiration and energy production occur. It is the "powerhouse" of the cell. The electron transport chain in the inner membrane of the mitochondrion is the last step of aerobic respiration that is involved in the synthesis of adenosine triphosphate (ATP). Electron transport chains are also major sites of premature electron leakage to oxygen, forming superoxide and leading to oxidative stress. This chain contains four protein complexes (Complex I, II, III, IV and V), among which Complex I is the first enzyme. Complex I is also referred to as NADH:ubiquinone, catalyzing the transfer of electrons from NADH to coenzyme Q10. It is also a major contributor to cellular production of reactive oxygen species. Complex I contains about 45 different subunits, whose major role is to synthesize ATP. Under normal condition, up to 5% of the mitochondrial oxygen

consumption can be transferred into reactive oxygen species. In model systems, electron transport chain dysfunction is often related to elevated reactive oxygen species generation and cellular damage (Ramamoorthy et al., 2014). Additionally, cell death is often considered as the main cause to the abnormal opening of the mitochondria permeability transition pore in many diseases (Adams and Turnbull, 1996; Ramamoorthy et al., 2014).

Mitochondrial dysfunction has been repeatedly observed in heart failure but its role in the development and progression of heart failure remains elusive (Marin-Garcia and Goldenthal, 2008; Neubauer, 2007; Rosca et al., 2008). Complex I is the primary site for electron transfer from NADH as well as the major source of O_2^-, thus a key player in the regulation of ATP synthesis, oxidative stress and cellular redox balance. Impaired function of Complex I in the mitochondrial respiratory chain is commonly observed in cardiomyopathy and heart failure of a variety of etiologies (Ide et al., 1999; Marin-Garcia et al., 2009; Scheubel et al., 2002). To determine the mechanism linking Complex I dysfunction and heart failure, a recent study generated a mouse model with cardiac-specific impairment of Complex I function by deleting the *Ndufs4* subunit (Ndusf4H-/-) (Karamanlidis et al., 2013). NADH dehydrogenase-ubiquinone-FeS 4 (*Ndufs4*) subunit gene is involved in proper Complex I function such that the loss of *Ndufs4* decrease Complex I activity. Thus, Complex I deficiency can be induced by deleting *Ndufs4*.

1.2.3 Accelerated heart failure

Mitochondria play a central role in the normal functioning of the heart, and in the pathogenesis and development of different types of heart diseases. Mitochondria-triggered cell death is a major cause of cardiac injury and heart failure during cardiac stress. Previous studies in mitochondrial related diseases have revealed that over 50% of

individuals with mutations in genes encoding mitochondrial proteins develop cardiovascular diseases (Greaves and Taylor, 2006; Meyers et al., 2013).

In the preliminary study, we found that the knockout of *Ndufs4* results in a significant loss of Complex I supported respiration in the heart, which is well tolerated with no major change of cardiac function under unstressed conditions. Nevertheless, Complex I deficiency elevates protein acetylation through changing the redox state and renders the heart vulnerable to cell death and heart failure during chronic increased stress. This suggests that malfunction of mitochondrial Complex I predisposes the myocardium to injury via changed protein acetylation and susceptibility to chronic stress. In turn, additional stress accelerates the heart failure. These observations suggest that Complex I deficiency plays a pivotal role in the development of heart failure.

1.3 Development of hematopoietic stem cell development

Hematopoietic stem cells (HSCs) are responsible for the continuous production of all mature blood cells during the entire life span of an individual. They are clinically important cells in transplantation protocols used in therapies for blood-related diseases (Doulatov et al., 2012). Hematopoiesis in the mouse embryo occurs in several tissues including the yolk sac (YS), aorta-gonad-mesonephros (AGM), placenta and liver. HSCs develop from hemogenic endothelial (HE) cells that are generated via a process known as endothelial-to-hematopoietic transition in both AGM and YS (Kaimakis et al., 2013).

1.3.1 Overview of hematopoiesis

Hematopoiesis is the process in which blood cells are formed. Establishment and maintenance of the blood system rely on HSCs that normally reside in the bone marrow and have the ability to give rise to all mature blood cell types. HSCs are self-renewing

cells which are able to produce identical daughter HSCs without differentiation. However, HSCs can also differentiate into one or more types of blood cells by following specific differentiation pathways (Orkin 2000). All blood cells are classified into three lineages which are erythroid cells, lymphocytes and myelocytes. Erythroid cells are the oxygen carrying red blood cells. Both reticulocytes and erythrocytes are red cells that are functional and released into blood. Lymphocytes are the essence of the adaptive immune system. The lymphoid lineage consists of T cells and B cells which are known as white blood cells. Myelocytes include granulocytes, megakaryocytes and macrophages that are generated from common myeloid progenitors participating in innate immunity, adaptive immunity and blood clotting.

The hematopoietic system provides a successful example in applied regenerative medicine. In the past several decades, stem cell transplantation has become a routine therapy for blood diseases. After the eradication of the patient's own hematopoietic system, the transplanted hematopoietic stem cells provide lifelong reconstitution of the blood system of the patient. Despite of the success, the transplantation method still needs to be improved to reduce the engraftment failure and post-transplant infections.

1.3.2 Developmental origins of hematopoietic stem cells

HSC development during embryogenesis is a complex process involving multiple anatomical sites. After HSC precursors have been specified from mesoderm, they develop into functional HSCs and expand into a pool of HSCs in the fetal liver. Blood cell specification occurs at least three separate time points in the mammalian embryo that leads to three waves of hematopoietic cells production. The three waves provide a way by

which the embryo is able to be supplied with rapidly produced hematopoietic cells (Kaimakis et al., 2013).

The first wave of blood generation produces short-lived primitive erythrocytes that involve transmitting oxygen through the rapidly growing conceptus and also primitive macrophages and megakaryocytes. Primitive erythrocytes are generated from aggregates of mesodermal precursors in the yolk sac blood islands. In the mouse embryo, the second wave of hematopoietic cell generation starts at embryonic day (E)8/8.5 overlaying with first wave. Definitive hematopoietic progenitors are generated and hematopoietic cells begin to form clusters in the major blood vessels at E9.5. The third wave of hematopoietic cell specification leads to the generation of adult HSCs. Hematopoietic cell clusters are found on the ventral wall of the dorsal aorta and major arteries of the chick embryo resulting in the proposition that HSCs for the adult hematopoietic system arise from vascular endothelial cells (Boisset and Robin, 2012; Swiers et al., 2010). However, HSCs found in the yolk sac lack the definitive hematopoietic stem cells that do not show long-term hematopoietic reconstitution activity in mouse embryo prior to E11.5 (Chotinantakul and Leeanansaksiri, 2012).

During normal embryonic development, hematopoietic progenitors (HPs) and HSCs arise from a small population of endothelial cells referred as hemogenic endothelium (Bertrand et al., 2010; Boisset et al., 2010; Eilken et al., 2009; Kissa and Herbomel, 2010). Hemogenic endothelial cells, which begin as flat cells in a monolayer interconnected by tight junctions, undergo a transition to form round cells that express hematopoietic markers (CD41, CD45), briefly accumulate in the form of clusters, detach from the endothelial layer, and enter the circulation. Hemogenic endothelium is found in

multiple anatomic sites in the embryo including the yolk sac, the vitelline and umbilical arteries, and the dorsal aorta where it is flanked by the developing urogenital ridges in the AGM region (North et al., 1999). The first hemogenic endothelial cells in the mouse appear in the yolk sac at approximately E8.5, they are most abundant in the major arteries at E9.5, and the majority of them complete their transition to hematopoietic cells between E9.5 and E12.5 (North et al., 1999; Yokomizo et al., 2012). Each endothelial to hematopoietic cell transition takes approximately 5 hours to execute, and all hemogenic endothelial cells undergo the transition into a blood cell during a 3-4 day period (Bertrand et al., 2010; Boisset et al., 2010; Eilken et al., 2009; Kissa and Herbomel, 2010). The formation of hematopoietic progenitors from ES cells, and through direct reprogramming of endothelial cells with Runx1, Gfi1, Fosb, and Spi1 also proceeds through a hemogenic endothelial intermediate, recapitulating the normal developmental process (Eilken et al., 2009; Lancrin et al., 2009; Sandler et al., 2014).

The endothelial to hematopoietic cell transition is a fascinating process to behold. However, almost nothing is known about the transcriptional regulatory network that is responsible for this transition. Hemogenic endothelium from yolk sac and major arteries differ with respect to the types of hematopoietic progenitors they produce. Yolk sac hemogenic endothelium produces primarily committed erythroid/myeloid progenitors, whereas embryonic endothelium produces HSCs (Vo and Daley, 2015).

A broad range of signaling pathways have been identified involving in the hemogenic endothelial cells development. The potent TGF-beta family member bone morphogenic protein 4 is expressed in the sites undergoing hematopoietic clusters formation in both mouse and human (Hirschi, 2012). Hedgehog and Notch family

members also play critical roles in the development of hematopoietic cells derived from endothelium (Hofmann et al., 2009; Pajcini et al., 2011). In addition, the transcription factor Runx1 is also essential for hemogenic endothelium development (Michael J. Chen, 2009). Conditional knockout of Runx1 in endothelial endothelium leads to a failure to proceed through the endothelial-hematopoietic transition. On the other hand, there is no such defect in the hematopoietic compartment with Runx1 was knockout. Runx1 is also known to regulate other transcriptional factors including Scl, Lmo2 and Gata2, which are also important in hemogenic endothelium development (Nottingham et al., 2007).

1.4 Genetic variants associated with diseases

Genetic variation refers to the variation in the DNA sequence within each individual's genome. Human genetic variation is the genetic difference both within and among populations. There are different types of variants located within gene body or non-coding regions in the human genome including small-scale sequence variation and large-scale structural variation. Small-scale sequence variations refer to variations less than 1 kb such as single nucleotide variations, small insertions and deletions. Whereas, large-scale structural variations are usually greater than 1 kb including copy number variations and chromosomal rearrangements (Feuk et al., 2006).

1.4.1 Overview of Genome Wide Association Studies

A key goal of human genetics is to find genetic risk factors for diseases. Genome Wide Association Studies (GWAS) have evolved into a powerful tool for investigating the genetic architecture of human diseases. In GWAS, hundreds of thousands of single-nucleotide polymorphisms (SNPs) are tested for association within hundreds or thousands individuals for a disease or trait. SNPs are single base pair changes in the DNA sequence

that occur with high frequency (>1%) in the human genome. To investigate genetic risk factors, SNPs are usually used as markers for a genomic region, with a large majority of them having a minimal impact on diseases or traits.

The most common experimental design of GWAS is the case-control setup, which compares two large groups of individuals, one healthy control group and one case group affected by a disease. All individuals in each group are genotyped for the set of common known SNPs. The typical number of SNPs is about one million but it may vary between different genotyping technologies. For each SNP, it is determined if the allele frequency is significantly different between the case and the control group. In this type of setup, the basic unit for reporting effect sizes is the odds ratio, which compares the odds of disease for individuals having a specific allele to the odds of disease for individuals who having a different allele. Furthermore, a p-value for the statistical significance of the odds ratio is estimated using a common statistical test, such as the Chi-squared test.

1.4.2 Mapping of causal disease variants

The vast majority (over 90%) of disease-associated variants from GWAS studies have been found to localize outside of known protein-coding sequences, thus impeding the direct interpretation of their functional effects (Maurano et al., 2012) (Figure 5A). These noncoding variants perturb binding sites of transcription factors, local chromatin structure or co-factor recruitment, ultimately resulting in changes of transcriptional output of the nearby gene(s). Among different categories of non-coding DNA sequences, transcriptional enhancers play a primary role in regulating gene expression, with many human diseases resulting from altered enhancer activities (Epstein, 2009; Freedman et al., 2011; Noonan and McCallion, 2010; Visel et al., 2009). A consensus finding is that

sequence variants associated with disease and expression variations are enriched in enhancers.

On the other hand, given the abundance of functional non-coding sequences in the genome, many SNPs, if they contribute to disease, likely act by impacting the noncoding genome and being involved in a regulatory function (Chorley et al., 2008; Freedman et al., 2011; Noonan and McCallion, 2010). Characterization of noncoding variants poses a significant challenge in human genetics. Among the different classes of noncoding regulatory sequences, transcriptional enhancers represent the primary basis for differential gene expression, with many human diseases resulting from altered enhancer action (Epstein, 2009; Freedman et al., 2011; Noonan and McCallion, 2010; Visel et al., 2009). Numerous recent studies (both large- and small-scale) had uncovered tens of thousands of enhancers in a diverse array of human cells and tissues (Levo and Segal, 2014; Verfaillie et al., 2016; White et al., 2013). A consensus finding is that sequence variants associated with diseases and expression variations are enriched in enhancers.

1.4.3 Non-coding genetic variants

1.4.3.1 Identifying gene targets of transcriptional enhancers

Functional non-coding variants can influence cis- or trans-regulatory elements, indicating that potential genome interactions can be detected using related molecular biology methods. Chromosome Conformation Capture (3C) technology is a biochemical strategy to analyze contact frequencies between selected genomic sites in cell populations, enabling the study of chromatin looping and genome architecture in three dimensions (Dekker et al., 2002). Further improvement was made based on the original 3C (one vs one) technology to develop 4C (one vs all), 5C (many vs many) and Hi-C (all vs all). 4C

(circular chromosome conformation capture) can capture all regions that physically contact a site of interest (Simonis et al., 2006), whereas 5C (carbon-copy chromosome conformation capture) can detect interactions between all restriction fragments within a given region that is usually less than one megabase away (Dostie et al., 2006). Furthermore, Hi-C uses high-throughput sequencing to comprehensively map genome-wide chromatin interactions by performing DNA-DNA proximity ligation in intact nuclei. Another technology, named chromatin interaction analysis by paired-end tag sequencing (ChIA-PET), allows the detection of long-range interactions mediated by a protein of interest (Fullwood et al., 2010).

1.4.3.2 Transcriptional influence of regulatory variants

One major role of functional non-coding variants is regulating gene expression. Expression quantitative trait loci (eQTL) mapping approach can link DNA sequence variants to expression level change of one or more genes (Rockman and Kruglyak, 2006). They are identified via studying a population of genetically different individuals. These individuals can be members of an outbred population or can be bred using experimental crosses. To identify variants that affect gene expression, two types of data are generated from each individual (Albert and Kruglyak, 2015). First, each individual needs to be genotyped. Second, the expression of each gene in the genome is measured in each individual using either expression microarrays or RNA sequencing. eQTLs are then identified by comparing the genotypes with expression levels using association or linkage analysis. To test if a given sequence variant has a significant impact on the expression of a given gene, a statistical test is performed. Individuals are grouped according to the allele they carry. If the gene has a significantly higher expression level in one group than

in the other group, we can conclude that the variant affects the expression of this gene. The test is repeated at each DNA variant across the entire genome, leading to a genome-wide scan for eQTLs for this gene (Kendziorski et al., 2006).

eQTL studies can locate SNPs within a given GWAS locus that affects target transcript levels, but there are a few features of eQTL data. For instance, eQLTs are cell-type specific and may only occur in disease relevant cell types (Holloway et al., 2011). Furthermore, the SNP associated with transcript level change may not be the causal one. Thus, the association between SNPs and genes expression is correlative and indirect (Gilad et al., 2008). Finally, eQTL analysis rarely considers the combinatorial effects of multiple SNPs at a given locus or SNPs from multiple loci.

1.4.3.3 Computational methodologies in regulatory variants analysis

Genetic variants within coding gene regions typically influence the host coding gene by producing different gene products. However, for those regulatory elements outside coding regions such as regulatory enhancers, determining their target genes is still challenging. Although "C"-based methods and ChIA-PET are useful for detecting enhancer-promoter interactions, they are still costly and available data sets are very limited to a few cell lines. To address this issue, computational approaches were developed to provide a relatively fast and cost-effective way for identifying putative enhancer-gene interactions in related cell types (Figure 5B). A common computational approach is to find the nearest TSS of an enhancer as its target. To improve this simple strategy, additional constraints are considered including insulator sites, histone modification patterns, and expression levels within a given genomic domain (Corradin et al., 2014; Ernst et al., 2011; Heintzman et al., 2009). Besides, an integrated method for

identifying enhancer targets IM-PET leveraging abundant omics data was developed to further improve the prediction (He et al., 2014).

On the other hand, regulatory genetic variants and their impact on protein functions can be predicted by computational approaches based on quantifying constraint on the affected residue from conserved protein sequences. However, such a method cannot be applied to non-coding variants. Thus, alternative computational methods have been developed to prioritize non-coding variants considering genomic and epigenomic features such as DNA sequence, histone modification binding sites, transcription factor binding sites and etc such as GWAVA, Funseq and Cadd (Khurana et al., 2013; Kircher et al., 2014; Ritchie et al., 2014). However, those approaches hardly help us understand more detailed molecular mechanisms for those diseases of interest. For instance, they do not take advantage of tissue/cell specific data.

1.5 Thesis objective

As described above, network-based approaches have been successfully employed to investigate a number of biological problems. Different types of networks have been constructed using either experimental or computational methods. In addition, many network inference methods have been developed to identify interesting genes and pathways associated with certain biological processes. Nevertheless, significant challenges remain in the field of network biology including understanding network dynamics, constructing networks with limited number of samples and identifying disease risk variants.

In this thesis work, I hypothesize that key transcriptional regulators and causal genetic variants affect gene expression through tissue-specific transcriptional regulatory

networks. I also hypothesize that gene expression change is coordinated at the level of gene modules. Here, gene modules are groups of genes or gene products that are functionally coordinated, physically interacting or co-regulated (Barabasi et al., 2011; Hartwell et al., 1999; Segal et al., 2003; Stuart et al., 2003). To test these hypotheses, working with Dr. Xiaoke Ma, I first developed a computational method called *i*nference of Multiple Differential Modules (*i*MDM) that can be used to extract gene functional modules from multiple input networks (Chapter 2). We demonstrated the utility of our method using the development of heart failure as our model system. Applying *i*MDM algorithm, we discovered condition-specific and shared gene modules across different conditions. We also showed that our method is able to help understand pathway dynamics and uncover causal genes during disease progression. Next, I developed a computational framework for constructing condition-specific TRNs utilizing a limited number of samples. I validated the framework using public gene expression profiling and protein-DNA interaction data. After constructing condition-specific TRN for different HSC developmental stages, I further developed a ranking method to find key transcriptional regulators in the TRN (Chapter 3). With this ranking strategy, I identified key TFs that are likely to play an important role in HSC development. Finally, I worked with Dr. Yasin Uzun to develop a general computational framework called Annotation of Regulatory Variants using Integrated Networks (ARVIN) for identifying causal non-coding variants using disease specific gene regulatory networks (Chapter 4). We characterized the performance of our method using known noncoding mutations in 20 diseases. We also applied our method to uncover novel causal variants associated with

seven autoimmune diseases. The proposed methods and studies significantly enhanced our understanding of healthy and disease development.

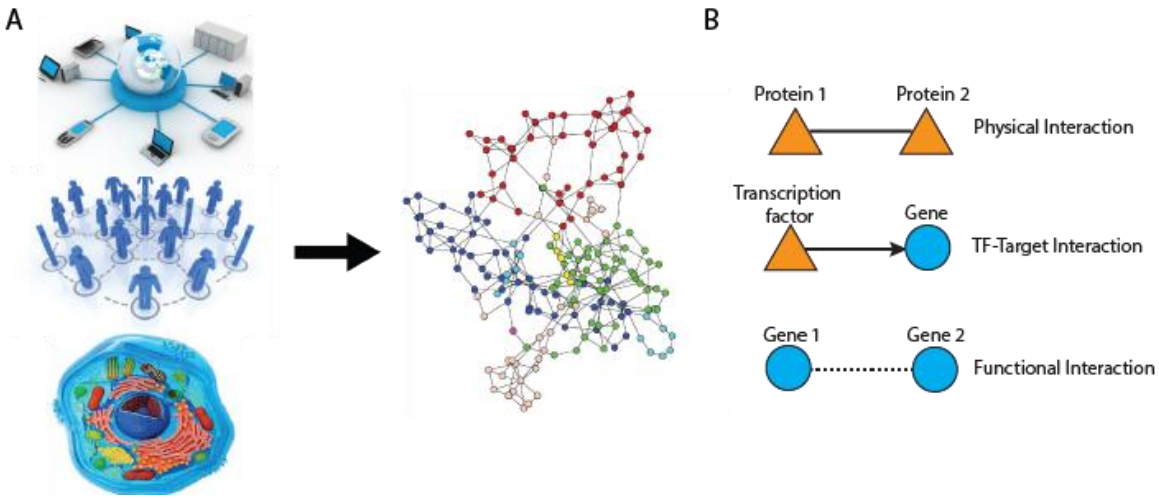

Figure 1. Network modeling.
A) Network modeling **B)** Three types of biological networks: protein-protein interactions; interactions between transcription factor and their targets; and functional interactions between genes. 1) In protein interaction networks, proteins are physically interacted with each other. 2) In transcriptional interaction networks, edges indicate interactions between transcription factors and their targets. 3) In gene functional interaction network, interactions integrate multiple evidences including physical interactions, co-expression, phenotypic/disease, and phylogenetic profiles.

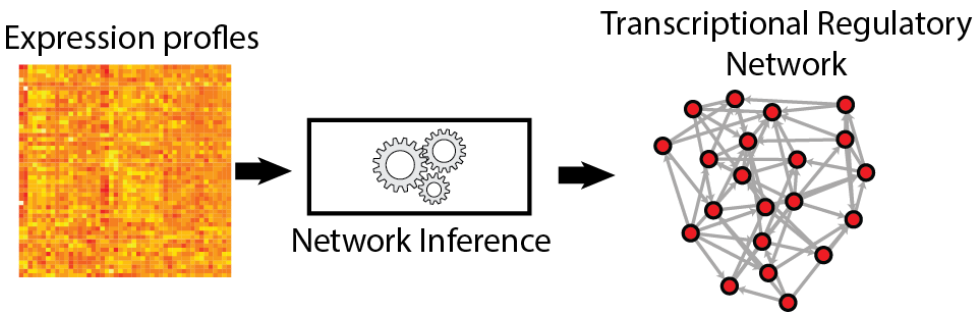

Figure 2. Network construction using computational methods with gene expression profiles.

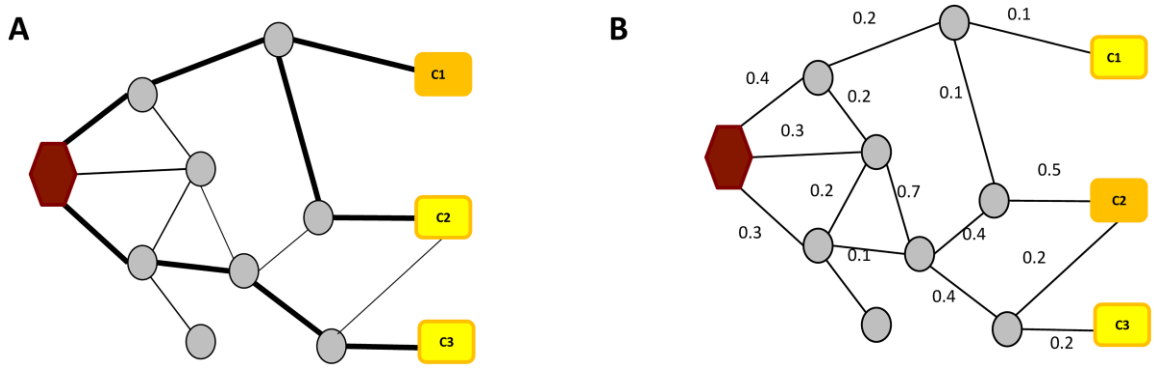

Figure 3. Network-based algorithms.
A) Distance-based method. The shortest paths from a target gene (with hexagon shape) to each of three candidate genes are shown. In this example, c1 has the minimum shortest distance to the root node. **B) Flow-based method**. The gene receiving the most significant amount of flow is identified as the disease gene. The information flow methods can be solved using Kirchhoff's current law. Edge weight is quantified as the amount of information flow going through this edge. In this examples, c2 receives the largest amount of information flow among candidate nodes.

Figure 3 is adapted from "Dong-Yeon Cho et al, (2012) Network biology approach to complex diseases. *PLOS* computational biology."

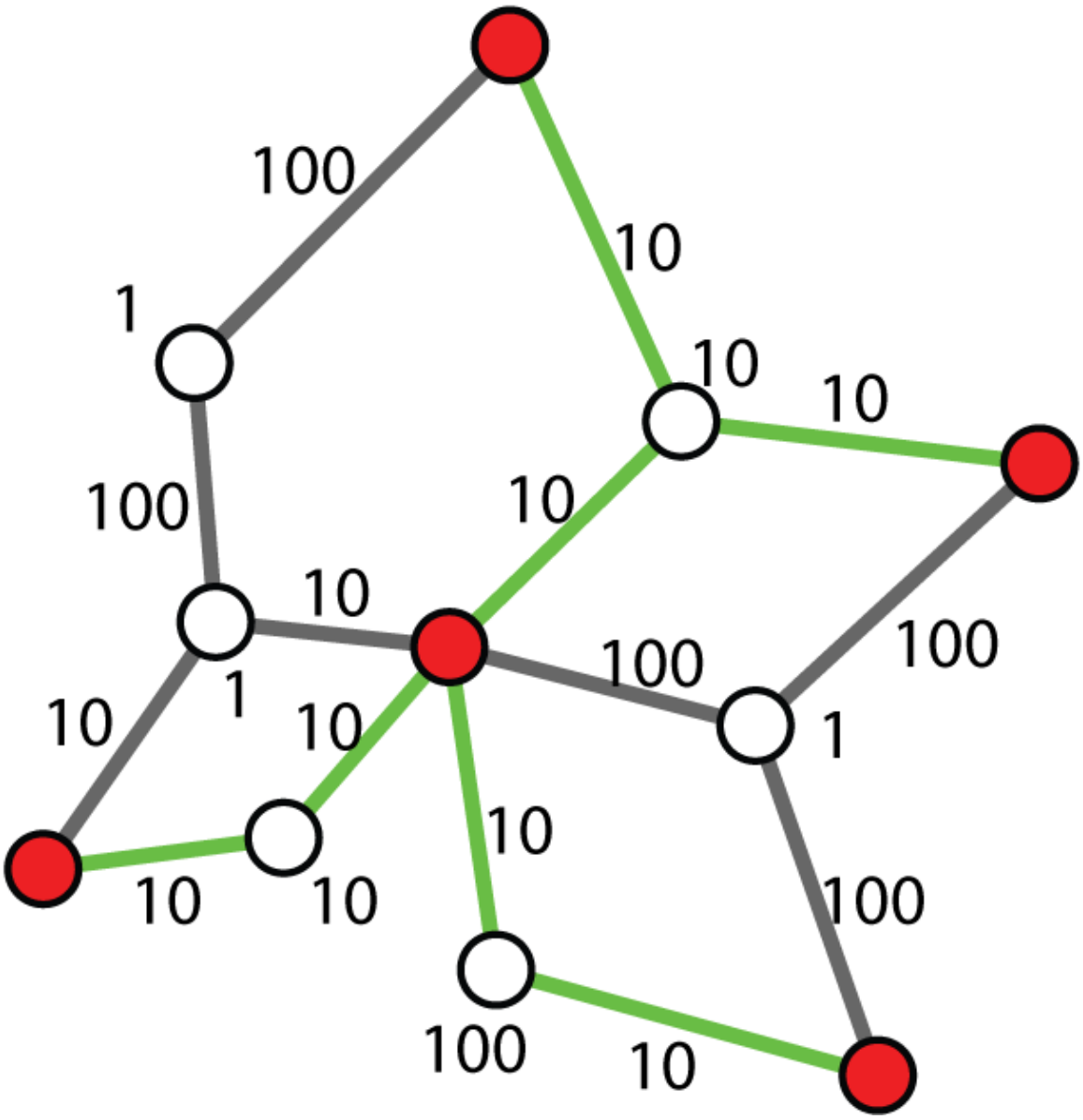

Figure 4. Prize Collecting Steiner Tree algorithm. Given an undirected graph with weighted nodes and edges, this algorithm aims to find a sub-graph (green edges and their connected nodes) with largest node weight minus the edge weight to connect selected terminal nodes (red).

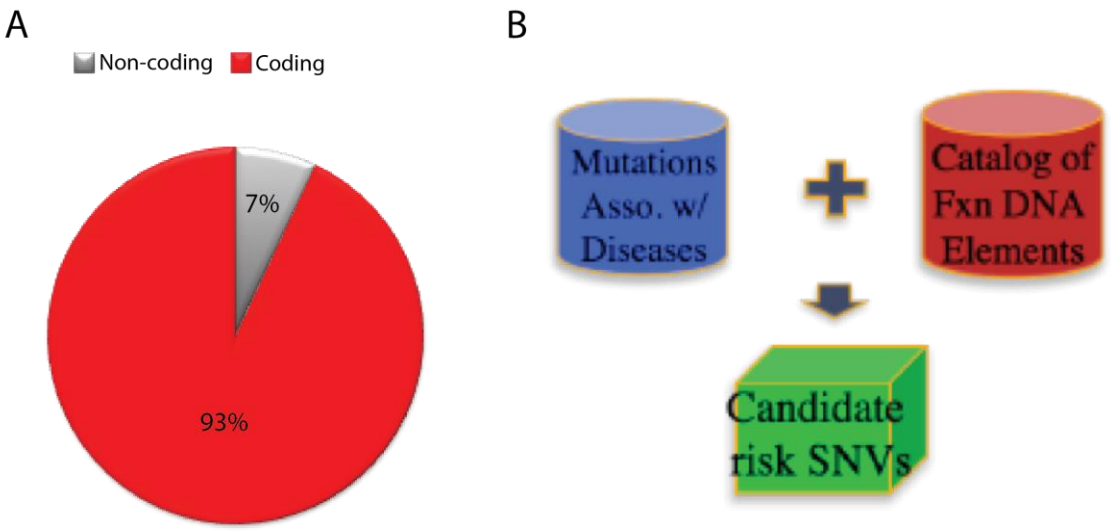

Figure 5. A) Distribution of GWAS SNPs with regard to genomic regions. B) Conventional approach for identifying risk SNPs.

Content of Chapter 2 reprinted from Ma X, Gao L et al., "Revealing pathway dynamics in heart diseases by analyzing multiple different networks." PLOS Computational Biology. 2015. 11:e1004332.

CHAPTER 2: A NETWORK-BASED APPROACH TO INVESTIGATE THE DYNAMICS OF CARDIAC TRANSCRIPTOME DURING ACCELERATED HEART FAILURE USING A MOUSE MODEL

Abstract

Development of heart diseases is driven by dynamic changes in both the activity and connectivity of gene pathways. Understanding these dynamic events is critical for understanding pathogenic mechanisms and development of effective treatment. Currently, there is a lack of computational methods that enable analysis of multiple gene networks, each of which exhibits differential activity compared to the network of the baseline/healthy condition. We describe the *i*MDM algorithm to identify both unique and shared gene modules across multiple differential co-expression networks, termed M-DMs (Multiple Differential Modules). We applied *i*MDM to a time-course RNA-Seq dataset generated using a murine heart failure model generated on two genotypes. We showed that *i*MDM achieves higher accuracy in inferring gene modules compared to using single or multiple co-expression networks. We found that condition-specific M-DMs exhibit differential activities, mediate different biological processes, and are enriched for genes with known cardiovascular phenotypes. By analyzing M-DMs that are present in multiple conditions, we revealed dynamic changes in pathway activity and connectivity across heart failure conditions. We further showed that module dynamics were correlated with the dynamics of disease phenotypes during the development of heart failure. Thus, pathway dynamics is a powerful measure for understanding pathogenesis. *i*MDM provides a principled way to dissect the dynamics of gene pathways and its relationship

to the dynamics of disease phenotype. With the exponential growth of omics data, our method can aid in generating systems-level insights into disease progression.

2.1 Introduction

Many heart diseases are attributable to both genetic and environmental factors (Kathiresan and Srivastava, 2012). These factors can perturb gene transcript levels, protein levels, and metabolite levels, which in turn perturbs the interactions among the molecules. Perturbation of the molecular network ultimately leads to perturbation of the cellular and physiological states, contributing to the diseases. Therefore, understanding molecular networks can lead to important insights into the pathogenic mechanisms of heart diseases.

The concept of network biology has been applied to studies of various cardiovascular diseases, including heart failure (Akavia and Benayahu, 2008; Dewey et al., 2011), atherosclerosis (Gargalovic et al., 2006), coronary heart disease (Huan et al., 2013), and atrial fibrillation (Tan et al., 2013), just to name a few. Because transcriptome data is the most abundant type of omics data, most studies used co-expression networks. In such networks, two genes are connected and assumed to functionally interact if their expression profiles are correlated across multiple conditions. Because genes in the same pathway tend to have correlated expression, analyzing co-expression network is an effective strategy for pathway inference. However, a limitation of previous studies is that networks were constructed using only co-expression information. This practice reduces the statistical power for identifying pathways that are perturbed under diseased conditions. It is more powerful to identify groups of genes that exhibit coherent

differential activities between healthy and diseased conditions. Such gene groups directly capture the perturbed pathways.

Here we described a novel computational framework, *i*nference of Multiple Differential Modules (*i*MDM) that enables simultaneous analysis of multiple differential co-expression networks (DCNs). *i*MDM finds coherently differentially expressed gene modules that are either unique or shared among multiple DCNs. By definition, sets of genes that are differentially expressed under diseased states but do not exhibit correlated expression pattern will not be identified as a module. This is consistent with the notation that the entire pathway is perturbed under disease condition. To capture dynamic changes in gene modules across conditions, we have applied a novel graph-theoretical measure that quantifies changes in both gene activity and gene connectivity.

We demonstrated the utility of our method using the development of heart failure as our model system. Using RNA-Seq, we measured the transcriptome of the heart at four critical stages during the development of heart failure. By applying *i*MDM to multiple differential co-expression networks constructed from our time-course RNA-Seq dataset, we discovered both condition-specific and shared gene modules in gene networks of different heart failure conditions. By quantifying connectivity changes in shared gene modules across different conditions, we showed that gene modules with higher connectivity dynamics have higher correlation with the dynamics of heart failure phenotypic measures, suggesting that studying pathway dynamics using *i*MDM is an effective strategy to uncover causal genes of disease progression. Given the vast amount of transcriptome data, there are ample opportunities to apply our method to better understand the role of network dynamics in the development of heart diseases.

2.2 Results

2.2.1 Profiling of the transcriptome during the development of heart failure using RNA sequencing

We performed a factorial RNA-Seq study to monitor the impact of mitochondrial respiratory complex I deficiency and chronic pressure overload on the heart transcriptome as it progressed from hypertrophy to failure (Figure 6A). Complex I deficiency was triggered by cardiac-specific deletion of Ndufs4 which encodes a structural component of complex I (Karamanlidis et al., 2013). Pressure overload was triggered by transverse aortic constriction (TAC). For each single perturbation or their combinations, we monitored disease progression at four time points, 1, 2, 4 and 8 weeks after the introduction of the perturbation. In total, we profiled the heart transcriptome under 4 major conditions. For the sake of discussion, we termed these conditions wild type sham (WTSH), wild type TAC (WTTAC), knock out sham (KOSH), and knock out TAC (KOTAC). For each time point, four biological replicate RNA-Seq data were generated using 4 hearts.

Hierarchical clustering revealed that the transcriptome profiles of the hearts segregate first by treatment conditions (TAC vs. SH) and then by genotypes (WT vs. KO) (Figure 6B). Further, we found the largest number of differentially expressed genes (DEGs) in the KOTAC vs. WTSH comparison (N = 6521, False Discovery Rate (FDR) < 0.05), followed by the WTTAC vs. WTSH comparison (N = 5238). In contrast, there were only 251 DEGs in the KOSH vs. WTSH comparison. This result suggests that the transcriptome of KOTAC hearts is most perturbed, which is also associated with accelerated heart failure. On the other hand, there is only a very modest perturbation to the transcriptome of KOSH hearts.

2.2.2 Application of iMDM to the heart failure RNA-Seq dataset

Although clustering and differential gene expression analyses can reveal global trend in transcriptome dynamics, such methods cannot reveal individual pathways and their dynamics across conditions, motivating us to develop the inference of multiple differential modules (*i*MDM) algorithm. The major algorithmic steps of iMDM are illustrated in Figure 7. We applied *i*MDM to our heart failure RNA-Seq data and identified M-DMs that occur in single as well as multiple DCNs. Using our RNA-Seq data, we first constructed three Differential Co-expression Networks (DCNs, see Materials and Methods), each of which contains 10929 genes. At a p-value threshold of 0.05, we identified a total of 232 M-DMs, including 109 1-DMs, 107 2-DMs, and 16 3-DMs (Figure 8). A summary of the discovered M-DMs is provided in Table 1. Consistent with the result of our differential expression analysis, the KOTAC DCN yielded the largest number of condition-specific 1-DMs (N = 56) followed by the WTTAC DCN (N = 46). In contrast, much smaller number of modules (N = 7) was identified in the KOSH DCN. *i*MDM also uncovered a large number of 2-DMs in both the KOTAC and the WTTAC DCNs (N = 73).

2.2.3 Performance benchmarking of the iMDM algorithm

We conducted two types of comparisons to demonstrate the advantages of *i*MDM over existing methods. To determine if using differential network can improve performance over using co-expression network alone, we have compared the performance of *i*MDM when fed with these two types of networks separately. We used our RNA-Seq data to construct three DCNs and three co-expression networks for WTTAC, KOTAC, and KOSH condition, respectively. The two algorithms were fed with

appropriate sets of input networks (i.e. DCNs for the *i*MDM, co-expression networks for the other algorithm). The outputs of the two algorithms were seven sets of modules that were discovered from seven sets of networks (three single networks, three sets of two networks, and one set of three networks). To determine if using multiple networks can improve performance over using a single co-expression network, we have compared *i*MDM to the popular WGCNA algorithm (Langfelder and Horvath, 2008), which is primarily designed for the analysis of a single co-expression network and has been used in several studies of heart diseases (Dewey et al., 2011; Gargalovic et al., 2006; Huan et al., 2013; Tan et al., 2013). We generated seven single co-expression networks using one, two, and three experimental conditions at a time, respectively. Each single co-expression network was fed to WGCNA to return a set of modules. Like the other two algorithms, seven sets of modules were computed by WGCNA. We evaluated the resulting seven sets of modules discovered by the different algorithms using multiple reference pathway annotations, including Gene Ontology (Ashburner et al., 2000), KEGG (Kanehisa et al., 2012), MGI pathways (Blake et al., 2014), Canonical pathways (Subramanian et al., 2005), Biocarta (Nishimura, 2001), and Reactome (Croft et al., 2011). *i*MDM achieved significantly higher specificity and sensitivity when evaluated using all except one reference sets (*p-value* < 0.05, one-sided Fisher's exact test, Figure 9A and 9B). Besides gold-standard pathway annotations, a higher percentage of gene modules identified by *i*MDM was enriched for genes whose deletions lead to cardiovascular phenotypes documented in the Mouse Phenome Database (Grubb et al., 2014) (Figure 9C). We concluded that compared to using co-expression networks,

simultaneous analysis of multiple differential co-expression networks improves the inference accuracy of gene pathways.

2.2.4 Condition-specific 1-DMs reveal unique pathways associated with different heart failure conditions

We found that the three sets of 1-DMs were enriched for different Gene Ontology (GO) annotations (Figure 10A). For instance, KOSH 1-DMs were enriched for nucleotide catabolism and localization of cell. WTTAC 1-DMs were enriched for terms such as tricarboxylic acid cycle, phospholipid metabolism, and enzyme linked receptor protein signaling. KOTAC 1-DMs were enriched for terms such as extracellular structure organization, fatty acid metabolism, hemostasis, and negative regulation of response to stimulus. In general, the different enriched terms were consistent with their specific phenotypes. Several of the rate-limiting steps of nucleotide metabolism take place in the mitochondria and can be affected by the fitness of the organelle (Desler et al., 2010). Nucleotide metabolism is generally regarded as a house-keeping process. This is likely the reason why genes involved in nucleotide metabolism were enriched in modules identified under KOSH condition. For the more severe phenotypes of WTTAC and KOTAC, other processes directly related to heart failure were enriched other than this house-keeping process. The other term unique to KOSH 1-DMs was "localization of cell". Genes annotated with this term are involved in communication with the extracellular matrix. Changes in extracellular matrix are linked to myocardial fibrosis and inflammation, which are earlier events of, hear failure development (Diez J, 2005). For the terms specifically enriched among WTTAC 1-DMs, both TCA cycle and phospholipid metabolism contribute to the general energy metabolism deficiency in failing heart, which has been observed before using the TAC model of heart failure [18].

For KOTAC condition, loss of *Ndufs4* leads to significant lower NAD+/NADH ratio in KOTAC hearts compared to WTTAC hearts (Karamanlidis et al., 2013). The low NAD+/NADH ratio inhibits fatty acid beta-oxidation (Ussher et al., 2012). This coordinated down-regulation of the fatty acid module likely contributes to the more severe deregulation of energy metabolism in KOTAC hearts. Hemostasis has been reported to be associated with more severe form of heart failure such as KOTAC (Davis et al., 2000).

Besides GO annotation, we found that a higher fraction of KOTAC 1-DMs (versus WTTAC) was enriched for genes whose disruption leads to cardiovascular phenotypes documented in the Mouse Phenome Database (Grubb et al., 2014) (17.9% vs. 4.3%, p-value = 0.03, one-sided Fisher's exact test, Figure 10B).

Because the edge weight in DCNs is a measure of differential gene expression between the disease and baseline conditions, a larger average edge weight of a 1-DM means a bigger difference in the expression of module genes. In other words, the average edge weight serves as a measure of differential activity of the module. We next compared the distributions of average edge weight in the 1-DMs for WTTAC and KOTAC. Our result shows that 1-DMs in the KOTAC network had a greater difference in the expression level than those in the WTTAC network (0.32 vs. 0.19, p-value = 2.4E-16, one-sided t-test, Figure 10C). We also compared the percentages of up- or down-regulated 1-DMs in the WTTAC and KOTAC networks. At a p-value cutoff of 0.01, we found that the percentage of differentially expressed (up- and down-regulated) KOTAC 1-DMs was significantly higher than that of WTTAC 1-DMs at all time points. For

example, the percentages at week 1 are 54.3% and 85.7% for WTTAC and KOTAC, respectively (p-value = 4.9E-4, one-sided Fisher's exact test, Figure 10D).

Figure 11 shows two example 1-DMs, one unique to the WTTAC network and one unique to the KOTAC network. The top panels of the figure show the visualization of the 1-DMs. The middle panels show the expression profiles of the module genes under four perturbation conditions over time. The bottom panels show the mean edge weights of the 1-DMs in the three non-baseline conditions. Together the middle and right panels explain why a 1-DM is uniquely observed in one condition. Taking the WTTAC 1-DM for example, although many genes of the module were differentially expressed in both WTTAC and KOTAC conditions, their expression correlation was much lower in the KOTAC condition. Notice the tighter correlation of expression profiles among WTTAC module genes compared to that of KOTAC module genes (Figure 11A middle panel). As a result, the edge weights among the module genes in the KOTAC DCN were significantly smaller than those in the WTTAC DCN (Figure 11B right panel). Thus, this module was only identified by *i*MDM in the WTTAC DCN.

The example WTTAC 1-DM is enriched for genes involved in the regulation of cell adhesion (Figure 11A, p-value = 1.1E-4). The example KOTAC 1-DM is enriched for genes involved in fatty acid metabolism (Fig 6B, p-value = 3.2E-8). The expression of this module is significantly lower in KOTAC compared to WTTAC at weeks 2, 4, and 8. A number of module genes encode enzymes for fatty acid metabolism and have significantly reduced expression, including *Acot1*, *Acot2*, *Acsl1*, *Cpt1b*, *Cpt2*, *Crat*, and *Decr1* (Figure 14). These observations are consistent with our previous finding that loss of *Ndufs4* leads to significant lower NAD+/NADH ratio in KOTAC hearts compared

to WTTAC hearts (Karamanlidis et al., 2013). The low NAD+/NADH ratio inhibits fatty acid beta-oxidation (Ussher et al., 2012). This coordinated down-regulation of the fatty acid module likely contributes to the more severe deregulation of energy metabolism in KOTAC hearts.

In summary, the above analyses demonstrate the power of simultaneous analysis of multiple DCNs for uncovering condition-specific pathways involved in heart failure. We found that 1-DMs in the KOTAC condition exhibited higher differential activities during heart failure progression and were enriched for higher fraction of genes with known cardiovascular phenotypes when disrupted. These KOTAC-specific 1-DMs provide new insights into the mechanisms for the accelerated heart failure in KOTAC mice.

2.2.5 M-DMs shared among multiple networks can be used to reveal pathway dynamics during the progression of heart failure

Pathway dynamics can be attributed to changes in both gene expression and connectivity among genes (i.e. pathway rewiring). Although less studied, the latter type of dynamics has recently been shown to play a critical role in disease progression and treatment response, such as the role of hub genes and rewiring of signaling pathways during cancer treatment and cardiac hypertrophy (Drozdov et al., 2013; Lee et al., 2012; Taylor et al., 2009). Here, we demonstrate that *i*MDM enables systematic analysis of pathway dynamics by considering both activity and connectivity changes among shared 2/3-DMs across networks. We further show pathway dynamics is correlated with the dynamic changes in disease phenotypes, which can provide better insights into molecular mechanisms of disease progression.

Because component modules of a 2/3-DM share the same set of genes in multiple DCNs but can differ in their connectivity, 2/3-DM provides a natural way to capture dynamic changes in pathway connectivity. We thus devised the Module Connectivity Dynamic Score (MCDS) to quantify the dynamics of M-DMs (see Materials and Methods for details). Since the DCNs are weighted based on the degree of correlated differential expression, MCDS quantifies not only the presence and absence of edges but also changes in edge weights that can be viewed as the interaction strength among genes.

To identify M-DMs that exhibit significant dynamics than expected by chance, we compared the MCDS values of real 2/3-DMs to a null distribution of MCDS values of random 2/3-DMs. At a p-value cutoff of 0.01, we found 102 dynamic 2/3-DMs. A list of the dynamic 2/3-DMs is provided in Table 1.

Figure 12A shows an example dynamic 2-DM, observed in both KOTAC and WTTAC DCNs. For clarity, only edges with significant weight changes ($p < 0.05$) are shown. For this module, the majority of the changed edges had increased weight in the KOTAC condition compared to the WTTAC condition (in red), due to more significant changes in the expression of the two genes under the KOTAC condition compared to the baseline. There were also a few edges (in green) with decreased weight in the KOTAC condition. These connectivity changes suggest that the pathway was rewired between different heart failure conditions. The degree of rewiring can be quantified by our MCDS metric. Additional example dynamic 2-DMs and 3-DMs are shown in Figure 15 and 16.

Previous studies have shown that certain pathways are more dynamic than others during disease progression or stress response (Bandyopadhyay et al., 2010; Bisson N, 2011; Taylor et al., 2009). To examine this issue in the context of heart failure, we

performed GO term enrichment analysis of the 2/3-DMs. Although certain GO terms were enriched among both dynamic and static 2/3-DMs, each type of M-DMs was also enriched for a unique set of GO terms. For instance, unique functions of the dynamic M-DMs included cell proliferation, trans-membrane transport, ion homeostais, and cell morphogenesis whereas those of static 2/3-DMs include regulation of transcription, chromosome organization and response to organic nitrogen (Figure 17).

The enrichment of unique functional annotations among dynamic modules suggests that dynamic modules may be effective markers for disease progression. We therefore asked how the observed dynamics of 2/3-DMs correlate with the change in cardiac function. We used the following three measures to monitor the function of the heart as it progressed to failure (Figure 18): heart weight normalized by tibial length (HW/TL), left ventricular internal dimension in diastole (LVID(d)) and LV fractional shortening (FS%). For each 2/3-DM, only using conditions from which the M-DM is derived, we computed the correlation between its average normalized gene expression level and each of the three cardiac function measures (Figure 12B, Table 1). Strikingly, we found that dynamic 2/3-DMs had significantly higher correlation with measures of cardiac function than static 2/3-DMs (Figure 12C). For example, the correlations with fractional shortening were 0.60 and 0.31 for dynamic and static modules, respectively (p-value = 5.7E-6, one-sided t-test). This result suggests that dynamic 2/3-DMs are better markers for disease progression.

2.3 Discussion

From a systems biology point of view, diseases are caused by perturbations to the gene network. Such perturbations change dynamically as the disease progresses. We developed a mathematical model to represent perturbed gene networks and a robust search algorithm to identify regions of the perturbed networks with differential activities and connectivity. Differential network analysis has been applied to protein-DNA interaction networks (Harbison et al., 2004; Luscombe et al., 2004), protein-protein interaction networks (Ellis et al., 2012; Workman et al., 2006), genetic interaction networks (Bandyopadhyay et al., 2010; Guenole A, 2013), and functional gene interaction networks (Gill et al., 2010; Zhang et al., 2011). However, in all previous work, only two conditions were considered (i.e. only one resulting differential network) in the computational methods. A key innovation in our method is the ability to identify unique and shared modules from multiple differential gene networks, each of which representing a different perturbation condition. By definition, *i*MDM finds coherently differentially expressed gene modules. Sets of genes that are differentially expressed under diseased states but do not exhibit correlated expression pattern will not be identified as a module. This is consistent with the notation that the entire pathway is perturbed under disease condition. From a computational point of view, it increases the specificity of the inference as we demonstrated in the benchmarking experiment (Fig 4A).

Another challenge in studying network dynamics is how to quantify the rewiring of the pathways. Previous studies only focused on highly connected genes in a pathway, the so-called hub genes, instead of the entire pathway (Pujana et al., 2007; Romanoski et

al., 2011; Taylor et al., 2009). Here, we have used the MCDS metric to quantify the dynamics of an entire pathway. MCDS examines all edges in a module. More importantly, it quantifies not only the presence and absence of edges but also changes in edge weights that can be viewed as interaction strength among genes.

By applying the *i*MDM algorithm to our heart failure RNA-Seq data, we found that condition-specific 1-DMs exhibit differential activities, mediate different biological processes, and are enriched for genes with known cardiovascular phenotypes. Unlike 1-DMs, 2/3-DMs are not condition-specific. A previous study has suggested that there were major differences in topological and biological properties among gene pairs that have global vs. conditional co-expression (Das et al., 2012). We thus compared 1-DMs to 2/3-DMs in terms of their topological features and their activity correlation with disease phenotypes. We found genes in 1-DMs had more connections and located in more central positions in the networks. Activities of 1-DMs also had higher correlation with the disease phenotype measures. This is consistent with the previous observation that conditional interactions are enriched for genes that are key to maintaining network integrity.

In contrast to 1-DMs, M-DMs identified in 2 or more conditions enabled us to study the dynamics of gene modules. By applying the MCDS metric, we were able to distinguish dynamic and static 2/3-DMs. We demonstrated that these two types of modules differ in multiple aspects, including their functional annotations. In particular, we have found that activities of dynamic 2/3-DMs have higher correlation with changes in cardiac disease phenotype, suggesting dynamic modules may play a more important

role during disease progression. Thus, studying pathway dynamics can lead to novel insights into disease pathogenesis.

*i*MDM only needs transcriptome profiling data as the input and both microarray and RNA-Seq data are applicable. Given the increasing amount of transcriptome data on various cardiovascular diseases, we envision that *i*MDM can be applied in several ways to reveal network dynamics under different conditions, including temporal dynamics during disease progression and dynamics between disease subtypes.

Besides comparing disease subtypes as what was done here, another interesting analysis is between-disease comparison, such as heart failure versus arrhythmia. The pioneering work on human disease network by Goh et al. (Goh et al., 2007) has revealed that genes associated with similar disorders show both higher likelihood of physical interactions between their products and higher expression profiling similarity for their transcripts, supporting the existence of distinct disease-specific functional modules. We envision that a pan-heart analysis using *i*MDM can lead to similar insights, in particular pathway signature and disease-specific pathways.

There are a couple of directions that the basic concept of *i*MDM can be extended in future work. First, integrating multiple types of molecular analytes beyond gene expression might further expand our ability to identify dynamic molecular events that are associated with phenotypic dynamics. Genetic mutation data such as those from exome and whole-genome sequencing can be used as prior information to guide module search under the assumption that mutated sequences are likely to be involved in the diseases. Epigenomic data can be integrated with transcriptome data to understand how environmental factors perturb gene networks. Second, comparing and contrasting

dynamic events involving different molecular types may yield new mechanistic insights into their interactions in the context of disease progression.

2.4 Materials and methods

2.4.1 Overview of the iMDM (inference of Multiple Differential Modules) method

Figure 7 provides a schematic of the *i*MDM algorithm. The algorithm takes as the input transcriptome profiles gathered under both healthy/baseline and disease conditions. Using the transcriptome profiles, *i*MDM first constructs multiple differential co-expression networks (DCNs), one for each condition. Two genes are connected in a DCN if they exhibit correlated expression profiles across conditions *and* their expression levels are significantly different between the disease and the baseline conditions. Next, we adapted the *M-module* algorithm (Ma X, 2014) to identify statistically significant multiple differential modules (M-DMs) present in multiple DCNs. *i*MDM is implemented in the R statistical programming language. The software is freely available upon request. In the following sections, we describe details of each algorithmic steps of the method.

2.4.2 Construction of differential co-expression networks (DCNs)

For each disease condition, construction of the DCN consists of two steps: 1) construction of a binary co-expression network; and 2) edge weight assignment based on differential gene expression between the disease and baseline conditions. To construct the binary gene co-expression network, edges are chosen based on the absolute value of the Pearson correlation of the expression profiles of two genes. To remove indirect correlation due to a third gene, we used the 1st order partial Pearson correlation coefficient (Watson-Haigh et al., 2010). Only edges whose correlations are equal or

greater than the pre-defined threshold δ are chosen. In this study, the value of δ was set at 0.8 such that maximal number of genes was connected in all DCNs to be constructed. In step 2, weights are assigned to edges in the binary co-expression network based on the p-value of differential gene expression between the disease and baseline conditions. Various methods can be used to detect differential gene expression for microarray or RNA-Seq data. Here, we used edgeR (Robinson et al., 2010). The weight $w_{i,j}$ on edge (i,j) in the differential network is defined as following: $w_{ij} =$

$$\begin{cases} \frac{(-logp_i - logp_j)^{\frac{1}{2}}}{(2 * max_{t \in V} |logp_l|)^{\frac{1}{2}}}, if cor(i,j) \geq \delta \\ 0, \ if \ cor(i,j) < \delta \end{cases}$$

where p_i and p_j are p-values of differential expression for genes i and j, respectively. V is the node set of the co-expression network, and $cor(i,j)$ is the absolute value of Pearson correlation between genes i,j based on their expression profiles. Under this weighting scheme, genes that are co-expressed and significantly differentially expressed are assigned higher weights, which satisfies our assumption that those genes likely participate in a pathway that exhibit differential activities between the two conditions being compared.

Mathematically, given M DCNs with the same node set but different edge sets, $G_k = (V,E_k)(1 \leq k \leq M)$, they can be represented by a 3-dimensional matrix $A = (a_{ijk})nxnxM$ where a_{ijk} denotes the weight on the edge $e(i,j)$, $w_{i,j}$, in network Gk. An M-DM, C, is defined as a set of genes whose connectivity within them is stronger than random expectation across all M DCNs under consideration.

2.4.3 Identification of multiple differential modules in multiple DCNs

We adapted our recently developed *M-module* algorithm to identify M-DMs. *M-module* is designed for identifying gene modules with common members but varied connectivity across multiple molecular interaction networks (Ma X, 2014).

M-DM search consists of three steps: seed prioritization, module search by seed expansion, and refinement of candidate modules. The seed prioritization step ranks genes in multiple networks by using the topological feature of the gene in the network. Briefly, for each network G_k = **(**V**,**E_k**)(**$1 \le k \le M$**)** with an adjacency matrix A_k = **(**a_{ijk}**)**nxn, we construct a function g**:**$V \rightarrow R$ such that g**(**i) denotes the importance of vertex i in the corresponding network. The function is defined as $g(i) = \sum_{j \in N_k(i)} A_{ijk} g(j)$ where N_k**(**i**)** denotes the set of neighbors of i in G_k; A'_kdenotes the degree normalized weighted adjacency matrix which is computed as $A'_k = D^{-1/2} A_k D^{1/2}$ where D is diagonal matrix with element $D_{ii} = \Sigma_j A_{ijk}$. The product, $A'g$, denotes the information propagation on network via the edges of networks, which means the importance of a node depends on the number of its neighbors, strength of connection and importance of its neighbors. The exact solution to the equation above is $(1 - A'_k)^{-1}$.

For each gene, after obtaining its ranks in all individual networks, denoted as g = **[**g**(1),…,**g**(M)]**, we calculate a z-score for each rank g**(l)**. Then we obtain the rank for that gene across all networks by averaging the z-scores across all networks. The top 10% genes were selected as the seeds although the search result is not sensitive to the fraction of seeds used (Ma X, 2014). Starting with each seed, the module search step iteratively includes genes whose addition leads to the maximum decrease in the graph entropy-based objective function until there is no decrease in the objective function. For a given

vertex $i \in C$, let $L_k(i)$ denotes the total weight between vertex i and other vertices of the M-DM C in the network G_k, i.e., $L_k(i) = \sum_{j \neq i, j \in C} a_{ijk}$. Similarly, let $-L_k(i) = \sum_{j \neq i, j \in V} C^{a_{ijk}}$ denotes the weight between i and vertices outside of C. We defined the entropy for the connectivity of vertex i to C as $H_k(C_i) = -p_i^{[k]} log p_i^{[k]} - \left(1 - p_i^{[k]}\right) \log\left(1 - p_i^{[k]}\right)$ where $p_i^{[k]} = {L_k(i)}/{(\bar{L}_k(i) + L_k(i))}$. The motivation for using graph entropy is that it quantifies the skewness of in-module connectivity versus out-module connectivity. Summing over all vertices in C and network k, we have $Hk(C) = \sum i \in C H(Ci)$. The graph entropy for C across all networks and normalized for the size of C is $H(C) = {(\sum_{k-1}^{M} H_k(C))}/{|C|}$

The objective function of the algorithm is defined as: $\sum_{i=1}^{\tau} minH(C_i)\ s.t. \begin{cases} x_{ij} \in \{0,1\} \\ \sum_{j=1}^{\tau} x_{ij} \geq 1 \\ \sum_{i=1}^{n} x_{ij} > 0 \end{cases}$

where $C_i (1 \leq i \leq \tau)$ is a candidate M-DM. $X = [x_1, \ldots, x_\tau]$ is an index matrix in which each column corresponds to an M-DM and each row corresponds to a gene. The constraints mean that each gene can belong to one or more modules and each module has to contain at least one gene.

During the refinement step, M-DMs whose sizes are smaller than five are removed. To merge overlapping M-DMs, we used Jaccard index which is the ratio of intersection over union for two sets. A Jaccard index of 0.5 was used in this study.

2.4.4 Calculation of the statistical significance of candidate M-DMs

The statistical significance of M-DMs is computed based on the null score distribution of M-DMs generated using randomized networks. Each network is completely randomized 100 times by degree-preserved edge shuffling. To obtain module

scores for the null distribution, we performed module search on the randomized networks. Using the null distribution, the empirical p-value of an M-DM is calculated as the probability of the module having the observed score or smaller by chance. P-values are corrected for multiple testing using the method of Benjamini-Hochberg (Benjamini Y, 1995). An adjusted p-value of 0.05 was used as the significance threshold.

2.4.5 Quantification of connectivity dynamics of shared M-DMs

By definition, each M-DM with M≥2 has multiple component modules from different DCNs. To quantify the change in the connectivity of component modules, we used a graph-theoretical measure, the Module Connectivity Dynamic Score (MCDS). Specifically, given an M-DM *C* whose weighted adjacent matrices of the corresponding induced subgraphs are denoted by$A_i^C (1 \le i \le M)$, the MCDS between two adjacent component modules is defined as the $L\mathbf{2}$norm of the matrix subtraction normalized by the number of genes in the M-DM, *i.e.*, $\Delta A_{i,i+1}^C = \left\| A_i^C - A_{i+1}^C \right\|_2 / |C|$where $\|\cdot\|_2$ is the matrix $L\mathbf{2}$ norm. The overall MCDS of an M-DM is defined as the average MCDS of all pairwise comparisons: $\tau(A^C) = \sum_{i=1}^{M-1} \Delta A_{i,i+1}^C / (M-1)$

The statistical significance of MCDS for an M-DM is computed in a similar way as that for M-DMs. Briefly, we first calculate the null distribution for MCDS scores based on random M-DMs. The empirical p-value of an MCDS is calculated using the null distribution. The method of Benjamini-Hochberg is used for multiple testing correction. An adjusted p-value of 0.05 was used as the significance threshold.

2.4.6 Transgenic mice, transverse aortic constriction surgery and echocardiography

Generation of transgenic mice with cardiac restricted *Ndufs4* deletion was described in our recent publication (Karamanlidis et al., 2013). Mice were fed on rodent

diet and water available *ad libitum* with a 12-hour light/dark cycle in a vivarium. Adult male mice (3–4 months old) received transverse aortic constriction (TAC) to induce chronic pressure overload or sham surgeries as previously described (Tarnavski O, 2004). Cardiac geometry and function (left ventricular internal dimension in diastole (LVID(d)) and LV fractional shortening (FS%)) were recorded at 1, 2, and 4 weeks using echocardiography with the VEVO 770 system equipped with a 707B scan head. All measurements were averaged from six cardiac cycles.

2.4.7 RNA sequencing and data processing

Total RNA was isolated from frozen cardiac tissue using the RNeasy Kit for fibrotic tissues (Qiagen) and treated with DNase to remove genomic DNA contamination. Quality and integrity of RNA were checked using Agilent Bioanalyzer 2100. All samples used for RNA sequencing had a Bioanalyzer RIN number of at least 8. Illumina TruSeq RNA sample preparation kit was used to generate multiplexed sequencing libraries. Libraries were loaded into a flowcell at a concentration of 5 pM and clustered on an Illumina cBot. Sequencing was done on a HiSeq2000 that generated paired-end reads of length 50 bp.

We mapped paired-end reads to the mouse genome (mm9) using Tophat (Trapnell et al., 2009). Only uniquely mapped reads with fewer than 2 mismatches were used for downstream analyses. Transcripts were assembled using Cufflinks (Trapnell et al., 2010) and Ensemble (release 66) as the source of annotated transcripts. Normalized transcript abundance was computed using Cufflinks and expressed as FPKM (Fragments Per Kilobase of transcripts per Million mapped reads). FPKM values were developed for paired-end RNA-seq data, whereas RPKM (Reads Per Kilobase of transcripts per Million

mapped reads) is used for single-end reads. Furthermore, FPKM already considers transcript length and library size (Garber et al., 2011). Thus, genes are comparable within or between samples using FPKM values. Gene-level FPKM values were computed by summing up FPKM values of their corresponding transcripts (Trapnell et al., 2010). FPKM values were used to compute gene co-expression networks.

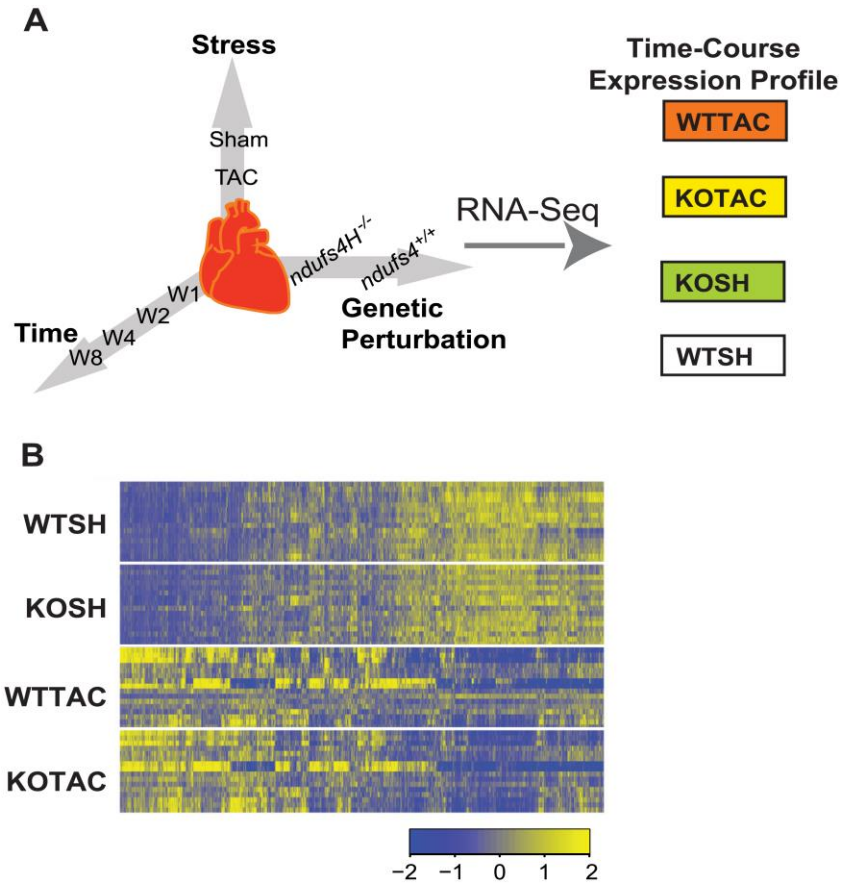

Figure 6. RNA-Seq experiment using a mouse heart failure model generated on two genotypes.
A, Time-course RNA-Seq data were generated using mouse hearts perturbed by stress (TAC) or genetic perturbation (Ndufs4 deletion) or both, resulting four conditions. WTSH, sham treatment of wild type mice (baseline); KOSH, sham treatment of KO mice; WTTAC, TAC treatment of wild type mice; KOTAC, TAC treatment of KO mice. B, Hierarchical clustering of top 5000 genes with highest variance of expression levels across conditions.

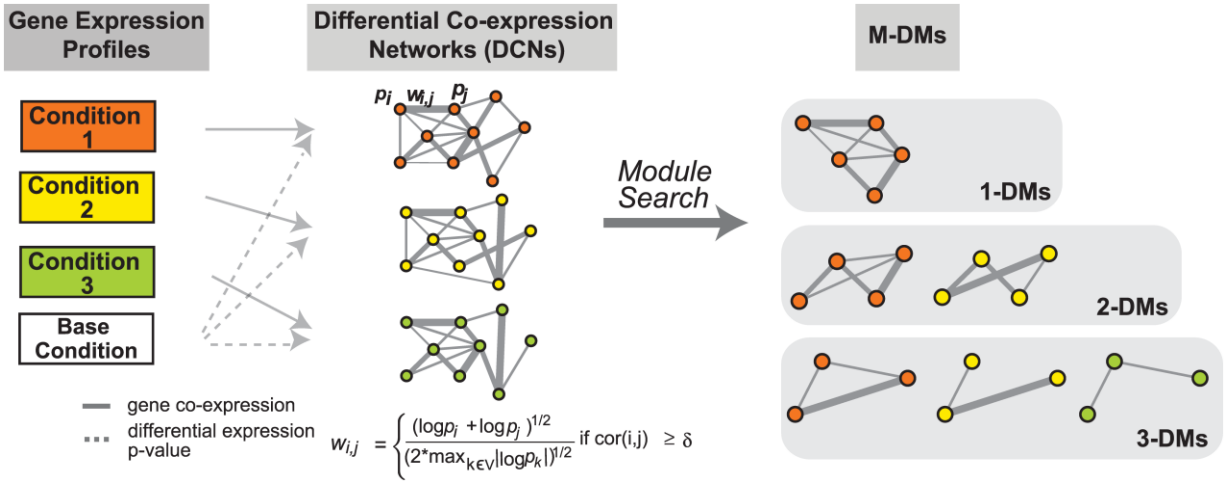

Figure 7. Overview of the *i*MDM algorithm.
The algorithm has two major steps. First, gene expression profiles across multiple conditions are used to build differential gene co-expression networks (DCNs). To build DCNs, a binary co-expression network is constructed first in which edges are chosen based on the absolute value of Pearson correlation of the expression profiles of two genes. Only edges whose correlation exceeds a pre-defined threshold $\boldsymbol{\delta}$ are included in the binary network. Edges in the binary network are then weighted ($\boldsymbol{w_{i,j}}$) based on the p-values ($\boldsymbol{p_i}$ and $\boldsymbol{p_j}$) of differential gene expression between the baseline and disease conditions. Second, multiple differential co-expression networks are analyzed to identify shared and unique multiple differential modules (M-DMs) under different conditions. 1-DM are modules that are only found in one condition whereas M-DMs with $M \geq 2$ are modules that are found in multiple conditions.

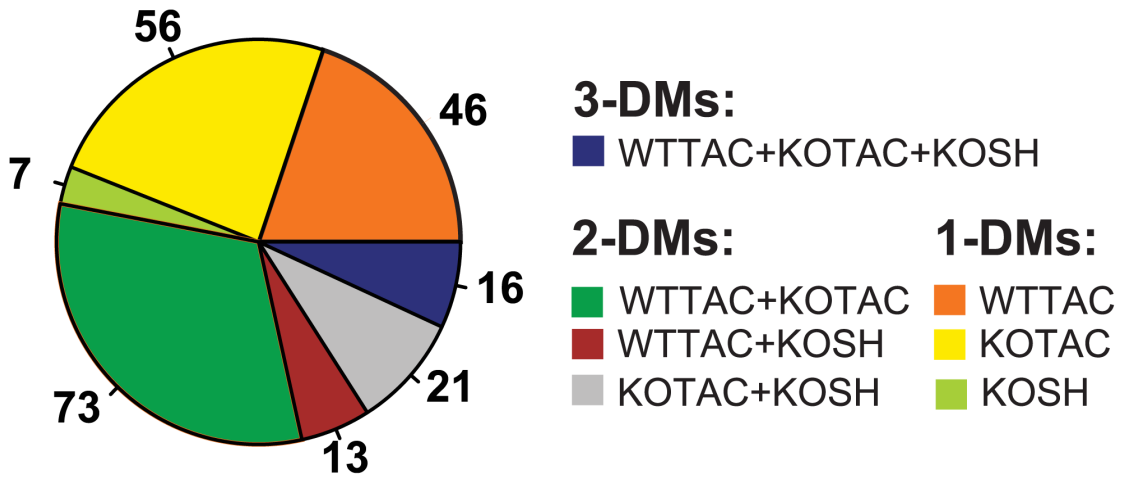

Figure 8. Application of the *i*MDM algorithm to the heart failure RNA-Seq dataset. Numbers of M-DMs detected in different DCNs. Each color represents a different type of M-DMs.

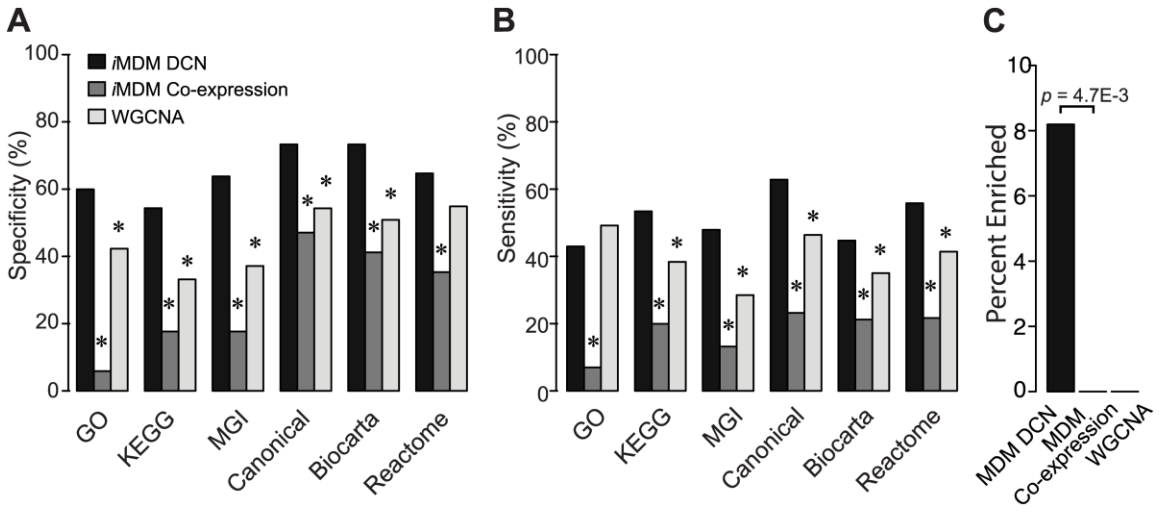

Figure 9. Performance comparison of the *i*MDM algorithm.
*i*MDM DCN, method using multiple differential co-expression networks; *i*MDM Co-expression, method using multiple co-expression networks but no differential gene expression information. A, Specificity of the algorithms. Gene modules found by each method were evaluated using a set of gold-standard pathway annotations. Specificity was defined as the fraction of predicted modules that significantly overlaps with reference pathways. B, Sensitivity of the algorithms. Sensitivity was defined as the fraction of reference pathways that significantly overlaps with predicted modules. Pathway overlap P-values were computed using the hypergeometric distribution. P-values for the difference in specificity and sensitivity were computed using Fisher's exact test. C, Percentage of predicted modules that significantly overlapped with genes whose deletions lead to cardiovascular phenotypes. P-values for the difference in the percentage of overlapped modules was computed using Fisher's exact test. All p-values were corrected for multiple testing using the method of Benjamin-Hochberg. *, *p-value* < 0.05.

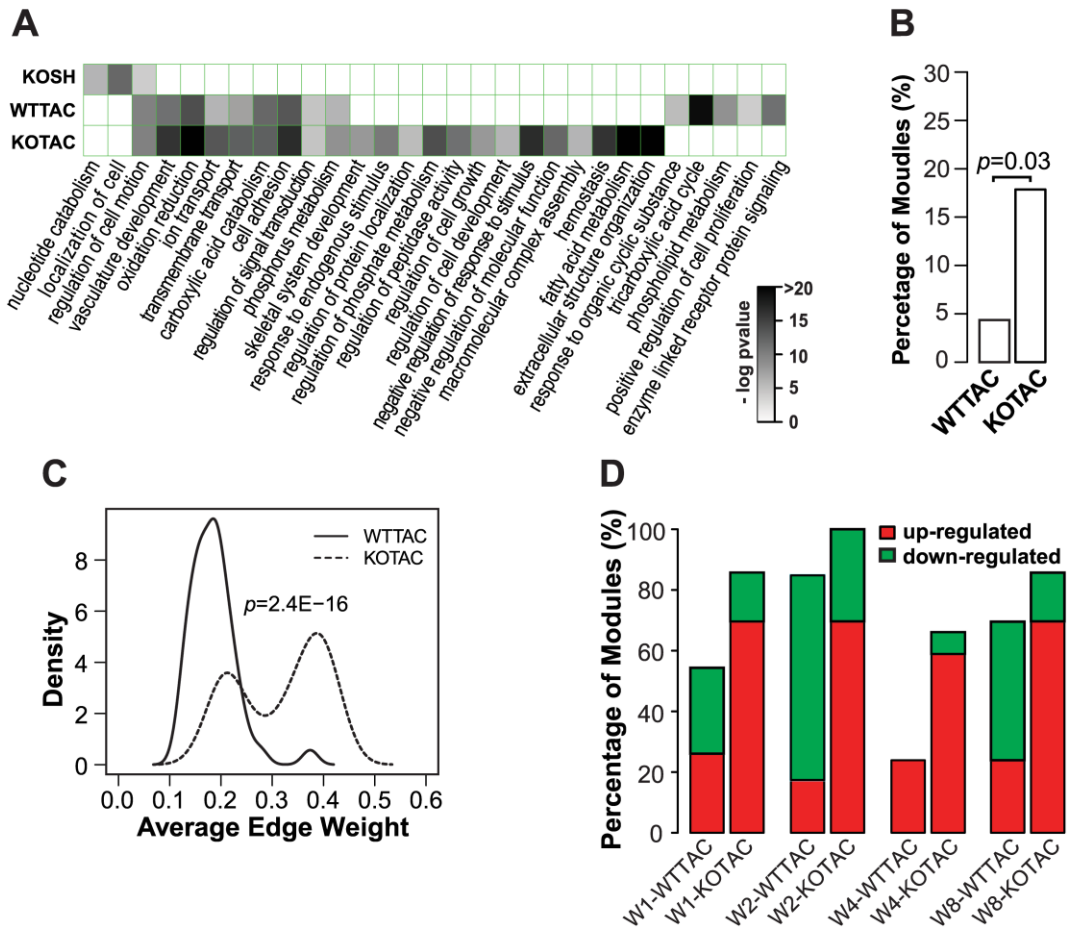

Figure 10. Global features of 1-DMs.
A, Enriched GO biological process terms in the three sets of 1-DMs. Enrichment p-value was computed using hypergeometric distribution. The grey scale is proportional to the minus logarithm of the enrichment p-value. B, Percentage of 1-DMs enriched for genes annotated to have cardiovascular system phenotypes when disrupted. C, Distributions of average edge weight of 1-DMs in WTTAC and KOTAC DCNs. D, Percentage of 1-DMs up- and down-regulated during the course of heart failure development. W1, week 1, etc.

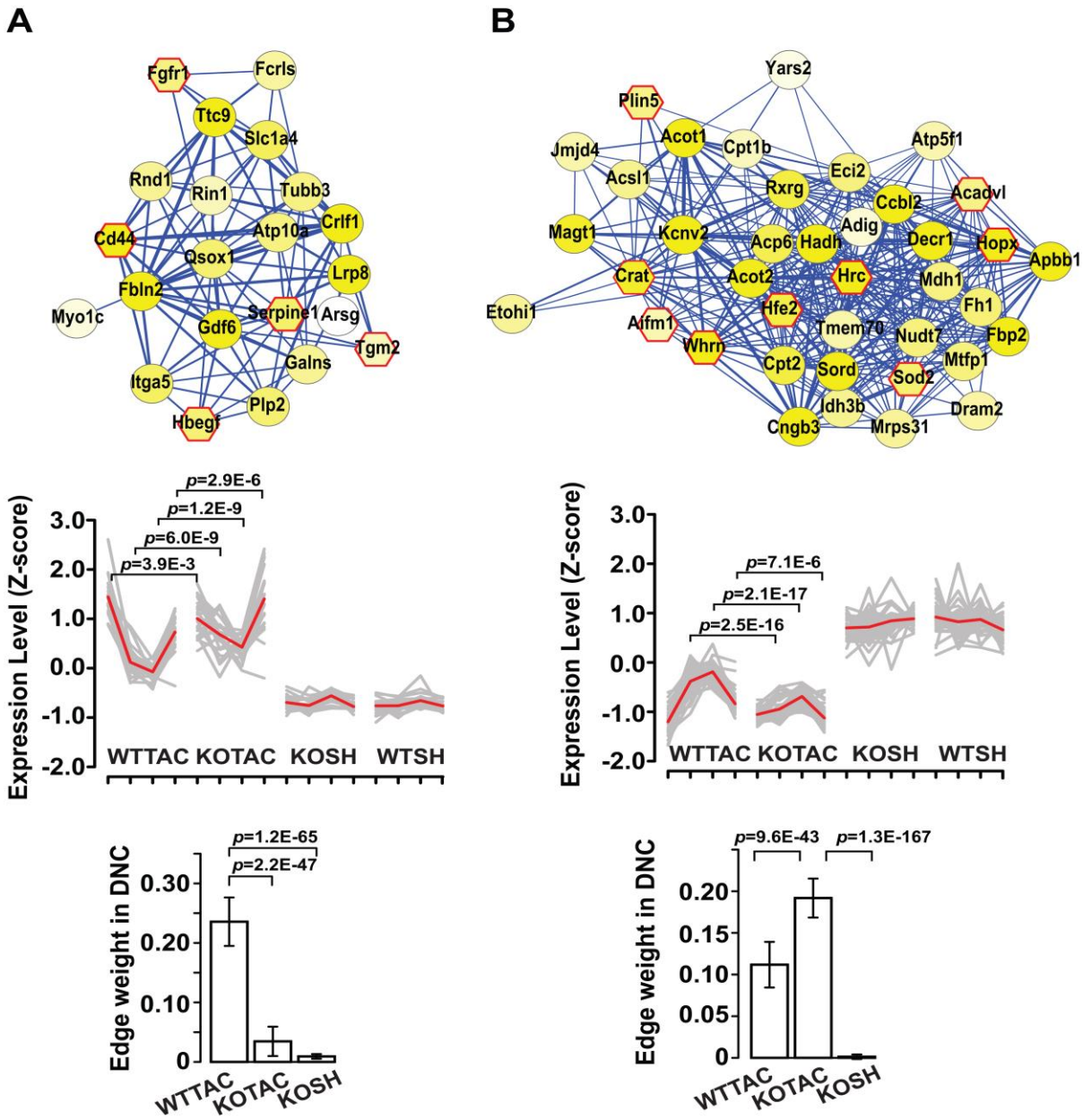

Figure 11. Example 1-DMs uniquely identified in WTTAC and KOTAC DCNs.
A, 1-DM unique to the WTTAC DCN and was enriched for genes involved in regulation of cell adhesion. B, 1-DM unique to the KOTAC DCN and was enriched for genes involved in fatty acid metabolism. Top panel, visualization of the module using Cytoscape [21]. Node color is proportional to the p-value of differential gene expression between disease and baseline (WTSH) conditions. Octagon with red border, genes whose mutations lead to cardiovascular phenotypes. Middle panel, expression profiles of module genes in four conditions. Each condition has four time points. Expression levels of each gene across all samples were normalized by Z-score transformation. P-values for the difference in gene expression level were based on t-test. Bottom panel, histogram for edge weights of discovered modules in the WTTAC and KOTAC DCNs.

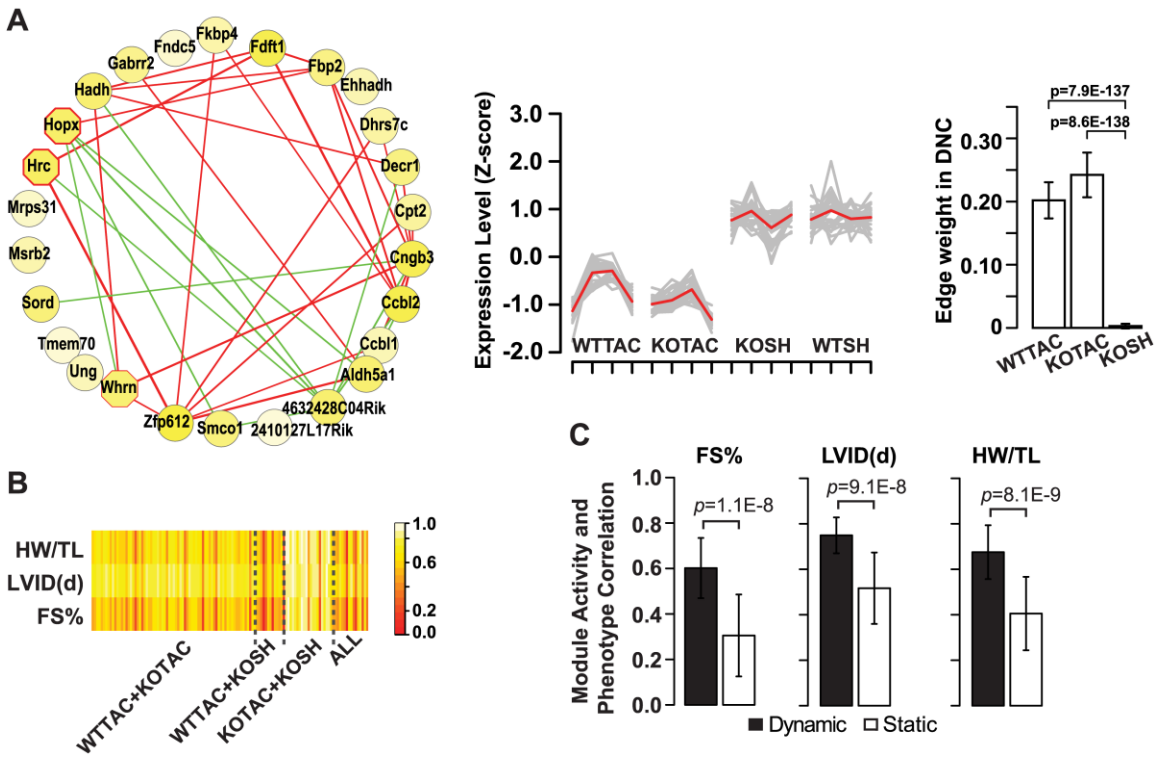

Figure 12. M-DMs identified from multiple differential co-expression networks. A, An example 2-DM identified in WTTAC and KOTAC DCNs. It was enriched for genes involved in oxidation reduction. Node color is proportional to the average p-value of differential gene expression between the two disease conditions and baseline (WTSH) condition. Octagon with red border, genes whose mutations lead to cardiovascular phenotypes. Left panel, Rewiring of the 2-DM. Only edges that exhibit significant changes in edge weights between the two DCNs are shown. Edge thickness is proportional to the absolute value of difference. Difference was calculated as "KOTAC—WTTAC". Red, increase; green, decrease. Unconnected nodes indicate there was no edge connected to the nodes that exhibit significance change in weight between the two conditions. Middle panel, expression profiles of module genes in four conditions. Each condition has four time points. Expression levels of each gene across all samples were normalized by Z-score transformation. P-values for gene expression level difference were based on t-test. Right panel, histogram for edge weights of the 2-DM in the respective networks. B, Correlation between module activity and phenotypic measures. Row, phenotypic measures; column, M-DMs. All, 3-DMs (WTTAC+KOTAC+KOSH). Module activity is the average normalized gene expression level of all member genes in a module. FS%, left ventricular fractional shortening; HW/TL, heart weight normalized by tibia length; LVID(d), left ventricular internal diameter in diastole. C, Histograms of the module activity and phenotype correlations for dynamic and static 2-DMs. P-values were based on one-sided t-test.

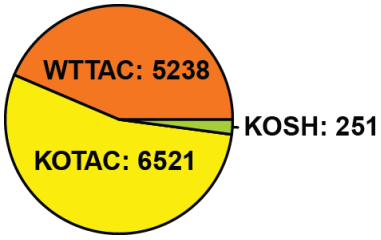

Figure 13. Number of differentially expressed genes in perturbed hearts compared to control hearts (WTSH).

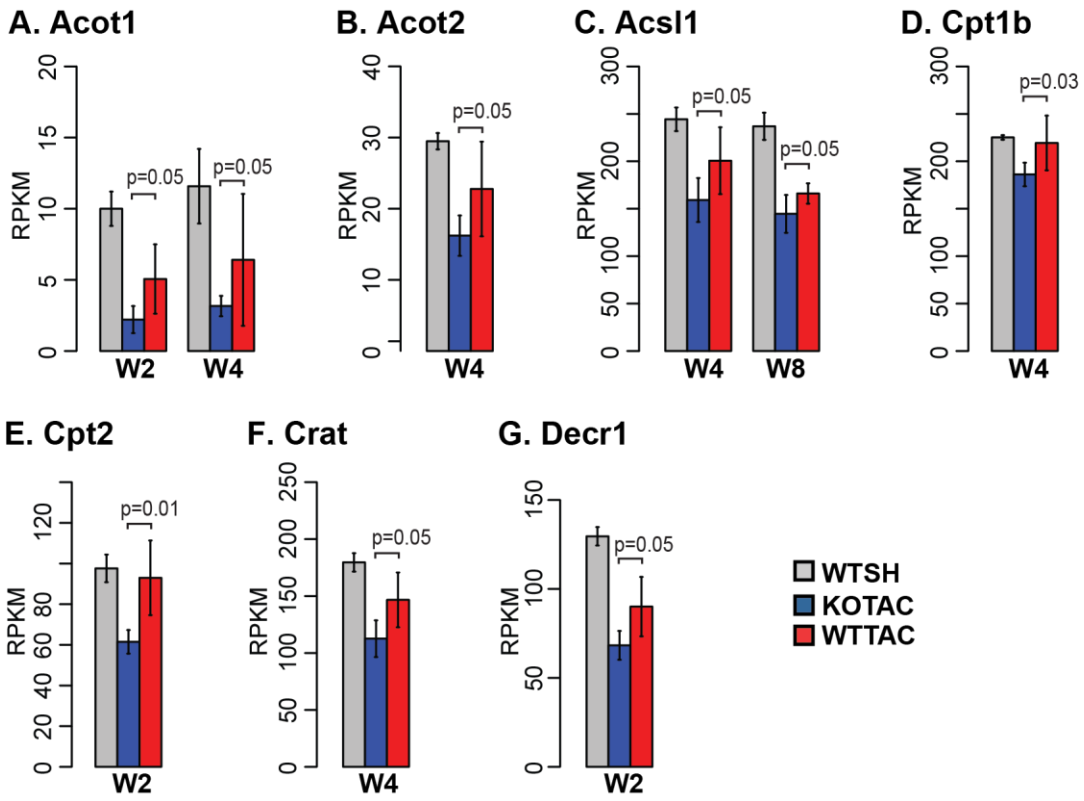

Figure 14. Expression levels of genes in an example KOTAC-specific 1-DM.

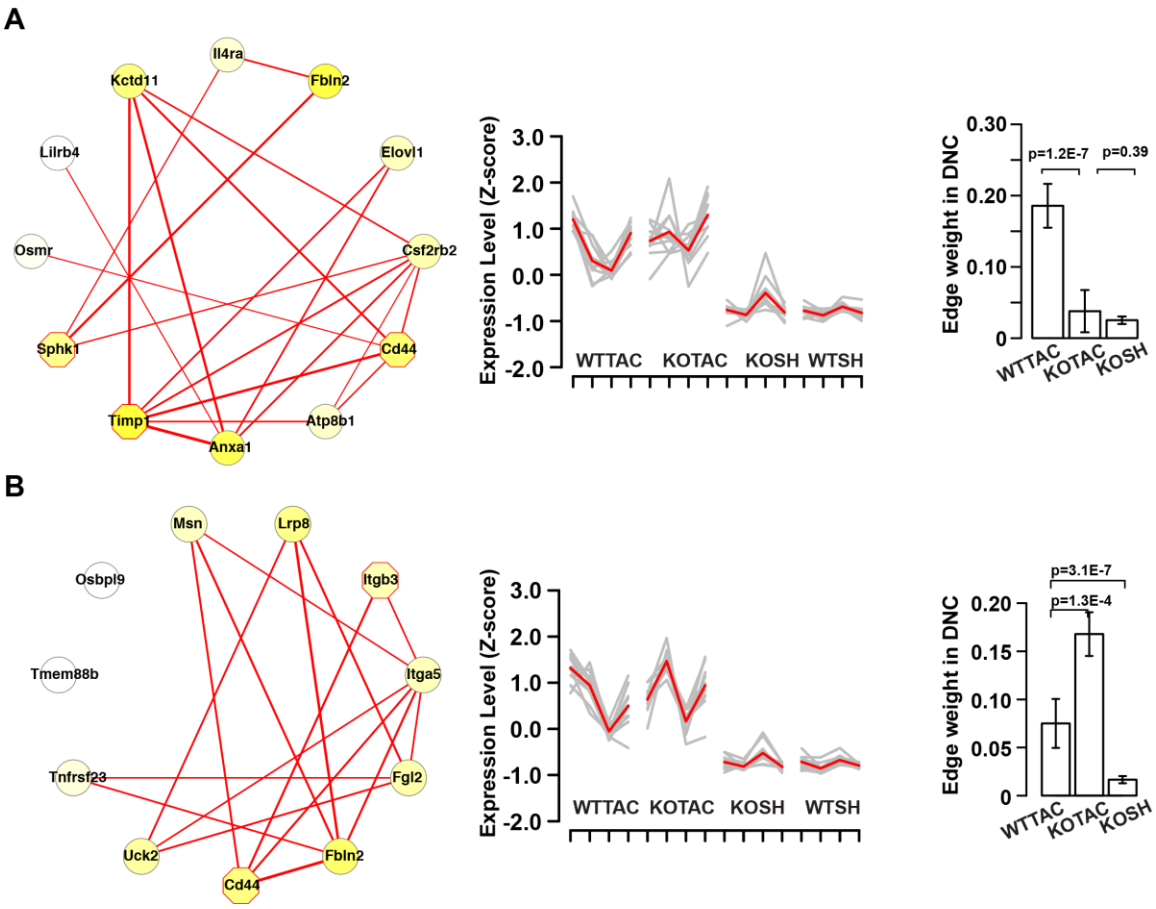

Figure 15. Example 2-DMs.
A, 2-DM found in KOTAC and KOSH DCNs. It is enriched for genes involved in the regulation of cell proliferation. Node color is proportional to the average p-value of differential gene expression between the two disease conditions and baseline (WTSH) condition. Octagon, genes whose mutation leads to cardiovascular phenotypes. Left panel, rewiring of the 2-DM. Only edges that exhibit significant changes in edge weights between the two DCNs are shown. Difference in edge weight is calculated as "KOTAC-KOSH". Red, increase, green, decrease. Unconnected nodes indicate there is no edge connected to the nodes that exhibit significance change in weight between the two conditions. Middle panel, expression profiles of module genes in four conditions. Expression levels of each gene across all samples are normalized by Z-score transformation. P-values for gene expression level difference are based on t-test. Right panel, histogram for edge weights of discovered 2-DMs in the respective networks. B, 2-DM found in WTTAC and KOSH DCNs. It is enriched for genes involved in cell migration. Difference in edge weight is calculated as "WTTAC-KOSH".

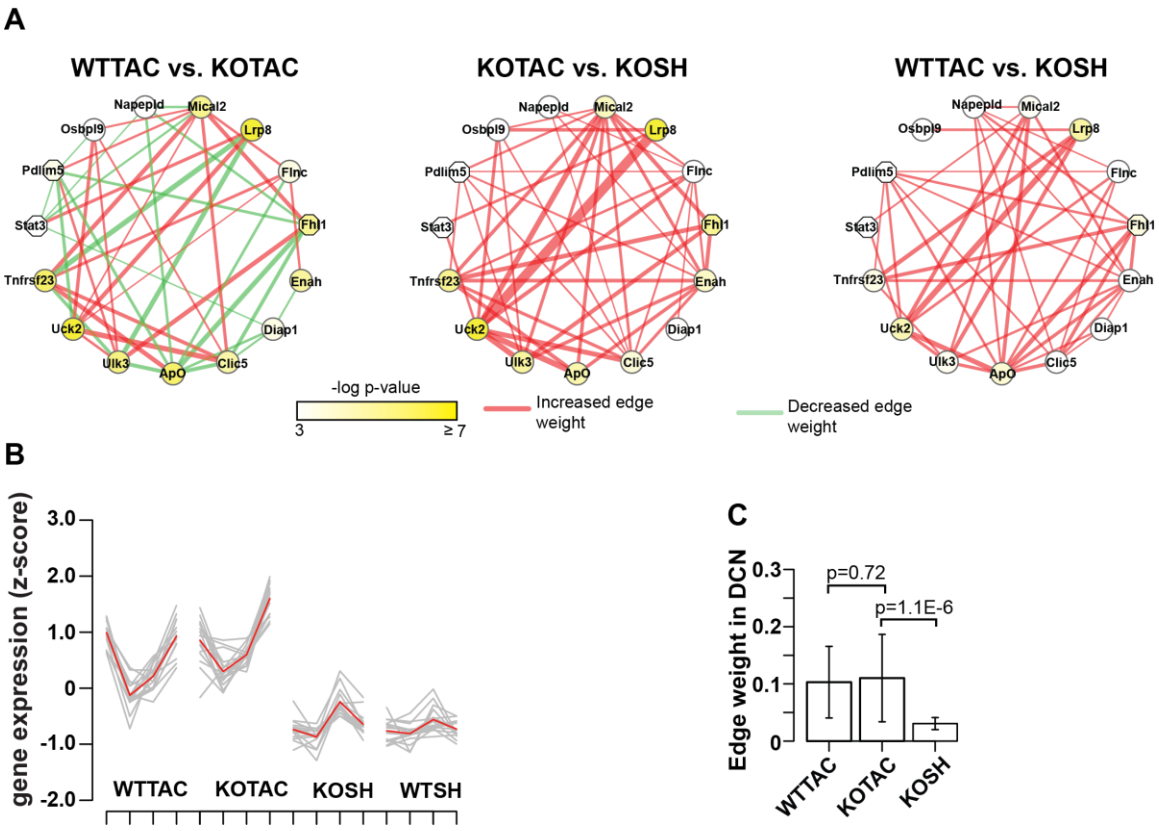

Figure 16. An example 3-DM.
It was enriched for genes involved in actin cytoskeleton organization. A. Rewiring of the 3-DM. Node color is proportional to the average p-value of differential gene expression between the two disease conditions and baseline (WTSH) condition. Octagon, genes whose mutation leads to cardiovascular phenotypes. Only edges that exhibit significant changes in edge weights between two DCNs are shown. Difference in edge weight was calculated as "KOTAC-WTTAC", "KOTAC-KOSH", and "WTTAC-KOSH". Red, increased edge weight in the comparison, green, decreased edge weight. Unconnected nodes indicate there was no edge connected to the nodes that exhibit significance change in weight between the two conditions. B. expression profiles of module genes in four conditions. Expression levels of each gene across all samples were normalized by Z-score transformation. C. Histogram for edge weights of the 3-DM in the respective networks.

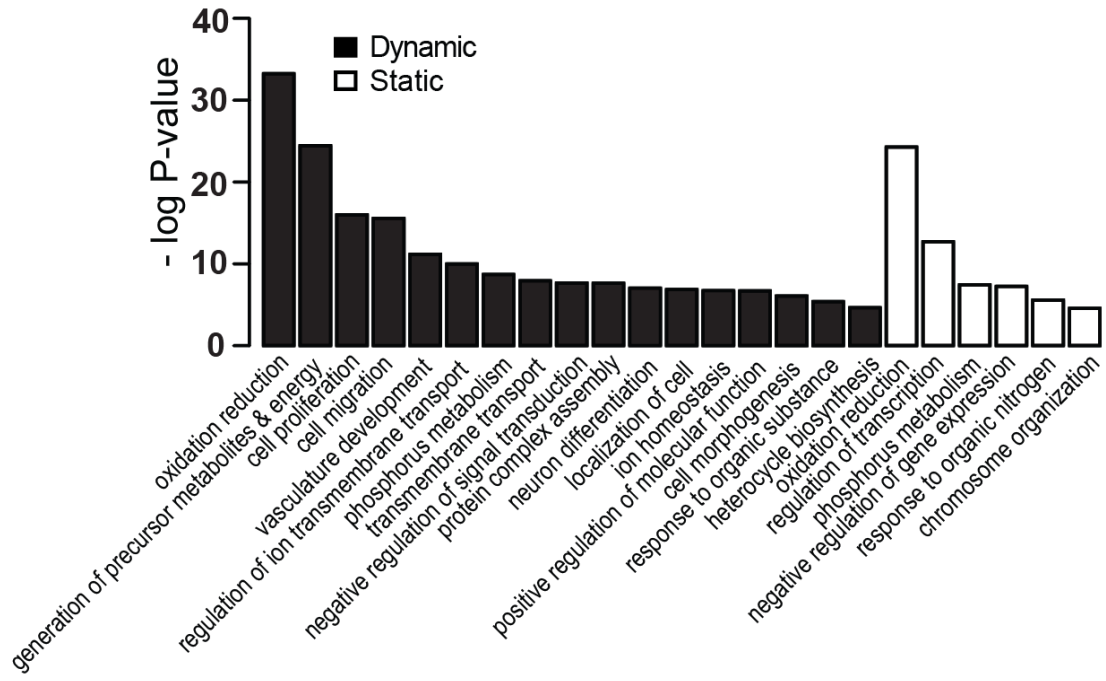

Figure 17. Enriched GO terms among dynamic and static M-DMs. Y-axis denotes the minus logarithm of the enrichment p-value.

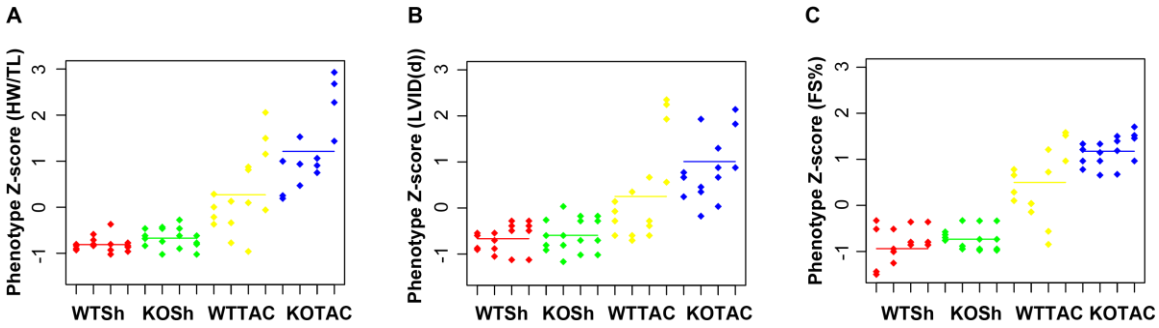

Figure 18. Heart functional measures of mice used for RNA-Seq profiling.
Each dot represents a heart. A, heart weight normalized by tibial length (HW/TL). B, left ventricular internal dimension in diastole (LVID(d). C, left ventricular fractional shortening (FS%). Because the value of FS% is between 0 and 1 and lower FS% values mean worse cardiac function whereas lower values of HW/TL and LVID(d) mean better cardiac function, we first transformed the raw FS% value as (1-FS%). Raw measures were then z-score transformed for each measurement type separately to make them comparable.

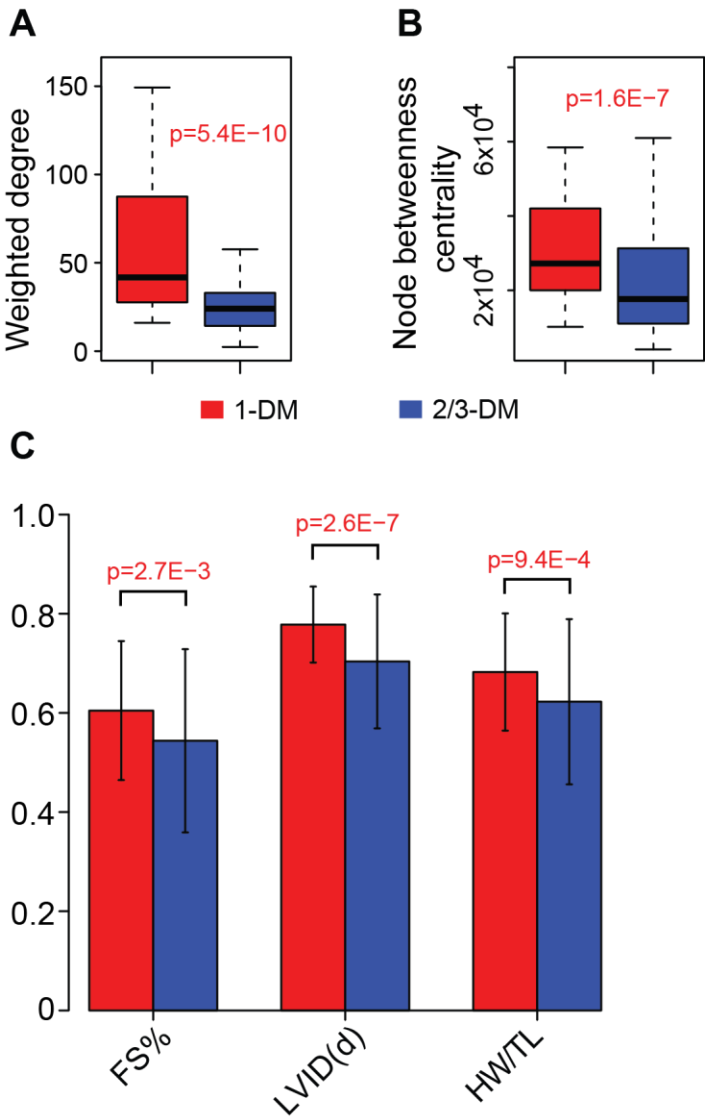

Figure 19. Topological and biological differences between 1-DMs and 2/3-DMs. A. Boxplot for the weighted degree of modules. B. Boxplot for the node betweenness centrality of modules. C. Histograms of module activity and disease phenotype correlations of the two types of M-DMs. Module activity is the average normalized gene expression level of all member genes in a module. P-values were based on one-sided t-test.

Table 1. Lists of multiple differential modules (M-DMs), one for each condition or condition combinations.

WTTAC Modules:

ID	Avg. degree	Module pval.	Module_gene
1	5.92	0.02	2310067B10Rik, Abcc9, Acad11, AI464131, Aldh5a1, Cpn2, Entpd5, Gal3st3, Gpr22, Grhl2, Klhl33, Lamb3, Lrrc15, Pde4a, Ppm1k, Sbk2, Scn4a, Slc16a10, Slc16a7, Slc22a3, Slc25a42, Smarcd1, Tbc1d24, Tbc1d4
2	7.17	0.03	2310067B10Rik, 9030617O03Rik, Acot1, Acot3, Acsl1, AI464131, Cpn2, Crhr2, Emilin2, Entpd5, Gal3st3, Hadh, Hfe2, Hspa12a, Impa2, Klhl23, Lingo3, Lpcat3, Lrrc15, Magt1, Mrgprh, Nqo2, Serac1, Slc16a10, Slc16a7, Slc5a6, Tbc1d4, Tmem135, Whrn
3	8.07	0.03	Ace, Actn1, Adcy7, Akap2, Arsg, Bhlhe40, Crlf1, Ctgf, Etv5, Fam46b, Fbln2, Fgl2, Fstl3, Galns, Gdf6, Hbegf, Itga5, Lrp8, Myo1c, Pfkp, Plp2, Qsox1, Rras2, Serpine1, Slc22a4, Tgm2, Tln1, Tnfrsf23, Tspan17, Ulk3
4	8.21	0.02	1700040L02Rik, Adhfe1, Akap7, Aldh4a1, Armc2, Asb14, Auh, Coq2, Cutc, D3Ertd751e, Dhrs4, Etfa, Hopx, Isoc2a, Ldhd, Lrrc39, Map3k5, Mipep, Phyh, Plk1s1, Ric8b, Sdhc, Sucla2, Suclg2
5	7.79	0.03	2310067B10Rik, 9030617O03Rik, Adra1b, Agpat3, Ank2, Cpn2, Crat, Crhr2, Entpd5, Gal3st3, Grhl2, Iqsec1, Jmjd4, Ky, Lingo3, Lpcat3, Lrrc15, Mme, Mrgprh, Pde4a, Sbk2, Scn4a, Serac1, Slc16a10, Slc22a3, Slc25a42, Tbc1d24, Tbc1d4, Tmem143
6	7.59	0.01	1500009L16Rik, Ace, Adamtsl2, Arfgap3, Arhgap22, Crlf1, Ctgf, Eya2, Fibin, Frzb, Fxyd6, Lrp1, Mfap5, Nox4, Nppa, Nupr1, Pkd2, Sh3rf3, Sphk1, Spry1, Tmbim1, Tspan17
7	4.68	0.04	Abca12, Abcc8, Actr3b, Arhgap5, Camk2a, Cnst, D2hgdh, Fgf13, Ghr, Gpr155, Grb14, Hlf, Kcnj3, Kcnj5, Kcnn2, Klhdc1, Ky, Mid1ip1, Nr1d2, Nr3c2, Sobp, Trmt2b
8	6.77	0.04	1500009L16Rik, Ace, Adamtsl2, Akr1b8, Bgn, Cacnb1, Crlf1, Ctgf, Dkk3, Etv5, Eya2, Fibin, Fxyd6, Gab2, Grk5, Lrp1, Lrrc16a, Nppa, Pon2, Prelp, Rasl11b, Sh3rf3, Sphk1, Spry1, Tmbim1, Tspan17
9	5.05	0.02	Acad10, Acad12, Acad8, Acsl6, Akap7, Asb14, Atg10, Auh, Cacna1s, Cdnf, Clpb, Cutc, D3Ertd751e, Dhrs4, Kcna7, Ldhd, Mpv17, Nampt, Pank1, Pdk2, Phyh, Ric8b
10	8.14	0.03	2310067B10Rik, 9030617O03Rik, Acot3, Agpat3, AI464131, Aldh5a1, Celsr2, Clcn3, Cpn2, Crhr2, Emilin2, Entpd6, Grhl2, Hfe2, Impa2, Kcnd2, Klhl23, Lingo3, Lpcat3, Lrrc15, Magt1, Mrgprh, Nqo2, Pla2g5, Slc16a10, Slc16a7, Slc25a22, Slc26a6, Tbc1d4
11	8.37	0.02	2310067B10Rik, Adra1b, Agpat3, Celsr2, Clcn3, Cox15, Dennd4b, Dus4l, Gal3st3, Iqsec1, Kcnd2, Klhl23, Klhl33, Ky, Lpcat3, Lrrc15, Mme, Oma1, Pde4a, Pigo, Sbk2, Scn4a, Slc16a10, Stat5a, Tbc1d4, Tmem135, Tmem25
12	9.00	0.03	2310010M20Rik, 3110002H16Rik, 4632428C04Rik, 9030612E09Rik, 9030617O03Rik, Aldh5a1, Celsr2, Cngb3, Cyb5rl, Dcun1d2, Dhrs7c, Dis3l, Fkbp4, G630090E17Rik, Hnmt, Klhl23, Nckap5, Nqo2, Nsmaf, Pkd2l2, Pla2g5, Plxnb1, Plxnb3, Prodh, Retsat, Slc22a5, Slc36a2, Stom, Zfp667
13	5.64	0.02	Abcc9, Acad10, Acad12, Actr3b, Adi1, Amigo1, Creg1, Dcaf12l1, Dhdh, Gpt2, Hdac11, Ldhd, Lgi3, Lsm14b, Mpv17, Mreg, Nampt, Nr1d2, Osbp2, P2ry1, Pfkm, Rps6ka5, Rxra, Slc2a12, Ttc38
14	3.67	0.03	Bmp7, Cadm4, Efcab2, Entpd5, Esrrg, Fign, Gal3st3, Gpr22, Grb14, Grhl2, Hlf, Kcnh2, Kcnj3, Kcnn2, Ky, Lrrc15, Nceh1, Pde4a, Slc16a10, Slc22a3, Tnfrsf19

Table 1 - Continued			
15	6.80	0.02	1500009L16Rik, Camkk1, Comp, Cygb, Dlg2, Fam114a1, Fibin, Fxyd6, Kbtbd11, Kctd15, Ltbp4, Med10, Mxra8, Nox4, Nupr1, Pi16, Pkd2, Pmp22, Pqlc3, Prelp, Serping1, Slc1a5, Sod3, Spry1, Tmem43
16	7.07	0.04	2310067B10Rik, 9030617O03Rik, Adra1b, Agpat3, Bmp7, Cadm4, Camk2a, Cenpf, Entpd5, Fastkd1, Fbxo3, Gal3st3, Gm6086, Gpr155, Ipo13, Iqsec1, Kcnd2, Lgr6, Lrrc15, Mme, Pigo, Ralgapa2, Sbk2, Scn4a, Slc16a10, Stat5a, Tbc1d4
17	5.96	0.02	Arsg, Atp10a, Cd44, Crlf1, Fbln2, Fcrls, Fgfr1, Galns, Gdf6, Hbegf, Itga5, Lrp8, Myo1c, Plp2, Qsox1, Rin1, Rnd1, Serpine1, Slc1a4, Tgm2, Ttc9, Tubb3
18	7.42	0.02	Adra1a, Agtpbp1, Arhgap26, Dbt, Dgat2, Ehhadh, Epb4.1, Fbxo31, Fnip1, Gm16119, Gpcpd1, Klf12, Klhl38, Mitf, Mylk4, Pcnt, Ppip5k2, Prkce, Rps6ka5, Scn4b, Slc2a4, Tgfbr3, Thrb, Wnk2
19	7.00	0.04	Atp6v1h, Camkk1, Capg, Carhsp1, Fxyd5, Glis2, Gpx1, Hn1, Kctd15, Lmna, Ltbp4, Med10, Mical1, Mtmr11, Nupr1, Panx1, Pmp22, Pqlc3, Rab23, Serping1, Slc1a5, Sod3, Tmem43
20	4.86	0.01	Actr3b, Adck3, Amigo1, Atp1a2, Creg1, Ctf1, Dcaf12l1, Gpt2, Gtf2i, Hdac11, Idh1, Il10rb, Lgi3, Mpv17, Mreg, Osbp2, P2ry1, Rxra, Slc2a12, Slc7a4, Syt3, Ttc38
21	5.46	0.01	1110034G24Rik, 2310010M20Rik, Acsm5, Ces1d, Echdc3, Fdft1, Gchfr, Hnmt, Hopx, Isoc2a, Mrpl39, Msrb2, Nckap5, Nsmaf, Pcmtd2, Phkg1, Plk5, Poln, Sord, Tcea3, Tha1, Tmod4
22	9.29	0.04	2310067B10Rik, Abcd3, Acads, Acot2, Adhfe1, AI464131, Aldh5a1, As3mt, Ccbl1, Ccbl2, Crat, Ehhadh, Gpt, Hopx, Hsdl2, Kcnj2, Klf12, Klhl33, Lamb3, Maob, Mdh1, Mylk4, Ppm1k, Rbfox1, Sgca, Slc16a10, Slc2a4, Slc5a6, Smarcd1, Sod2, Wnk2
23	7.39	0.03	2310067B10Rik, Abcc9, AI464131, Aldh5a1, As3mt, Ccbl1, Ccbl2, Eci2, Efnb3, Ehhadh, Gpr22, Grhl2, Hopx, Isca1, Klhl33, Lamb3, Lgi1, Lrrc15, Maob, Oma1, Rbfox1, Sgca, Slc16a10, Slc25a29, Slc2a4, Slc5a6, Smarcd1, Sp4
24	5.89	0.02	9030617O03Rik, Actr3b, Atp1a2, Camk2a, Cldn12, Creg1, Ctf1, Dcaf12l1, Gpr155, Gpt2, Hdac11, Idh1, Impa2, Isoc1, Lgi3, Mpv17, Npc1, Nr1d2, P2ry1, Reep1, Rps6ka5, Rxra, Slc2a12, Slc36a2, Syt3, Ttc38
25	6.88	0.02	1700040L02Rik, Acacb, Adhfe1, Atp5a1, Bzw2, Cpeb3, Dhrs11, Efnb3, Fam174b, Gpr22, Hadha, Kcnd3, Lamb3, Maob, Me3, Mitf, Nmnat3, Ppm1k, Rbfox1, Rmnd1, Slc25a29, Slc25a42, Slc27a1, Slc2a4
26	7.71	0.03	Akap2, Bhlhe40, Fbln2, Fgl2, Galns, Gdf6, Hbegf, Hectd2, Itga5, Klhl29, Lrp8, Mical2, Myo1c, Pfkp, Plcg2, Prnp, Qsox1, Rnf19b, Rras2, Serpinb1c, Serpine1, Shb, Slc38a2, Tgm2, Tln1, Tnfrsf23, Uck2, Ulk3
27	6.30	0.04	Adamts12, Arl4c, Fbln2, Galns, Gdf6, Hbegf, Hectd2, Itga5, Lox, Loxl2, Lrp8, Mical2, Myo1c, Pfkp, Plp2, Qsox1, Rin1, Rras2, Serpine1, Tgm2, Tln1, Uck2, Ulk3
28	5.41	0.02	2310067B10Rik, Adra1b, Cadm4, Cnga3, Cpn2, Entpd5, Fign, Gal3st3, Gpr155, Gpr22, Grb14, Grhl2, Iqsec1, Kcnn2, Ky, Lrpprc, Lrrc14b, Lrrc15, Mfap3l, Mme, Mrgprh, Myh6, Pde4a, Sbk2, Slc16a10, Slc22a3, Tmem135
29	6.27	0.03	2310067B10Rik, 9030617O03Rik, Acot1, Acot3, Cnga3, Cpn2, Crhr2, Entpd5, Foxo6, Gal3st3, Grhl2, Hspa12a, Impa2, Lingo3, Lpcat3, Lrrc14b, Lrrc15, Lysmd4, Magt1, Mrgprh, Mrm1, Nt5dc2, Slc16a10, Slc16a7, Slc22a3, Slc26a6, Slc5a6, Smarcd1, Sord, Tmem135
30	6.78	0.02	9030617O03Rik, Atp1a2, Celsr2, Ctf1, Emilin2, Entpd6, Gpt2, Idh1, Impa2, Kcnd2, Klhl23, Klhl33, Lgi3, Mtfr1, Nqo2, Pkd2l2, Plbd1, Plxnb3, Reep1, Rxra, Slc2a12, Slc36a2, Slc7a4, Syt3, Taf4a, Tbc1d4, Zfp667

Table 1 - Continued			
31	5.67	0.02	Acsl1, Adra1b, Ankrd32, Cadm4, Camk2a, Cnst, Entpd5, Fign, Gal3st3, Gca, Gpr22, Hlf, Kcnh2, Kcnj3, Kcnn2, Ky, Lrrc15, Mtm1, Nr3c2, Pde4a, Ppat, Sbk2, Slc16a10, Slc25a42
32	4.95	0.01	2310010M20Rik, 4632428C04Rik, AI464131, Ccbl2, Cngb3, Ehhadh, Gabrr2, Gm16119, Gpt, Klhl33, Lgi1, Lifr, Mylk4, Plxnb1, Ppip5k2, Prodh, Sec31b, Sp4, Tgfbr3, Trim7, Zfp612
33	11.21	0.05	1700040L02Rik, Adhfe1, Akap7, Armc2, As3mt, Asb14, Atp5a1, Auh, Ccbl2, D3Ertd751e, Dcaf11, Dhrs4, Egln1, Etfa, Hadha, Hopx, Idh3b, Lamb3, Lrrc39, Mccc2, Mdh1, Mipep, Nfs1, Pccb, Pdk2, Phyh, Pln, Rbfox1, Ric8b, Sdhc, Sgca, Slc25a34, Sucla2, Suclg2
34	6.29	0.03	1700040L02Rik, Acadm, Adhfe1, Atp5a1, Atpaf1, Cpeb3, Dhrs11, Fam174b, Gpr22, Hadha, Lamb3, Me3, Mitf, Nmnat3, Pcca, Pccb, Pln, Ppm1k, Slc25a11, Slc27a1, Slc2a4
35	5.36	0.02	2310067B10Rik, Adra1b, Cadm4, Cpn2, Entpd5, Fign, Gal3st3, Gpr22, Grhl2, Iqsec1, Kcnh2, Kcnn2, Ky, Lrrc15, Magt1, Mtr, Pde4a, Pdss2, Pigo, Sbk2, Slc16a10, Slc16a7, Slc22a3, Slc25a42, Smarcd1
36	5.13	0.02	Ankrd32, Cadm4, Camk2a, Cnst, Entpd5, Fign, Gal3st3, Gpr155, Gpr22, Grb14, Kcnh2, Kcnj3, Kcnn2, Ky, Lrrc15, Lynx1, Mfap3l, Paqr9, Pde4a, Pitrm1, Ppp2r3a, Sbk2, Slc22a3, Wnt5a
37	9.52	0.03	Adra1a, Agtpbp1, Aldh6a1, Arhgap26, Cacna1s, Cobll1, Dbt, Dgat2, Epb4.1, Fbxo31, Galm, Gm16119, Ldhd, Mylk4, Nampt, Nnt, P2ry1, Pank1, Pcnt, Pnpla8, Prkce, Rps6ka5, Scn4b, Slc2a4, Slco5a1, Svip, Thrb, Tmem150c, Wnk2
38	8.87	0.04	2310067B10Rik, 9030617O03Rik, Abcc9, Agpat3, Bmp7, Celsr2, Clcn3, Dus4l, Entpd6, Eral1, Impa2, Kcnd2, Klhl23, Klhl33, Lrrc15, Mme, Nqo2, Oma1, Plxnb3, Ptcd3, Scn4a, Slc16a10, Slc25a22, Slc2a12, Tbc1d4, Tmem135, Tmem143, Tmem25, Uqcrc2, Vars2, Zfp667
39	5.57	0.02	2310067B10Rik, Bzw2, Cpeb3, Cpn2, Crhr2, Entpd5, Fam174b, Gpr22, Hadha, Idh3a, Kcnn2, Lrrc15, Nmnat3, Pde4a, Ppm1k, Rap1gap2, Rtn4ip1, Sbk2, Slc16a10, Slc22a3, Slc25a42, Slc27a1, Smarcd1
40	5.21	0.02	1500009L16Rik, C1qtnf7, Camkk1, Carhsp1, Comp, Cygb, Gipc2, Hn1, Kbtbd11, Kctd15, Ltbp4, Map3k14, Med10, Mical1, Nupr1, Panx1, Pi16, Pmp22, Prelp, Radil, Rbp1, Serping1, Sod3, Wisp2
41	9.66	0.03	1700040L02Rik, Adhfe1, As3mt, Asb14, Atp5a1, C030006K11Rik, Ccbl2, Dcaf11, Dhrs11, Efnb3, Etfa, Fastk, Hopx, Lamb3, Lrrc39, Maob, Mccc2, Mdh1, Me3, Mitf, Pcp4l1, Pln, Rbfox1, Rhot2, Ric8b, Sdhc, Sgca, Slc25a29, Slc2a4
42	6.48	0.02	Adamtsl2, Arl8a, Cilp, Comp, Cygb, Dkk3, Dlg2, Eya2, Fam26e, Fibin, Fxyd6, Gab2, Il17rd, Ltbp4, Nox4, Nppa, Nupr1, Pi16, Pkd2, Pmp22, Prelp, Serpine2, Serping1, Sphk1, Spry1
43	7.75	0.03	1500009L16Rik, Adamtsl2, Cacnb1, Comp, Ctgf, Cyp1b1, Cyr61, Dkk3, Eya2, Fam26e, Fibin, Fxyd6, Gab2, Itgbl1, Lrp1, Nox4, Nppa, Nupr1, Pkd2, Prelp, Ptprn, Serpinf1, Sh3rf3, Slc30a4, Sphk1, Spry1, Tmbim1, Tspan17
44	5.19	0.03	Abca12, Abcc8, Actr3b, Adi1, Amigo1, Cnst, Creg1, D2hgdh, Dcaf12l1, Ghr, Gtf2i, Hdac11, Lgi3, Mtr, Nr1d2, Osbp2, Ppp1r26, Psap, Slc20a2, Trmt2b, Ttc38
45	6.48	0.01	Bhlhe40, Fgl2, Fhl1, Fstl3, Galns, Gdf6, Hbegf, Itga5, Lrp8, Myo1c, Pfkp, Plp2, Rras2, Slc22a4, Slc41a2, Tgm2, Tln1, Tmem88b, Tnfrsf23, Uck2, Ulk3
46	6.13	0.02	2310067B10Rik, Abcd3, Adra1b, Aifm1, Ankrd32, Cyb5r2, Entpd5, Gal3st3, Gpr22, Grhl2, Idh3a, Iqsec1, Kcnn2, Ky, Lrrc15, Mmachc, Pde4a, Sbk2, Scn4a, Slc16a10, Slc22a3, Slc25a42, Smarcd1, Tbc1d16

Table 1 - Continued

KOTAC Modules:

ID	Avg. degree	Module pval.	Module_gene
1	12.103	0.0003	1110018G07Rik, Adi1, Adra1a, Aldh6a1, Arhgap26, Asb15, Atp2a2, Bcl7a, Camk2a, Cd28, Clec18a, Dbt, Dcaf12l1, Dnajc28, Efcab2, Entpd5, Fgf13, Fyco1, Isoc1, Kbtbd7, Kcna7, L2hgdh, Ldhd, Mccc1, Mitf, Nampt, Ndufs1, Nr3c2, Osbp2, Pank1, Pcca, Pfkm, Pln, Ppm1k, Ptpn3, Rbfox1, Sdha, Tbc1d4, Thrb
2	12.026	0.0003	Acadvl, Acot1, Acot2, Acp6, Acsl1, Adig, Aifm1, Apbb1, Atp5f1, Ccbl2, Cngb3, Cpt1b, Cpt2, Crat, Decr1, Dram2, Eci2, Etohi1, Fbp2, Fh1, Hadh, Hfe2, Hopx, Hrc, Idh3b, Jmjd4, Kcnv2, Magt1, Mdh1, Mrps31, Mtfp1, Nudt7, Plin5, Rxrg, Sod2, Sord, Tmem70, Whrn, Yars2
3	11.485	0.0003	3110057O12Rik, Abcc9, Adi1, Adra1a, Agl, Asb15, Atp2a2, Bcl7a, Camk2a, Cd28, Clec18a, Dbt, Dnajc28, Efcab2, Fastkd1, Fyco1, Kcna7, Kcnip2, L2hgdh, Nampt, Nr3c2, Pank1, Pcca, Pdk2, Pln, Ppp2r3a, Prkce, Prlr, Ptpn3, Rbfox1, Sgol2, Slc22a5, Tbc1d4
4	12.2	0.0003	0610009O20Rik, 1700040L02Rik, 3110057O12Rik, Acadm, Adhfe1, Adi1, Adra1a, Aldh6a1, Asb14, Bcl7a, Blcap, Clec18a, Coq9, Dbt, Echs1, Etfa, Etfdh, Fundc2, Fyco1, Hadha, Hspa5, L2hgdh, Ldhd, Mlycd, Mylk4, Pank1, Pcca, Pln, Ppm1k, Pxmp4, Ric8b, Sdha, Slc25a34, Thrb, Tnni3k
5	8.545	0.0003	2310010M20Rik, 3110002H16Rik, 9030617O03Rik, Acot1, Acot3, Adra1b, Aldh5a1, Ank, Cadm4, Camk2a, Car14, Gabrr2, Gal3st3, Idh1, Impa2, Klhl33, Lpcat3, Lrrc15, Mme, Pdp2, Plbd1, Plxnb1, Plxnb3, Pomt1, Qsox2, Retsat, Scn4b, Serac1, Slc25a22, Slc36a2, Stard10, Tmem25, Whrn
6	7.909	0.0003	2200002D01Rik, Ace, Anxa3, Arfgap3, Ass1, Atp6v1h, Bcr, Bgn, Chpf2, Cttn, Cx3cl1, Dap, Dlgap4, Evi5, F2r, Foxc2, Frzb, Fxyd6, Gm5424, Ltbp4, Mical1, Msn, Mtpn, Myo1d, Nupr1, Plekha4, Sgms2, Slc16a3, Slc1a4, Tmem43, Trim47, Unc5b, Zyx
7	17.5	0.0003	Actn1, Anxa3, Ap3s1, Arf3, Arhgef40, Atp8b2, B4galnt1, Bgn, Cdr2l, Cenpt, Col18a1, Dap, Dnm3os, Ecm1, Edem1, Emilin1, Emp1, Fam129b, Fstl1, Gba, Loxl1, Meox1, Mmp23, Nek6, Numbl, Pcolce, Phlda3, Pi4k2b, Ppp1r9b, Ptgis, Rai14, Rcc2, Rsu1, Sdcbp, Sept5, Serpine2, Serpinf1, Sh3gl1, Sparc, Tgfbr2, Tgif1, Zbtb7c
8	17.83	0.0003	Acot10, Actn1, Adcy7, Anxa4, Arhgap1, Arhgef40, Arsb, B4galnt1, Bgn, Ckap4, Col5a2, Col8a1, Colec12, Enpp1, Ephb6, Fam129a, Farp1, Fbln5, Fgfr1, Gpc6, Jazf1, Kcnj15, Leprel2, Lox, Loxl3, Lrrk1, Ltbp3, Megf10, Mgp, Mrc2, Myof, Naalad2, Orai2, Plp2, Postn, Ptprf, Rab31, Rasa3, Rnf149, Runx1, Slc1a3, Slc41a2, Srpx2, Taok3, Tgfbr1, Tnfaip6, Tulp3

Table 1 - Continued

9	10.757	0.0003	9030617O03Rik, Acot1, Acot3, Adra1b, Agpat3, Aldh5a1, Asb15, Clcn1, Cpn2, Crat, Dhrs11, Entpd6, Gal3st3, Hfe2, Idh3a, Impa2, Jmjd4, Kcnh2, Kcnv2, Lingo3, Lpcat3, Lrrc14b, Lrrc15, Magt1, Mme, Oma1, Pdp2, Plbd1, Plxnb1, Retsat, Rxrg, Slc22a3, Slc36a2, Tmem25, Ttc38, Whrn, Zfp759
10	12.914	0.0003	Adcy7, Aebp1, Anxa4, Aspn, Col12a1, Col5a2, Col8a2, Colec12, Crispld1, Ephb6, Eya2, Fam129a, Fat1, Fbln5, Ism1, Itgb5, Itgbl1, Kcnj15, Klhl29, Lox, Loxl3, Ltbp3, Naalad2, Omd, Pamr1, Pcdh9, Pkd2, Ptprf, Runx2, Sfrp2, Sh3d19, Srgap3, Tgfbr1, Tnfaip6, Tro
11	16.902	0.0003	Bcat1, Bmp1, C1qa, Capza1, Ccdc80, Cercam, Clip3, Col15a1, Col3a1, Col5a1, Col6a1, Col6a2, Ctss, Dlg4, Eln, Emilin1, Fbn1, Fstl1, Gpc6, Gpr153, Gpx8, Itih5, Kdelr3, Lepre1, Leprel2, Lhfpl2, Lpar1, Ly86, Mrc2, Mxra7, Nid1, Oaf, Pcolce, Rcn3, S1pr2, Sept11, Soat1, Sox9, Sparc, Twist1, Txndc5
12	15.75	0.0003	Aebp1, Akr1b8, Ano10, Anxa4, Aspn, C1qtnf6, Chpf, Clu, Col14a1, Col16a1, Col8a2, Dact3, Ddah1, Des, Enpp1, Ephb6, Fam198b, Gpx3, Gxylt2, Itga11, Itgbl1, Lox, Ltbp3, Mvp, Myh10, Pamr1, Plp2, Pycr1, Rnf149, Rtn4, Runx1, Scn1b, Sfrp2, Shisa4, Srgap3, Srpx2, Star, Tpd52, Uchl1, Wisp1
13	15.976	0.0003	2610002J02Rik, Aff3, Anxa3, App, Bend6, Bgn, C1ra, C1s, Cd52, Ctsk, D630003M21Rik, Dok1, Ecm1, Eid1, Fam20a, Gm6548, Gpm6b, Islr, Lhfp, Loxl1, Megf10, Mgp, Mmp23, Mxra8, Numbl, Olfml3, Pmepa1, Ppic, Prss23, Pstpip1, Ptgis, Rbp1, Serpinb6a, Serpinf1, Sfrp1, Thbs3, Timp2, Tmem176b, Tmem45a, Tspo, Vat1, Zbtb7c
14	11.286	0.0003	Adcy7, Antxr1, B4galnt1, Col12a1, Col16a1, Dpysl3, Enpp1, Ephb6, Eya2, Fam129a, Fat1, Fgfr1, Ildr2, Itgb5, Itgbl1, Klhl29, Lox, Loxl3, Ltbp2, Ltbp3, Myof, Pamr1, Pcdh9, Pkd2, Pon2, Prmt2, Ptprf, Qsox1, Rnf149, Shc4, Shisa3, Slc41a2, Sulf1, Svep1, Tgfbr1
15	12.946	0.0003	5430435G22Rik, Anxa1, Anxa5, Apbb1ip, Arhgef40, Axl, Bgn, Coro1a, Csf2rb, Ctdspl, Ecm1, Edem1, Efemp2, Emp1, F2r, Fam129b, Gba, Iqgap1, Layn, Mfap5, Mob3a, Myo1d, Nek6, Numbl, Pde1a, Pdpn, Ptgis, Rai14, Rtkn, Serpine2, Sox9, Tgfbr2, Tgif1, Tmem43, Tmem45a, Zbtb7c, Zyx
16	15.333	0.0003	Adamts2, Arsb, Bcat1, Ccrl1, Clec11a, Col1a1, Col1a2, Col3a1, Col5a1, Col5a2, Col6a1, Col6a2, Col8a1, Ctsk, Ctss, Dclk1, Dlg4, Dnm1, Dse, Enpp1, Fam198a, Fbn1, Fn1, Fndc1, Fzd1, Gli2, Gpc6, Gpr153, Gria3, Itih5, Lepre1, Mrc2, Myof, Naalad2, Pcolce, Postn, Ptgfrn, Tnfaip6, Zmat3
17	9.667	0.0003	1700040L02Rik, Adhfe1, As3mt, Asb14, Atp5a1, BC025920, Cecr2, Dcun1d2, Dhrs4, Efnb3, Epha4, Etfa, Fign, Gm1078, Gpr22, Hadh, Hadhb, Hnmt, Hopx, Lrrc39, Mccc2, Mdh1, Mitf, Mov10l1, Oxa1l, Phyh, Pln, Ppm1k, Ric8b, Rxrg, Sdhc, Svip, Tnni3k

Table 1 - Continued

18	12.375	0.0003	2310067B10Rik, 9030617O03Rik, Abcc9, Acot1, Acsm5, Aldh5a1, Atp2a2, Crat, Dennd4b, Dis3l, Galm, Hibadh, Ipo13, Kcnip2, Kcnj3, Maob, Mme, Myh6, Nsmaf, Osgepl1, P2ry1, Pdk2, Pdp2, Pkd2l2, Plbd1, Plxnb1, Plxnb3, Retsat, Rhot2, Rmnd1, Scn4a, Slc22a5, Slc25a12, Slc2a12, Slc36a2, Tmem150c, Tmem65, Trim7, Ttc38, Whrn
19	9.545	0.0003	9030617O03Rik, Acot1, Acot3, Adig, Adra1b, Aldh5a1, Car14, Ccbl2, Clcn1, Cpn2, Dhodh, Dlst, Gal3st3, Hfe2, Idh1, Impa2, Kcnv2, Lamb3, Lingo3, Lrrc15, Magt1, Nqo2, Plbd1, Plxnb1, Rxrg, Sbk2, Slc16a7, Slc22a3, Slc36a2, Stard10, Whrn, Zfp629, Zfp759
20	15.3	0.0003	2200002D01Rik, 2610002J02Rik, Anxa3, App, Arhgef40, Bgn, Cenpt, Ctdspl, Ecm1, Emp3, F2r, Fam129b, Gba, Megf10, Meox1, Mfap5, Mgp, Mmp23, Mob3a, Myo1d, Numbl, Olfml3, Phlda3, Pmepa1, Ptgis, Rai14, Serpinb6a, Serpine2, Serpinf1, Sfrp1, Sox9, Tgif1, Timp2, Tmem43, Tmem45a, Tnfrsf1b, Twist1, Vat1, Zbtb7c, Zyx
21	11.029	0.0003	0610009O20Rik, 1700040L02Rik, 3110057O12Rik, Acadm, Acads, Acat1, Adhfe1, Asb14, Auh, Coq9, Crat, Dbt, Dcaf11, Etfdh, Fkrp, Fundc2, Hadha, Hspa5, Ldhd, Mipep, Mylk4, Ndufs1, Pcca, Plin5, Pln, Pxmp4, Ric8b, Sdhc, Slc22a5, Slc25a11, Slc25a20, Slc25a34, Stom, Suclg2, Tnni3k
22	15.405	0.0003	Anxa1, Anxa5, Bace2, Cd109, Cd63, Creb3, Ctdspl, Dap, Dchs1, Ecm1, Elmo1, Fam129b, Fscn1, Gba, Gm5506, Igfbp7, Irf8, Meox1, Mfap5, Mob3a, Mtmr11, Myo1d, Npdc1, Olfml3, Pcdhb17, Pdpn, Plk3, Pqlc3, Ptgis, Rab4b, Rai14, Rbp1, Rtkn, S100a11, Serpine2, Serpinf1, Tmem45a, Tnfrsf1b, Tsku, Vat1, Zbtb7c, Zyx
23	14.857	0.0003	0610009O20Rik, 1700040L02Rik, 3110057O12Rik, Abcc9, Acadm, Adhfe1, Adra1a, Agpat9, Asb14, Atp2a2, Auh, Camk2a, Coq9, Crat, Dbt, Echs1, Etfa, Etfdh, Gfm1, Got2, Hadha, Ldhd, Mitf, Mlycd, Mylk4, Pank1, Paqr9, Pcca, Pdk2, Pfkm, Pln, Ppm1k, Prlr, Pxmp4, Rbfox1, Ric8b, Scn4a, Slc22a5, Slc25a20, Tmem143, Tnni3k, Usp2
24	10.424	0.0003	1700040L02Rik, Acadm, Acat1, Adhfe1, Asb14, Atp5a1, BC025920, Chrna2, Coq9, Dbt, Dcaf11, Dcun1d2, Dhrs4, Echs1, Etfa, Etfdh, Fundc2, Gnpat, Hadha, Hspa5, Ldhd, Lrrc39, Mlycd, Mylk4, Pcca, Peo1, Pln, Ppm1k, Pxmp4, Ric8b, Slc25a11, Slc25a34, Suclg2
25	11.162	0.0003	1700040L02Rik, Acot1, Acot2, Acp6, Acsm5, Adig, As3mt, BC025920, Ccbl2, Cecr2, Clcn1, Cngb3, Cpt2, D3Ertd751e, Decr1, Efnb3, Fbp2, Fndc5, Hadh, Hfe2, Hopx, Hrc, Hsdl2, Kcnv2, Lrrc15, Lrrc39, Magt1, Mdh1, Nudt7, Plbd1, Rxrg, Slc36a2, Sord, Sucla2, Svip, Tmem70, Whrn

Table 1 - Continued

26	16.744	0.0003	Adcy7, Anxa4, Bgn, Col16a1, Cyb5r3, Dlg2, Dnm1, Dok1, Ephb6, Fam198b, Farp1, Fbln5, Fgfr1, Gmfb, Gpx3, Gxylt2, Islr, Itgbl1, Kdelr3, Kirrel, Lox, Ltbp3, Megf10, Mpeg1, Nkd2, Pamr1, Plp2, Plxdc2, Postn, Rab31, Rasa3, Rtn4, Runx1, Sfrp1, Sfrp2, Slc1a3, Slc41a2, Srpx2, Taok3, Tgfbr1, Timp2, Tpd52, Uchl1
27	12.03	0.0003	2200002D01Rik, 2610002J02Rik, Ass1, Bgn, Cd63, Ctdspl, Cyb5r3, Dok1, Eid1, Emp1, Gpx8, Lgals1, Megf10, Mgp, Nkd2, Olfml3, Pmepa1, Pqlc3, Prss23, Ptgis, Ptprn, Radil, Rbp1, Serpinf1, Sfrp1, Timp2, Tmem43, Tmem45a, Tnfrsf1b, Trim46, Tspo, Vat1, Zbtb7c
28	13.606	0.0003	Anxa5, Arf3, Atp8b2, Bak1, Bin1, Cdr2l, Cenpt, Clip2, Ctsz, Dap, Ecm1, Emp1, Fam129b, Gba, Kdelr3, Lxn, Mcm6, Meox1, Mmp23, Oaf, Pcolce, Phlda3, Ppp1r9b, Rab4b, Rcc2, Rhoc, S100a11, Serpine2, Serpinf1, Sh3gl1, Sparc, Tgif1, Vim
29	13.189	0.0003	2200002D01Rik, 2610002J02Rik, App, Arhgap22, Arl4c, Bend6, C1ra, C1s, Ctdspl, Cyb5r3, Dok1, Eid1, Fam20a, Islr, Mapk7, Megf10, Mgp, Mmp23, Myo1d, Nkd2, Olfml3, Phlda3, Pmepa1, Prss23, Pstpip1, Ptgis, Ptprn, Rbp1, Serpinb6a, Sfrp1, Spsb2, Timp2, Tmem43, Tmem45a, Tspo, Vat1, Zbtb7c
30	16.605	0.0003	Actn1, Adamts2, Arsb, B4galnt1, Bcat1, Bmp1, Capza1, Ckap4, Col15a1, Col3a1, Col5a1, Col6a1, Col6a2, Col8a1, Dclk1, Eln, Emilin1, Emp1, Epb4.1l2, Fbn1, Fcgr1, Fjx1, Fn1, Fstl1, Fzd1, Gpc6, Gpr153, Gpx8, Itih5, Lepre1, Lpar1, Marcksl1, Mfap4, Mrc2, Myof, P2rx7, Postn, Prrg3, Ptgfrn, Sept11, Soat1, Txndc5, Ubtd2
31	13.079	0.0003	Anxa3, Arhgef40, Ass1, Bend6, Bgn, C1ra, C1s, Ctdspl, Dchs1, Ecm1, F2r, Fam129b, Gba, Gpm6b, Megf10, Mfap5, Mgp, Mob3a, Mpzl1, Myo1d, Nucb2, Olfml3, Pcdhb17, Pdpn, Pqlc3, Ptgis, Rai14, Rbp1, Serpinf1, Sfrp1, Timp2, Tmem43, Tmem45a, Tnfrsf1b, Tuba1a, Vat1, Zbtb7c, Zyx
32	11.359	0.0003	3110057O12Rik, Acadm, Aco2, Acss1, Adhfe1, Aldh4a1, Aldh6a1, Bcl7a, Blcap, Chrna2, Clec18a, Dbt, Dsg2, Echs1, Egln1, Epm2aip1, Etfdh, Fam174b, Fitm2, Fyco1, Gbas, Hadha, Hspa5, Ivd, L2hgdh, Ldhd, Mtm1, Mylk4, Ndufs1, Nnt, Pank1, Pcca, Pdss2, Pkia, Ppm1k, Sdha, Sec31b, Slc25a34, Slc27a1
33	10.727	0.0003	1110038B12Rik, 2200002D01Rik, Acot9, Arhgef40, Ass1, Bgn, Ctdspl, Cyb5r3, Dok1, Eid1, F2r, Gmfb, Megf10, Meox1, Mgp, Msn, Myo1d, Nkd2, Npc2, Olfml3, Pmepa1, Ptgis, Ptprn, Rbp1, Rcc2, Slc16a3, Tmem43, Tmem45a, Trim46, Tspo, Unc5b, Vat1, Zbtb7c
34	12.351	0.0003	6330416G13Rik, Abcd3, Acad11, Acot1, Acsl1, Aldh6a1, Arhgap26, Atp2a2, Crat, Csdc2, Dennd4b, Dld, Dpyd, Etfdh, Etohi1, Fbxo31, Gpsm1, Hadha, Hsdl2, Jmjd4, Kcnv2, Mtfr1, Mylk4, Nceh1, Nudt12, P2ry1, Pdp2, Ppip5k2, Rhot2, Rmnd1, Sec31b, Slc25a12, Slc25a20, Slc2a4, Tbc1d16, Tmem143, Wnk2

Table 1 - Continued

35	17.244	0.0003	Adcy7, Anxa4, Arsb, Col14a1, Col16a1, Col5a2, Col8a2, Colec12, Ddah1, Dlg2, Enpp1, Ephb6, Fam129a, Fam198a, Fbln5, Fcgr4, Fgfr1, Fndc1, Gli2, Gpx3, Gxylt2, Ism1, Itgbl1, Kcnj15, Kirrel, Klhl29, Lox, Lrrk1, Ltbp3, Naalad2, Olfm1, Pamr1, Plp2, Rasa3, Sfrp2, Slc1a3, Srgap3, Srpx2, Taok3, Tgfbr1, Tmem119
36	11.143	0.0003	1700040L02Rik, 3110057O12Rik, Acat1, Adhfe1, As3mt, Asb14, Atp5a1, BC025920, Cecr2, Clec18a, D3Ertd751e, Dhrs4, Efnb3, Epha4, Etfa, Fundc2, Hadh, Hadha, Hadhb, Hibadh, Hopx, Hrc, Idh3b, Lrrc39, Mipep, Pdss2, Phf20l1, Phyh, Pln, Ppm1k, Pxmp4, Ric8b, Rxrg, Slc25a11, Suclg2
37	9.788	0.0003	2200002D01Rik, 2410006H16Rik, 2610002J02Rik, Acot9, Arfgap3, Arhgap22, Arl4c, Bend6, Creb3, Cx3cl1, Cyb5r3, Dok1, Ehmt2, Eid1, Frzb, Fxyd6, Gm5424, Klc1, Map3k14, Mgp, Myo1d, Nkd2, Phlda3, Pmepa1, Pold4, Ptprn, Rbp1, Slc16a3, Slc1a4, Spsb2, Tspo, Unc5b, Vat1
38	10.4	0.0003	2310067B10Rik, Acot1, Acot3, Adra1b, Agbl2, Aldh5a1, Cpn2, Dis3l, Gal3st3, Galm, Ipo13, Kcnip2, Kcnj3, Klhl33, Lingo3, Lrrc15, Maob, Mfsd7c, Mme, Myh6, Pdp2, Pkd2l2, Plxnb1, Ptgr2, Retsat, Rmnd1, Scn4a, Slc22a3, Slc2a12, Tarsl2, Tmem135, Tmem25, Ttc38, Uqcc, Whrn
39	13.667	0.0003	Anxa2, Arhgdia, Arpc1b, Arpc3, Bace2, Capg, Carhsp1, Cd34, Ch25h, Ddx39, Elmo1, Endod1, Entpd2, Fxyd5, Gm5506, Kctd15, Litaf, Lmna, Npc2, Npdc1, Pdpn, Pi16, Plat, Pmp22, Ppp1r18, Ptma, Rap1b, Rhoc, S100a11, Tagln2, Tax1bp3, Tnfrsf1a, Tsku
40	14.308	0.0003	Anxa1, Arhgef40, Ass1, Bgn, Cd63, Clip2, Creb3, Ctdspl, Dap, Dpep2, Ecm1, Fam129b, Gba, Igfbp7, Lpcat2, Mfap5, Mgp, Mob3a, Myo1d, Nckap5l, Numbl, Olfml3, Phlda3, Plk3, Pmepa1, Pqlc3, Ptgis, Ptms, Rai14, Rbp1, Sfrp1, Tmem176a, Tmem43, Tmem45a, Tnfrsf1b, Trim46, Vat1, Zbtb7c, Zyx
41	19.48	0.0003	Adamts2, Aff3, Anxa4, Arsb, Aspn, Bgn, Cercam, Col14a1, Col27a1, Ctsk, Ctss, D630003M21Rik, Dnm1, Eln, Fam198a, Farp1, Fbln5, Fcgr4, Gpc6, Gpm6b, Gpr124, Gpr153, Islr, Leprel2, Loxl1, Lrrk1, Megf10, Mfap4, Mmp23, Mrc2, Mxra7, Myof, P2rx7, Pdgfrl, Plp2, Plxdc2, Postn, Rab31, Rasa3, Renbp, Rsu1, Runx1, S1pr2, Scpep1, Shc2, Slc1a3, Srpx2, Thbs3, Tmem119, Twist1
42	18.568	0.0003	Adamts2, Adcy7, Aebp1, Anxa4, Col14a1, Col16a1, Col8a2, Colec12, Ddah1, Enpp1, Ephb6, Fam198b, Farp1, Fbln5, Fcgr4, Fgfr1, Fzd1, Fzd2, Gpx3, Gxylt2, Itga11, Lox, Ltbp3, Naalad2, Pdgfrl, Plp2, Rab31, Rasa3, Rgs10, Rnf149, Rtn4, Runx1, Sfrp2, Shc2, Slc1a3, Srgap3, Srpx2, Star, Taok3, Tgfb3, Tmem119, Tns3, Uchl1, Wisp1

Table 1 - Continued

43	11.424	0.0003	Ankrd27, Anxa1, Anxa2, Arhgef40, Axl, Bean1, Ctdspl, Dap, Dchs1, F2r, Fbxw17, Gba, Igfbp7, Inpp5d, Iqgap1, Layn, Ltbp4, Meox1, Mfap5, Mob3a, Myo1d, Nckap5l, Nfkbiz, Panx1, Pdpn, Ptgis, Rai14, Rtkn, S100a11, Tmem43, Tnfrsf1b, Zbtb7c, Zyx
44	17.432	0.0003	Adcy7, Aebp1, Anxa4, Arsb, Aspn, Clec11a, Col14a1, Col5a2, Col8a2, Colec12, Ddah1, Dsel, Enpp1, Ephb6, Fam129a, Fbln5, Fn1, Fndc1, Fzd1, Gli2, Golim4, Gria3, Gxylt2, Igsf10, Ism1, Itih5, Kcnj15, Klhl29, Lox, Lrrk1, Naalad2, Pamr1, Pdgfrl, Ptprf, Runx2, S1pr2, Sfrp2, Sh3d19, Shisa4, Srgap3, Srpx2, Tgfbr1, Tnfaip6, Tns3
45	18.63	0.0003	Aspn, C1qtnf6, Ccdc80, Clec11a, Clip3, Col14a1, Col1a1, Col1a2, Col27a1, Col3a1, Col5a1, Col6a1, Ctsk, Ctss, D630003M21Rik, Dse, Dsel, Eln, Fam198a, Fcgr4, Fn1, Fstl1, Gli2, Gm6548, Gpm6b, Gpr153, Gpx8, Igsf10, Il10ra, Itih5, Loxl1, Lpar1, Mfap4, Mmp2, Mxra7, Ncf1, Nid1, Pak1, Pdgfrl, Prx, Rcn3, Scpep1, Serp1, Tmem119, Tpbg, Vav1
46	18.609	0.0003	Adamts2, Aebp1, Arsb, Aspn, Ccdc80, Chpf, Ckap4, Clec11a, Col14a1, Col16a1, Col1a1, Col1a2, Col3a1, Col5a1, Col5a2, Col6a1, Col8a1, Ddah1, Dse, Dsel, Enpp1, Fam198a, Fcgr4, Fn1, Fndc1, Fstl1, Fzd1, Gli2, Gpc6, Gpr153, Gpx8, Grn, Hmha1, Itih5, Jazf1, Klhl29, Lepre1, Mrc2, Myof, Naalad2, Pdgfrl, Postn, Prkcd, Sox9, Srpx2, Tns3
47	10.303	0.0003	Adam9, Arfgap3, Arl8a, Atp9b, Boc, Col12a1, Cttn, Eya2, Fam46b, Fat1, Frzb, Gab2, Itgb5, Itgbl1, Loxl3, Loxl4, Lrp1, Ltbp2, Ltbp3, Nkd2, Nox4, Pkd2, Pon2, Ptprf, Qsox1, Reck, Scx, Sh3rf3, Shc4, Slc41a2, Srpx2, Svep1, Zfyve28
48	14.389	0.0003	Actn4, Anxa1, Anxa3, Arhgef40, Atp8b2, Axl, Bean1, Bgn, Ctdspl, Cxcl10, Dap, Ecm1, Edem1, Emp1, Fam129b, Gba, Igfbp7, Iqgap1, Layn, Meox1, Mfap5, Mob3a, Myo1d, Nckap5l, Nek6, Numbl, Pdpn, Rai14, Rtkn, Serpine2, Serpinf1, Sparc, Tgfbr2, Tgif1, Tmem45a, Tnfrsf1b
49	9.657	0.0003	Adc, Arfgap3, Arl8a, Atp9b, Cacnb1, Cttn, Cx3cl1, Dok1, Farp1, Frzb, Fxyd6, Ildr2, Itgb5, Itgbl1, Loxl3, Lrp1, Ltbp2, Ltbp3, Nkd2, Nox4, Pkd2, Pon2, Qsox1, Rasa3, Scx, Sh3rf3, Shc4, Slc1a4, Slc41a2, Svep1, Thbs3, Tspan17, Uchl1, Wbscr27, Zfyve28
50	18.918	0.0003	1110038B12Rik, 2610002J02Rik, Anxa3, App, Arhgef40, Bgn, Ctsz, Cyb5r3, Dok3, Efemp2, Emilin1, Emp1, Fam129b, Fam198a, Fstl1, Gm6548, Gpx8, Islr, Leprel2, Lhfp, Lhfpl2, Loxl1, Lrrk1, Ly86, Megf10, Mfap4, Mgp, Mmp23, Numbl, Olfml3, Pcolce, Pi4k2b, Pmepa1, Prss23, Ptgis, Sept5, Serpinf1, Sfrp1, Sox9, Thbs3, Timp2, Tkt, Tmem176a, Tmem45a, Tpbg, Tspo, Twist1, Vat1, Zbtb7c

Table 1 - Continued			
51	20.216	0.0003	Adamts2, Adcy7, Aff3, Arsb, Aspn, Col14a1, Col27a1, Col5a2, Col8a2, Colec12, Ctsk, Ctss, D630003M21Rik, Ddah1, Dse, Ephb6, Fam198a, Fbln5, Fcgr4, Fn1, Fndc1, Fzd2, Gli2, Gpx3, Gxylt2, Hexb, Kirrel, Lox, Loxl1, Lrrk1, Ltbp3, Mfap4, Mrc2, Naalad2, Nradd, P2rx7, Pak1, Pdgfrl, Plp2, Plxdc2, Prkcd, Rab31, Rasa3, Sfrp2, Shc2, Slc1a3, Srgap3, Srpx2, Tmem119, Tmem176a, Wisp1
52	20.7	0.0003	Aebp1, Ano10, Aspn, Ccdc80, Ccr2, Chpf, Clec11a, Col14a1, Col1a1, Col1a2, Col27a1, Col8a2, Colec12, Ctsk, D630003M21Rik, Dact3, Ddah1, Dse, Eln, Enpp1, Fam198a, Fbln5, Fcgr4, Fn1, Fzd2, Gpx8, Gria3, Gxylt2, Igsf10, Il10ra, Itga11, Lhfpl2, Mfap4, Mmp2, Mpeg1, Mrc2, Pak1, Pdgfrl, Ptn, Sec16b, Sfrp2, Shc2, Srgap3, Srpx2, Star, Tgfb3, Tlr2, Tmem119, Trem2, Wisp1
53	16.86	0.0003	2610002J02Rik, Anxa1, Anxa3, Arhgef40, Asns, Cd63, Creb3, Dap, Ecm1, Fam129b, Fam171a2, Gba, Igfbp7, Meox1, Mfap5, Mgp, Mob3a, Npdc1, Olfm1, Olfml3, Pdpn, Phlda3, Plk3, Pqlc3, Ptgis, Rai14, Rhoc, Rras, Rtkn, S100a11, S100a13, Serpine2, Serpinf1, Sfrp1, Sox9, Tgif1, Tmem43, Tmem45a, Tnfrsf1b, Tsku, Twf1, Vat1, Zyx
54	14.556	0.0003	Arhgap1, Bmp1, C1qtnf6, Capza1, Ccl8, Cdr2l, Col15a1, Col18a1, Col5a3, Col6a1, Col6a2, Ctss, Ctsz, Dlg4, Eln, Emilin1, Emp1, Epb4.1l2, Fbn1, Fcgr1, Fcrls, Fstl1, Gpr153, Itpripl2, Kdelr3, Lpar1, Mcm6, Oaf, Pabpc1, Pcolce, Prrg3, Sept11, Soat1, Sparc, Tril, Tyms
55	10.343	0.0003	2200002D01Rik, 2410006H16Rik, Acot9, Arhgap22, Ass1, C1qtnf7, Cx3cl1, Cyb5r3, Dok1, Eid1, F2r, Frzb, Fxyd6, Gm5424, Gmfb, Megf10, Mgp, Msn, Myo1d, Nkd2, Pcgf2, Pdk3, Pmepa1, Prss23, Ptgis, Ptprn, Rbp1, Sfrp1, Slc16a3, Slc1a4, Slc1a5, Tmem43, Unc5b, Vat1, Zbtb7c
56	11.758	0.0003	5430435G22Rik, Anxa1, Arhgef40, Atl3, Atp8b2, Axl, Bean1, Cd53, Ckb, Csf2rb, Dap, Ecm1, Edem1, Emp1, Gba, Gnai3, Inpp5d, Iqgap1, Layn, Mfap5, Mob3a, Nek6, Pde1a, Pdpn, Prrg3, Ptgis, Rai14, Rtkn, Serpine2, Tgfbr2, Tmem43, Wipf1, Zyx

KOSH Modules:

ID	Avg. degree	Module pval.	Module_gene
1	15.49	0.002	1700025G04Rik, Adam12, Axl, Azin1, Bean1, Ccl6, Ccl9, Cd63, Crlf1, Fam179a, Fkbp11, Gda, Gli1, Gm13889, Gpd1, Lman1l, Lrp8, Mt2, Prg4, Rcan1, Rnf19b, Rom1, Sbno2, Serpinb1c, Serpine1, Serpine2, Slc3a2, Sphk1, Star, Syt12, Timp1, Tnc, Tnfrsf23, Tubb2b, Uck2, Ulk3, Xpo1

Table 1 - Continued

2	15.29	0.002	2900052N01Rik, Acta1, Atf3, Capza1, Cox19, Crlf1, Dgat1, Efhd2, Fstl3, Gdf15, Gm13889, Gpd1, Hist1h1c, Itga7, Itih4, Lman1l, Lox, Lrp8, Mal, Mmp19, Pdpn, Ptgds, Rab3ip, Rcan1, Rnf19b, Rom1, Serpinb1c, Star, Syt12, Timp1, Tmem88b, Tnc, Tnfrsf23, Uck2, Ulk3
3	16.65	0.002	Acta1, Atf3, Axl, Baalc, Ccl6, Ccl9, Cd63, Cdkn1a, Col8a1, Crlf1, Efhd2, Fam179a, Gli1, Gm13889, Itga7, Itih4, Krt80, Lox, Lrp8, Met, Mt2, Myot, Pdpn, Prg4, Rcan1, Rnf19b, Serpinb1c, Serpine1, Slc3a2, Sphk1, Star, Syt12, Timp1, Tmem88b, Tnc, Uck2, Ulk3
4	16.21	0.002	1700025G04Rik, Atf3, Bean1, Ccl6, Cd63, Ddx39, Efhd2, Fkbp11, Gli1, Gm13889, Gpd1, Hand2, Itih4, Lman1l, Lrp8, Mt2, Myot, Nlrc3, Pdlim1, Pdpn, Prg4, Rab3ip, Rcan1, Rnf19b, Serpinb1c, Serpine1, Slc25a5, Slc3a2, Srl, Star, Syn2, Syt12, Timp1, Tnc, Tnfrsf23, Tubb2b, Uck2, Ulk3
5	15.26	0.002	Ahctf1, Arhgap20, Bcl9, Brpf1, Cdk19, Epb4.1, Fry, Fryl, Fzd4, Glul, Gpr116, Gpr137c, Itpr2, Klf12, Lifr, Man2a1, Mknk2, Pde1c, Per2, Plekhg3, Prex2, Rapgef5, Rasal2, Rictor, Slc4a8, Smg1, Stc1, Tfdp2, Tgfbr3, Thrb, Timp3, Tmem57, Tmtc1, Tnrc6a, Wee1
6	14.77	0.002	2900052N01Rik, Acta1, Atf3, Atp6v1d, Cd63, Cox19, Ddx39, Dgat1, Dstn, Edn3, Fstl3, Gm13889, Gpatch4, Gpd1, Hars, Hist1h1c, Lman1l, Nlrc3, Pdpn, Ptgds, Rab3ip, Rcan1, Rexo2, Rom1, Rras2, Serpinb1c, Star, Syn2, Syt12, Timp1, Tnc, Tnfrsf12a, Tnfrsf23, Uck2
7	14.32	0.002	8430408G22Rik, Agfg2, Atp10d, Cebpb, Cebpd, Chd6, Errfi1, Fry, Glul, Gpr116, Lyve1, Man2a1, Max, Mbd1, Mbnl2, Mgat4a, Mylip, Pde1c, Pdgfd, Pik3r1, Ppp1r3a, Rasal2, Rerg, Sf3b1, Slc10a6, Slc43a3, Slc4a8, Spry1, Stc1, Tgfbr3, Thrb, Tmem100, Tmem52, Tmtc1

WTTAC + KOTAC Modules:

ID	Module pval.	MCDS pval.	MCDS score	Module_genes
1	0.03	0	0.18	0610009O20Rik, Adra1a, Aldh6a1, Anks1, Arhgap26, Atl2, Atp2a2, Epb4.1, Fbxo31, Gpsm1, Ldhd, P2ry1, Pank1, Ppip5k2, Ppp1r3a, Rbm38, Scn4b, Tmem150c, Wnk2
2	0.02	0	0.20	1110018G07Rik, Agtr1a, Ankrd32, Cnst, D2hgdh, Entpd5, Fgf13, Fign, Gpr22, Hlf, Kcnh2, Kcnj5, Kcnn2, Nr3c2, Ppat, Sobp, Zadh2
3	0.03	0	0.22	1810026J23Rik, 2310067B10Rik, Adra1b, Asb15, Cadm4, Cpn2, Cyb5r2, Entpd5, Fign, Gal3st3, Gm11992, Gpr22, Kcnh2, Kcnn2, Ky, Lrrc15, Pde4a, Slc16a10, Slc22a3, Slc25a42, Wnt5a

Table 1 - Continued

4	0.03	0	0.21	2310067B10Rik, Abcb7, Adra1b, Agpat3, Cadm4, Camk2a, Clcn3, Cldn12, Entpd5, Gal3st3, Gpr155, Iqsec1, Isoc1, Kcnh2, Kcnj3, Kcnn2, Lgr6, Lpcat3, Lrrc15, Pde4a, Sbk2, Slc16a10, Slc25a42, Stat5a
5	0.04	0	0.14	1700040L02Rik, Acacb, Acadm, Acads, Acat1, Adhfe1, Asb14, Atp5a1, Auh, Dcaf11, Etfa, Hadha, Idh3b, Isoc2a, Lrrc39, Mccc2, Mipep, Mlycd, Mylk4, Phyh, Pln, Ric8b, Sdhc, Slc25a11, Sucla2, Suclg2
6	0.03	0	0.23	Abcd3, Acad11, Car14, Cpn2, Dhrs11, Entpd5, Fign, Gal3st3, Gpr22, Grhl2, Idh3a, Kcnh2, Kcnj3, Klhl33, Lamb3, Lrrc15, Pdp2, Rbfox1, Sbk2, Slc16a10, Slc25a42, Slc2a4, Wnk2
7	0.03	0	0.21	2310067B10Rik, Acsl1, Adra1b, Cpn2, Crhr2, Entpd5, Gal3st3, Gm11992, Gpr22, Grhl2, Hadh, Kcnn2, Ky, Lrrc15, Magt1, Sbk2, Slc16a10, Slc22a3, Slc25a42, Smarcd1
8	0.03	0	0.18	1700040L02Rik, Adhfe1, Asb14, Atp5a1, Ccbl2, Coq2, Efnb3, Etfa, Fdft1, Fh1, Hadhb, Hopx, Hrc, Isoc2a, Lrrc39, Mccc2, Mdh1, Mipep, Mlycd, Phyh, Sdhc, Slc25a29, Sord, Ube2b
9	0.03	0.023001	0.12	Abca12, Acad10, Acad12, Agl, Aldh4a1, Atp8a1, Calcoco1, Crbn, Dhdh, Dnajb9, Fgf13, Klhdc1, Nampt, Trap1
10	0.03	0	0.26	2310010M20Rik, 3110002H16Rik, 4632428C04Rik, 9030617O03Rik, Aldh5a1, Ccbl1, Cngb3, Cyb5rl, Dhrs7c, Ehhadh, Fkbp4, Gabrr2, Hrc, Klhl33, Nqo2, Plxnb1, Qsox2, Retsat, Slc36a2, Sord, Ung, Whrn, Zfp612
11	0.03	0	0.17	Adcy6, Adhfe1, Adra1a, Aldh6a1, Arhgap26, Atl2, Epb4.1, Fbxo31, Gm14420, Ldhd, Mylk4, Pank1, Pcnt, Ppara, Ppip5k2, Ppp1r3a, Prkce, Rasl10b, Rbm38, Thrb, Ubr2, Wnk2
12	0.03	0	0.22	Agtr1a, Ankrd32, Arhgef9, Camk2a, Crhr2, Entpd5, Fign, Gca, Gpr22, Grhl2, Hlf, Kcnh2, Kcnj3, Kcnn2, Lamb3, Lrrc15, Nr3c2, Pln, Prlr, Rbfox1, Slc16a10, Slc25a42
13	0.03	0	0.24	Adamts6, Adamtsl2, App, Boc, Dkk3, Eya2, Frzb, Ildr2, Itgbl1, Lrp1, Lrrc16a, Nox4, Pkd2, Shc4, Slit3, Svep1, Tgfbr1, Tmem119
14	0.02	0	0.23	Agtr1a, Arhgap5, BC037032, Cpeb3, Dsg2, Entpd5, Fam174b, Fgf13, Fign, Gpr22, Kcnh2, Nmnat3, Nr3c2, Pm20d1, Ppat, Ppm1k, Slc25a42, Tnni3k
15	0.04	0	0.25	2200002D01Rik, Acot9, Akr1b8, Arfgap3, B4galt5, Bgn, Cttn, Cx3cl1, Frzb, Fxyd6, Gm5424, Ltbp2, Msn, Mtpn, Nkd2, Pon2, Scx, Slc16a3, Slc1a4, Tmbim1, Tspan17, Unc5b

Table 1 - Continued

16	0.02	0.000384375	0.15	Acadm, Aco2, Acss1, BC037032, Ccdc141, Cpeb3, Dsg2, Fam174b, Gbas, Got1, Nnt, Ppm1k, Ppp1r3b, Sdha, Slc27a1, Tmem179
17	0.04	0	0.22	Car14, Ccdc47, Cox10, Cpeb3, Dbt, Entpd5, Fign, Gal3st3, Gpr22, Idh3a, Kcnh2, Kcnj3, Kcnn2, Lrrc15, Nmnat3, Ppm1k, Rmnd1, Sbk2, Slc16a10, Slc25a42, Tnni3k
18	0.03	0	0.23	2310067B10Rik, 9030617O03Rik, Acot3, AI464131, Aldh5a1, Ccdc90a, Cpn2, Entpd5, Gal3st3, Grhl2, Hars2, Klhl33, Lamb3, Lingo3, Lrrc15, Osbp2, Pla2g5, Scn4a, Serac1, Slc16a10, Slc25a42, Slc2a4, Tmem25
19	0.03	0.029227723	0.13	Acad10, Acad12, Adra1a, Aldh4a1, Aldh6a1, Atp8a1, Calcoco1, Cdnf, Clpb, Kcna7, Ldhd, Mccc1, Nampt, Pank1, Pdk2, Pfkm
20	0.04	0	0.27	Adam12, Arhgap23, Cdr2l, Col15a1, Col4a1, Col4a2, Dclk1, Fbn1, Lamc1, Loxl2, Mrc2, Ptgfrn, Rab15, Ubash3b
21	0.03	0	0.38	Atp8b1, Capg, Cd44, Ch25h, Csf2rb2, Kctd11, Panx1, Snai1, Tgfb1, Timp1, Ttc9, Tubb2b
22	0.02	0.059208824	0.12	Adra1b, Agpat3, Aig1, Arhgef19, Clcn3, Cpn2, Gal3st3, Klhl30, Nhsl1, Sbk2, Stat5a, Tsc22d4
23	0.02	0.123354808	0.11	Adcy6, Adra1a, Arhgap26, Chd6, Cobll1, Epb4.1, Fbxo31, Herpud1, Klf15, Pik3r1, Ppip5k2, Ppp1r3a, Sesn1, Spsb1, Tgfbr3, Thrb, Tnrc6c, Wipf3
24	0.02	0.424292523	0.11	Ccdc127, Cnst, Cyfip2, D2hgdh, Gab1, Ghr, Kcnj5, Osbpl2, Pex26, Prr12, Rhobtb2, Rnf150, Sh3rf2, Sobp
25	0.03	0	0.19	2310067B10Rik, 9030617O03Rik, Atp1a2, Bmp7, Celsr2, Dars2, Dus4l, Gpr155, Gpt2, Idh1, Impa2, Ipo13, Klhl23, Mme, Nqo2, Oma1, Osgepl1, Plxnb3, Reep1, Retsat, Sgol2, Slc2a12, Tmem135, Unc45b
26	0.03	0	0.15	1700040L02Rik, Acadvl, Adhfe1, As3mt, Asb14, Atp5a1, Dbt, Dhrs11, Efnb3, Etfa, Hadhb, Hopx, Hrc, Lrrc39, Mccc2, Mdh1, Me3, Mitf, Mlycd, Oxa1l, Phyh, Prkag1, Sdhc, Sirt3, Slc25a29
27	0.01	0	0.16	0610009O20Rik, 1700040L02Rik, Abcd3, Acads, Adhfe1, Asb14, Atp2a2, Auh, Coq9, Crat, Dbt, Dennd4b, Etfa, Etfdh, Hadh, Hadha, Hibadh, Hspa9, Idh3b, Ldhd, Mipep, Mlycd, Mmab, Mylk4, Pdk2, Phyh, Pln, Prdx3, Pxmp4, Ric8b, Sdhc, Slc25a20, Sucla2, Suclg2, Svip, Tmem143, Uqcc
28	0.03	0	0.29	1500009L16Rik, Ace, Arfgap3, Aspscr1, Atp6v1h, Chpf2, Dlgap4, Dok1, Ecm1, F2r, Frzb, Fxyd6, Gipc2, Mical1, Nppa, Nupr1, Panx1, Plekha4, Slc39a7, Tmem43, Trim47, Tspan17
29	0.02	0	0.28	1500009L16Rik, Aspscr1, Atp6v1h, Csf2rb2, Ctdspl, Dlgap4, Eif1a, Endod1, Fam105b, Gipc2, Mical1, Nupr1, Panx1, Plekha4, Slc39a7, St6galnac4, Trim47, Ttc9

Table 1 - Continued

30	0.03	0	0.20	2310067B10Rik, 3110002H16Rik, 9030617O03Rik, Atp1a2, Bmp7, Ctf1, Dcun1d2, Dus4l, Fastkd1, Idh1, Impa2, Klhl23, Mme, Mrgprh, Nqo2, Pkd2l2, Plbd1, Plxnb3, Reep1, Retsat, Slc2a12, Slc36a2, Tmem135, Zfp667
31	0.03	0	0.24	Arfgap3, Col12a1, Dkk3, Eya2, Fbln5, Frzb, Ildr2, Itgb5, Itgbl1, Loxl3, Loxl4, Lrp1, Ltbp2, Nkd2, Pkd2, Pon2, Sh3rf3, Shc4, Ski, Skil, Slc41a2, Svep1, Tmbim1
32	0.02	0	0.14	Adra1b, Arhgef19, Cpn2, Fbxo21, Foxo6, Gal3st3, Glrx, Gm11992, Gm6086, Gpr155, Jarid2, Klhl30, Ky, Nhsl1, Sbk2, Stat5a, Tet1, Tsc22d4
33	0.02	0	0.28	Abhd15, Ace, Adamtsl2, Fgfr1, Frzb, Fzd7, Il17ra, Lrp1, Lrrc16a, Ltbp2, Mst1r, Pkd1, Sgms2, Ski, Svep1
34	0.02	0.917860526	0.10	2310042D19Rik, Acadm, Acat1, Bckdk, D10Jhu81e, Dcaf11, Egln1, Esrra, Fam174b, Fitm2, Gbas, Nnt, Pccb, Peo1, Pnpla2, Tmem66
35	0.04	0	0.21	2310067B10Rik, 3110002H16Rik, 9030617O03Rik, Bmp7, Cadm4, Fbxo21, Gal3st3, Gpr155, Gpr157, Grb14, Hibadh, Idh1, Impa2, Klhl33, Ky, Lrrc15, Mme, Nqo2, Pdp2, Plxnb3, Rps6ka5, Slc22a3, Slc2a12, Tmem135, Tmem25, Ttc38
36	0.02	0.000507216	0.13	5830417I10Rik, Adra1b, Agpat3, Arhgef19, Cpn2, Gal3st3, Gon4l, Gtf3c1, Iqsec1, Klhl30, Lgr6, Lpcat3, Nhsl1, Parp1, Pde4a, Sbk2, Stat5a, Tsc22d4
37	0.04	0	0.23	2310010M20Rik, 2410127L17Rik, 4632428C04Rik, Aldh5a1, Ccbl1, Ccbl2, Cngb3, Cpt2, Decr1, Dhrs7c, Ehhadh, Fbp2, Fdft1, Fkbp4, Fndc5, Gabrr2, Hadh, Hopx, Hrc, Mrps31, Msrb2, Sord, Tmem70, Ung, Whrn, Zfp612
38	0.03	0	0.20	Adra1b, Alas1, Arhgef9, Car14, Cpn2, Crhr2, Entpd5, Fign, Gal3st3, Gpr22, Grhl2, Hccs, Idh3a, Kcnh2, Kcnn2, Lrrc15, Nmnat3, Ppat, Sbk2, Slc16a10, Slc25a42, Uqcc
39	0.03	0.002510204	0.13	Acadl, Acadm, Acat1, Adhfe1, Asb10, Atp5a1, C030006K11Rik, D10Jhu81e, Dcaf11, Dhrs4, Etfa, Fundc2, Hadha, Idh3b, Immt, Lrrc39, Me3, Nnt, Pccb, Pln, Pnpla2, Ric8b, Sdhc, Slc25a11, Suclg2
40	0.02	0	0.23	3110002H16Rik, 9030617O03Rik, Atp1a2, Dis3l, Emilin2, Idh1, Impa2, Ipo13, Mme, Nqo2, Pkd2l2, Plbd1, Plxnb1, Plxnb3, Qsox2, Retsat, Slc2a12, Slc36a2, Wdr24
41	0.03	0	0.14	Cadm4, Cldn12, Ezr, Gal3st3, Gm11992, Gpr155, Grb14, Hspa12a, Isoc1, Kcnn2, Klhl30, Ky, Lmod3, Mfap3l, Osbpl6, Rbbp5, Slc22a3, Sox12, Stat5a, Tsc22d4, Ube2d1, Wnt5a, Zbtb20
42	0.02	0.863745536	0.10	Acadsb, Acsl6, Calcoco1, Cdc42bpg, Crebl2, Dcaf8, Gbas, Ivd, Ndrg2, Optn, Rab12, Tmem179, Txlnb, Wfikkn2

Table 1 - Continued

43	0.03	0	0.29	Adamts4, Atp8b1, Capg, Csf2rb2, Fxyd5, Gas2l1, Il4ra, Kctd11, Litaf, Timp1, Ttc9, Tubb2b
44	0.02	0	0.23	Cpn2, Crhr2, Dhrs11, Entpd5, Epha4, Fign, Gal3st3, Gm11992, Gpr22, Grhl2, Lamb3, Lrrc14b, Lrrc15, Magt1, Slc16a10, Slc16a7, Slc22a3, Smarcd1
45	0.02	0	0.24	Adra1b, Cadm4, Cpn2, Crhr2, Entpd5, Fgf16, Fign, Gal3st3, Gm11992, Gpr22, Grhl2, Kcnh2, Lrrc15, Magt1, Nhsl1, Sbk2, Slc16a10
46	0.02	0	0.24	Ctgf, Dennd4a, Dkk3, Frzb, Fzd7, Gab2, Leprel1, Lrp1, Ltbp2, Ltbp3, Nox4, Pkd1, Pkd2, Sh3rf3, Ski, Svep1, Tgfb2, Tmbim1
47	0.03	0	0.15	Adra1b, Agpat3, Aig1, Arhgef19, Cadm4, Gal3st3, Gm11992, Gpr155, Grb14, Kcnn2, Klhl30, Ky, Lrrc15, Lynx1, Pde4a, Rtn4ip1, Slc16a10, Slc22a3, Stat5a, Tmem25, Tsc22d4, Ttc30a1, Ttc30b, Wnt5a
48	0.03	0	0.22	Agtr1a, Car14, Cox10, Cpeb3, Cpn2, Crhr2, Dhrs11, Entpd5, Fign, Gm11992, Got2, Gpr22, Grhl2, Kcnh2, Kcnk3, Lamb3, Lrrc15, Mfn1, Nmnat3, Ppm1k, Sbk2, Slc16a10, Slc25a42
49	0.03	0	0.21	Adra1b, Ankrd32, Arhgef9, Cyb5r2, Entpd5, Fign, Gca, Gpr22, Idh3a, Kcnh2, Kcnj3, Kcnn2, Lrrc15, Lynx1, Msi2, Nr3c2, Pde4a, Ppat, Sbk2, Slc16a10, Slc25a42, Stard7
50	0.04	1	0.07	Ablim1, Adcy6, Chd7, D830031N03Rik, Dusp18, Epb4.1, Fry, Glt28d2, Gm14420, Gpcpd1, Klf15, Klf9, March6, Mbnl2, Mll3, Mlxip, Ncoa2, Pcnt, Plin4, Ppara, Ppp1r3a, Snrnp200, Spen, Tmem170b, Trim56, Ubr2
51	0.02	0	0.26	C1qtnf7, Camkk1, Cd302, Cygb, Kctd15, Ltbp4, Npc2, Nupr1, Panx1, Pi16, Pmp22, Pqlc3, Rab23, Serping1, Sod3
52	0.01	0	0.21	0610009O20Rik, 1700040L02Rik, Abcc9, Acads, Adhfe1, Adra1a, As3mt, Asb14, Atp5a1, Auh, Camk2a, Dhrs4, Efnb3, Etfa, Hadh, Hopx, Hrc, Idh3b, Ldhd, Lrrc39, Mdh1, Mipep, Mylk4, Nsmaf, P2ry1, Pdk2, Phyh, Pln, Prdx3, Rbfox1, Ric8b, Sdhc, Slc22a5, Sord, Sucla2, Suclg2, Svip
53	0.04	0.44701 3889	0.10	Amigo1, Capn7, Ctage5, D2hgdh, Fbxl5, Flcn, Intu, Klhdc1, Lmtk2, Mospd1, Osbp2, Pkdrej, Pnpla7, Rfc1, Suox
54	0.03	0	0.17	Adra1b, Cnga3, Cyb5r2, Entpd5, Gal3st3, Grb14, Kcnn2, Klhl30, Ky, Lrrc15, Pde4a, Pkdrej, Pkia, Raf1, Slc16a10, Slc22a3, Stat5a, Tsc22d4, Ttc30b, Wnt5a
55	0.02	0.64422 6606	0.11	2900097C17Rik, Acad12, Atp8a1, Calcoco1, Cdnf, Crbn, Dhdh, Dnajb9, Drosha, Klhdc1, Nampt, Pnpla8, Rai2, Trap1

Table 1 - Continued

56	0.03	0.07917 3786	0.12	Adcy6, Arhgap26, Dusp18, Epb4.1, Fbxo31, Fry, Gm14420, Gm8898, Klf15, Klf9, Pcnt, Plin4, Ppara, Ppip5k2, Ppm1l, Ppp1r3a, Snrnp200, Tfdp2, Trim56, Ubr2, Wnk2
57	0.03	0	0.18	3110002H16Rik, 9030617O03Rik, Atp1a2, Celsr2, Gatc, Gpt2, Idh1, Impa2, Klhl23, Lman2l, Mme, Mrgprh, Pkd2l2, Plxnb3, Pnpt1, Ppp2r3a, Reep1, Retsat, Sgol2, Slc2a12, Slc36a2, Tmem135, Unc45b
58	0.01	0	0.21	0610009O20Rik, 3110057O12Rik, Abcc9, Adhfe1, Adra1a, Aldh4a1, Aldh6a1, Asb15, Atl2, Atp2a2, Bcl7a, Camk2a, Ccdc47, Coq9, Dbt, Dnajc28, Echs1, Hadha, L2hgdh, Ldhd, Mccc1, Mitf, Mylk4, Nampt, P2ry1, Pank1, Pdk2, Pfkm, Pln, Ppm1k, Ptpn3, Rbfox1, Ric8b, Sdha, Thrb, Tmem65
59	0.03	0.00013 0851	0.14	5830417I10Rik, Adra1b, Gal3st3, Gon4l, Iqsec1, Kcnd2, Lgr6, Lpcat3, Med12l, Ncoa1, Pde4a, Ralgapa2, Scn4a, Stat5a, Tsc22d4, Ttc30b
60	0.03	0	0.15	Agtr1a, BC037032, Cacnb2, Cnst, Epm2aip1, Fgf13, Gab1, Gpr22, Kcnj5, Nr3c2, Pm20d1, Ppat, Ralgps2, Rbm20, Sobp, Zadh2
61	0.02	0	0.25	1500009L16Rik, Arfgap1, Arfgap3, Aspscr1, Fxyd6, Gipc2, Gm5424, Mical1, Nppa, Nupr1, Ormdl3, Ptprn, Rbp1, Slc39a7, Tspan17
62	0.03	0	0.29	Camkk1, Capg, Carhsp1, Hn1, Kctd15, Lmna, Ltbp4, Mapre1, Mical1, Nupr1, Panx1, Pmp22, Rab23, Rhoc, Serping1, Slc27a3, Sod3, Tmem43
63	0.03	0	0.25	2310067B10Rik, 9030617O03Rik, Acot1, Adra1b, Aldh5a1, Cpn2, Crhr2, Gal3st3, Grhl2, Impa2, Lingo3, Lrrc15, Magt1, Mme, Mrgprh, Qrsl1, Rhot2, Sbk2, Slc16a10, Slc16a7, Slc22a3, Slc5a6, Smarcd1, Whrn
64	0.01	0	0.16	0610009O20Rik, 1700040L02Rik, Acads, Adck1, Adhfe1, AI118078, Aldh5a1, As3mt, Asb14, Auh, Chac2, Chchd3, Crat, D3Ertd751e, Dcun1d2, Dnajc28, Etfa, Galm, Hadh, Hopx, Idh3b, Lrrc14b, Mipep, Osgepl1, P2ry1, Pex7, Phyh, Pln, Prdx3, Ric8b, Rmnd1, Sdhc, Slc22a5, Slc25a26, Sucla2, Suclg2, Svip, Tmem143
65	0.02	0	0.29	Acot9, B4galt5, Cilp, Cttn, Mvp, Pon2, Rtn4, Serpinb1c, Snx10, Ttll12, Ube2z
66	0.03	0	0.30	Adam9, B4galt5, Col12a1, Dpysl3, Loxl3, Ltbp2, Pon2, Prmt2, Qsox1, Runx1, Serpinb1c, Shc4, Shisa3, Slc10a3, Slc1a4, Spcs3, Tgfb2, Thbs4, Tubb3
67	0.03	0	0.23	2410006H16Rik, 5730409E04Rik, Arfgap3, Atp6v1h, Boc, Cx3cl1, Ddx50, Dok1, Eid1, Frzb, Fxyd6, Mgp, Mical1, Nkd2, Nox4, Nupr1, Ormdl3, Pkd2, Pmepa1, Ptprn, Scarf2, Scx, Tmbim1, Tspan17, Wbscr27

Table 1 - Continued				
68	0.03	0	0.24	2200002D01Rik, Acot9, Arfgap3, Atp6v1h, Chpf2, Cttn, Cx3cl1, Dlgap4, Frzb, Fxyd6, Ltbp2, Mical1, Nkd2, Pon2, Sfrp1, Slc16a3, Slc1a4, Slc30a4, Tmbim1, Trim47, Tspan17, Wbscr27
69	0.03	0.394528302	0.11	1700040L02Rik, Acadl, Acadvl, Atp5a1, C030006K11Rik, Cpt1b, Etfa, Gstm7, Hadhb, Idh3b, Lrrc39, Mccc2, Mdh1, Mlycd, Myadml2, Oxa1l, Prpf19, Ptcd2, Sdhc, Sirt3, Slc25a11, Slc25a20, Sod2
70	0.03	0	0.18	Adra1b, Alas1, Arhgef19, Arhgef9, Cpn2, Crhr2, Entpd5, Fign, Foxo6, Gal3st3, Gm11992, Grhl2, Jarid2, Kcnn2, Ky, Lrrc15, Nhsl1, Sbk2, Slc16a10, Slc22a3, Smarcd1, Stard7
71	0.03	1	0.08	Acss1, Adcy6, Aldh6a1, Arhgap26, Dusp18, Gm14420, Kif1c, Klf15, Pcnt, Plin4, Ppara, Ppp1r3a, Tbc1d16, Tfdp2, Trim56, Ubr2
72	0.03	0	0.30	1500009L16Rik, 2200002D01Rik, Arfgap3, Aspscr1, Atp6v1h, Bcr, Crlf1, Cx3cl1, Frzb, Fxyd6, Gipc2, Gm5424, Mical1, Nkd2, Nupr1, Ormdl3, Plekha4, Ptprn, Slc16a3, Slc1a4, Thbs4, Tmem45a, Trim47, Tspan17, Unc5b
73	0.02	0	0.15	Cadm4, Dnmt3a, F3, Gal3st3, Gm11992, Gpr155, Grb14, Hcn4, Hspa12a, Kcnn2, Ky, Lmod3, Nbn, Slc22a3, Trmt2b, Ube2d1, Wnt5a

WTTAC + KOSH Modules

ID	Module pval.	MCDS pval.	MCDS score	Module_genes
1	0.05	0	0.19	2310010M20Rik, 3110002H16Rik, Ces1d, Cmbl, Dhrs7c, Echdc3, Fdft1, Npepl1, Peli2, Plk5, Zfp612
2	0.03	0	0.22	Actr3b, Adi1, Ctage5, Dcaf12l1, Dirc2, Ghr, Gpr155, Intu, Pnpla7, Rps6ka5, Ttc38
3	0.05	0	0.19	Agtpbp1, Aldh6a1, Arhgap20, Arhgap26, Dbt, Fbxo31, Gm16119, Nampt, Pde7a, Ppip5k2, Ppm1k, Rps6ka5, Spsb1, Thrb
4	0.03	1	0.06	Adam22, Ahctf1, Arhgap20, Ccrn4l, Chd6, Epb4.1, Fry, Fzd4, Pcnt, Pik3r1, Plin4, Smg1, Tfdp2
5	0.03	0	0.14	1110034G24Rik, Arhgap26, Ccrl2, Fbxo31, Gm16119, Herpud1, Mylk4, Pde7a, Ppip5k2, Slc25a33, Stard13, Tgfbr3
6	0.03	0.000130851	0.13	2010111I01Rik, 6430548M08Rik, Aars, Clic5, Diap1, Dync1li1, Fam131c, Flnc, Hspa1l, Nuak1, Rrp12, Xirp1, Xirp2
7	0.03	0	0.29	Cd44, Fbln2, Fgl2, Itga5, Itgb3, Lrp8, Msn, Osbpl9, Tmem88b, Tnfrsf23, Uck2

Table 1 - Continued				
8	0.03	0.907043363	0.08	Adam22, Ahctf1, Arhgap20, Epb4.1, Fbxo31, Gm14420, Herc1, Pdpr, Tfdp2, Thrb, Tnrc6c, Ttn, Ubr2
9	0.05	0	0.25	2010111I01Rik, Akap2, Azin1, Enah, Fgl2, Fhl1, Fstl3, Itga5, Lrp8, Pfkp, Rhod, Tmem88b, Tnfrsf23, Uck2, Ulk3
10	0.03	0	0.29	Fbln2, Fgl2, Galns, Gdf6, Lrp8, Rnd1, Rnf19b, Rtn4, Star, Tmem88b, Tnfrsf23, Uck2, Ulk3
11	0.05	0	0.24	Asb14, Auh, D3Ertd751e, Gstz1, Hopx, Lrrc39, Mipep, Msrb2, Pdf, Pln, Svip
12	0.05	0	0.16	2010111I01Rik, Aars, Ankrd23, Clic5, Ddx21, Fgl2, Flnc, Hspa1l, Nuak1, Pdlim5, Tmem88b, Tomm70a, Uck2, Usp16, Xirp2
13	0.03	1	0.05	Acot11, Adamtsl4, Ahctf1, Arhgap20, Arhgap26, Bcl9, Ccrn4l, Fbxo31, Foxo3, Fry, Mbd1, Per2, Pik3r1, Plin4, Ppp1r3a, Rasal2, Rhobtb1, Stard13, Tef, Tfdp2, Ubr2, Vps13a, Wee1, Zbtb16

KOTAC + KOSH Modules:

ID	Module pval.	MCDS pval.	MCDS score	Module_genes
1	0.03	0	0.25	1700025G04Rik, Aars, Ankrd23, Ehbp1l1, Hspa1l, Lrp8, Nlrc3, Nuak1, Rcan1, Slc38a2, Stat3, Uck2
2	0.03	0	0.32	Akap2, Crlf1, Fstl3, Gdf15, Hbegf, Ifrd1, Itga5, Lman1l, Picalm, Serpina3n, Serpine1, Synpo2l, Zfp697
3	0.04	0	0.38	Adamts4, Ap3s1, Atp8b1, Cd44, Dap, Kctd11, Lrrc59, Mfap5, Scml4, Timp1, Tubb2b
4	0.03	0	0.26	Acad8, As3mt, Asb14, Auh, Decr1, Fh1, Hadh, Lrrc39, Mipep, Msrb2, Sord, Suclg2, Svip
5	0.03	0	0.22	Akap2, Azin1, Enah, Fhl1, Nlrc3, Osbpl9, Parp3, Prnp, Rras2, Shroom3, Tnfrsf23, Uck2, Ulk3
6	0.03	0	0.40	Adamts12, Atp10a, Col15a1, Col4a1, Col4a2, Col5a1, Col6a2, Dclk1, Fbn1, Loxl2, Ptgfrn
7	0.04	0	0.22	2010111I01Rik, Aars, Clic5, Dync1li1, Enah, Flnc, Hspa1l, Nlrc3, Pkn1, Tmem62, Tnfrsf23, Ubxn4, Uck2
8	0.03	0	0.31	Arfgap3, Crlf1, Efhd2, Fstl3, Ifrd1, Itga5, Krt80, Lman1l, Pbxip1, Serpina3n, Zfp697
9	0.03	0	0.27	Akap2, Anxa7, Atp8a2, Enah, Nlrc3, Parp3, Pfkp, Prnp, Rras2, Shroom3, Tgfb2, Ulk3
10	0.03	0	0.31	3110002H16Rik, 9030617O03Rik, Atp1a2, Celsr2, Ces1d, Gpr155, Gpt2, Mme, Plxnb3, Retsat, Slc40a1, Zfp612

Table 1 - Continued				
11	0.03	1	0.04	Ash1l, D830031N03Rik, Dgkh, Ep400, Fryl, Herc1, Mll2, Ncor1, Smg1, Snrnp200, Ttn
12	0.04	0	0.20	Ankrd23, Ctps, Dusp27, Grk5, Hk1, Mical2, Pfkp, Prkar1a, Rasl11b, Rras2, Shb, Shroom3, Spsb4, Synpo2l, Tgfb2
13	0.03	0	0.22	2310010M20Rik, 3110002H16Rik, Celsr2, Gpt2, Ipo13, Klf12, Klhl33, Lifr, Npc1, Spata2l, Zfp612
14	0.04	0	0.29	Anxa1, Atp8b1, Cd44, Csf2rb2, Elovl1, Fbln2, Il4ra, Kctd11, Lilrb4, Osmr, Sphk1, Timp1
15	0.03	0	0.40	Actn1, Atp10a, Col15a1, Col4a1, Col4a2, Col6a1, Dclk1, Fbn1, Lepre1, Loxl2, Lpar1, Myof, Ptgfrn
16	0.03	1	0.02	Aff1, Ahctf1, Fry, Fzd4, Hipk1, Lancl3, Lrrc8a, Nipbl, Pdgfd, Plxna2, Shroom4, Tmcc3
17	0.03	0	0.24	Anxa7, Cttn, Dpysl3, Fhl1, Loxl4, Myo1c, Pfkp, Prmt2, Rab35, Rtn4, Slc10a3, Tgfb2, Ulk3
18	0.03	0	0.28	Adamts4, Atp8b1, Cd44, Csf2rb2, Kctd11, Lrrc59, Osmr, S100a4, Scml4, Snai1, Timp1, Tubb2b, Vcan
19	0.04	0	0.26	Adhfe1, As3mt, Asb14, Atp5a1, Echs1, Fh1, Lrrc39, Mccc2, Me3, Ppm1k, Sdhc, Sord, Suclg2, Svip, Tnni3k
20	0.03	0	0.28	2010111I01Rik, 2310057M21Rik, Clic5, Gdf15, Hbegf, Hspa1l, Nuak1, Rcan1, Serpine1, Syn2, Synpo2l, Uck2
21	0.04	0.272591429	0.07	Epb4.1, Fbxo31, Fry, Glt28d2, Glul, Gm14420, Gm8898, Plin3, Ppp1r3a, Tmem182, Tmtc1, Ubr2

WTTAC + KOTAC + KOSH Modules:

ID	Module pval.	MCDS pval.	MCDS score	Module_genes
1	0.03	0.004472727	0.15	Acacb, Arhgap20, Arhgap26, Cobll1, Epb4.1, Fbxo31, Fyco1, Glt28d2, Pcnt, Ppip5k2, Tnik, Ubr2
2	0.03	0	0.14	Agtpbp1, Aldh6a1, Arhgap20, Arhgap26, Epb4.1, Fbxo31, Glt28d2, Gm14420, Klf9, Nampt, Pank1, Ppip5k2, Ppp1r3a, Rmnd5a, Tfdp2, Thrb, Tnik, Ubr2
3	0.03	0.669231818	0.10	Arhgap20, Arhgap26, Epb4.1, Fbxo31, Fry, Glt28d2, Pik3r1, Ppip5k2, Spsb1, Tgfbr3, Zbtb16
4	0.03	0	0.21	Adhfe1, As3mt, Auh, Clpx, Echs1, Lrrc39, Mccc1, Pln, Ppm1k, Sdhc, Suclg2, Svip
5	0.03	0.000384375	0.15	Adra1a, Aldh6a1, Arhgap26, Cobll1, Epb4.1, Fbxo31, Glt28d2, Gpcpd1, Ppip5k2, Spsb1, Tfdp2, Thrb, Tnrc6c
6	0.04	0	0.28	1500015O10Rik, Col8a2, Colec12, Dkk3, Fibin, Fmod, Ism1, Itgbl1, Olfml1, Pamr1, Sfrp2

Table 1 - Continued				
7	0.03	0.809140541	0.10	Cd74, Ciita, Ctss, Fcgr4, H2-Aa, H2-Ab1, H2-DMa, H2-Eb1, Itgal, Itgb7, Mpeg1
8	0.03	0	0.16	2010111I01Rik, Clic5, Diap1, Enah, Fhl1, Flnc, Lrp8, Mical2, Napepld, Osbpl9, Pdlim5, Stat3, Tnfrsf23, Uck2, Ulk3, Wsb2
9	0.03	0	0.22	3110002H16Rik, 9030617O03Rik, Actr3b, Atp1a2, Ctage5, Fbxo21, Gpr155, Gpt2, Ipo13, Mme, Rps6ka5, Ttc38
10	0.03	0	0.35	1500015O10Rik, Abi3bp, Adcy7, Cilp, Col8a2, Colec12, Dkk3, Fbln5, Fibin, Fmod, Fzd2, Gria3, Ism1, Itgbl1, Ltbp3, Mfap4, Pamr1, Pdgfrl, Ptn, Sfrp2
11	0.04	1	0.03	Ash1l, Birc6, Cep350, Dopey1, Ep300, Heg1, Huwe1, Lnpep, Med13, Med13l, Mga, Mib1, Mon2, Mycbp2, Nipbl, Nr2c2, Phc3, Prrc2c, Scaf11, Vps13b, Wdfy3, Xrn1, Zbed6
12	0.04	0	0.20	Acadm, As3mt, Atp5a1, Auh, Echs1, Lrrc39, Ndufs1, Ric8b, Suclg2, Svip, Tnni3k
13	0.03	0	0.22	1500017E21Rik, 1700025G04Rik, 2310057M21Rik, Ankrd23, Clic5, Hspa1l, Kif5b, Krt80, Nlrc3, Nuak1, Synpo2l
14	0.03	1	0.03	Birc6, Crebbp, Ep300, Hivep2, Itsn2, Klf3, Med13, Mga, Mon2, Nipbl, Prpf8, Prrc2c, Scaf11, Vps13b, Wdfy3
15	0.03	0	0.24	2010111I01Rik, 2310057M21Rik, Akap2, Clic5, Enah, Fhl1, Hbegf, Mical2, Nlrc3, Nuak1, Pfkp, Rras2, Shroom3, Slc38a2, Synpo2l, Tnfrsf23, Ulk3
16	0.03	1	0.09	Aff1, Arhgap26, Cobll1, Foxo3, Fry, Glt28d2, Pcnt, Ppara, Spen, Thrb, Tnik, Tnrc6c, Ubr2, Vprbp

CHAPTER 3: A NOVEL METHOD FOR CONSTRUCTING CONDITION-SPECIFIC TRNS AND ITS APPLICATION TO THE DEVELOPMENT OF HEMATOPOIETIC STEM CELLS

Abstract

Hematopoietic stem cells (HSCs) in the embryo are derived from hemogenic endothelium (HE) of the arterial wall from the aorta-gonad-mesonephros (AGM) region and yolk sac (YS). HE from AGM and YS has different developmental potentials. HE from YS primarily produces committed erythroid/myeloid progenitor and HE from AGM can produce lymphoid progenitors and HSCs. The transcriptional regulatory networks (TRNs) that control the endothelial-to-hematopoietic transition in AGM and YS are poorly understood. Here we compared the transcriptomes of endothelium and hemogenic endothelium from embryonic (E) day 9.5 and E10.5 AGM and YS by RNA-Seq. We developed a novel computational method for constructing condition-specific TRNs with limited number of samples by sample elimination and network comparison. By modeling developmental-stage-specific TRNs, we identified a number of transcription factors that regulate the endothelial-to-hematopoietic transition, including Runx1, Spi1, Gfi1 that are known as endothelial cell reprogramming factors. In addition, we also found TEAD factors and HOX family genes are important transcription factors that regulate the endothelial-to-hematopoietic transition.

3.1 Introduction

During normal embryonic development, hematopoietic progenitors (HPs) and HSCs differentiate from a small population of endothelial cells referred to as hemogenic endothelium (Bertrand et al., 2010; Boisset et al., 2010; Eilken et al., 2009; Kissa and Herbomel, 2010). Hemogenic endothelial cells, which begin as flat cells in a monolayer

interconnected by tight junctions, undergo a transition to form round cells that express hematopoietic markers (CD41, CD45), briefly accumulate in the form of clusters, detach from the endothelial layer, and enter the circulation. Hemogenic endothelium is found in multiple anatomic sites in the embryo including the yolk sac, the vitelline and umbilical arteries, and the dorsal aorta where it is flanked by the developing urogenital ridges in the so-called aorta/gonad/mesonephros (AGM) region (North et al., 1999). The first hemogenic endothelial cells in the mouse appear in the yolk sac at approximately embryonic day (E) 8.5, they are most abundant in the major arteries at E9.5, and the majority of them complete their transition to hematopoietic cells between E9.5 and E12.5 (North et al., 1999; Yokomizo et al., 2012). Each endothelial to hematopoietic cell transition takes approximately 5 hours to execute, and all hemogenic endothelial cells undergo the transition into a blood cell during a 3-4 day period (Bertrand et al., 2010; Boisset et al., 2010; Eilken et al., 2009; Kissa and Herbomel, 2010). The formation of hematopoietic progenitors from ES cells, and through direct reprogramming of endothelial cells with Runx1, Gfi1, Fosb, and Spi1 also proceeds through a hemogenic endothelial intermediate, recapitulating the normal developmental process (Eilken et al., 2009; Lancrin et al., 2009; Sandler et al., 2014).

The endothelial to hemogenic endothelial cell transition is a fascinating process to behold. However, almost nothing is known about the transcriptional regulatory network that is responsible for this transition. Hemogenic endothelium from yolk sac and major arteries differ with respect to the types of hematopoietic progenitors they produce. Yolk sac hemogenic endothelium produces primarily committed erythroid/myeloid progenitors, whereas embryonic endothelium produces HSCs (Vo and Daley, 2015).

3.2 Results

3.2.1 Overall transcriptome similarity between non-hemogenic endothelium and hemogenic endothelium

AGM HE vs. YS HE comparison revealed more genes involved in other developmental lineage whereas comparison of HE vs. E from the same tissue did not reveal as many genes involved in other developmental lineages. We first performed PCA analysis by incorporating additional fetal liver (FL) and bone marrow (BM) HSC samples which are considered more matured. We found that samples within the same tissue tend to group together (Figure 21A). To understand the difference between HE and E and the difference between different HEs, we conducted 3 pairwise comparisons at each time points (Figure 21B), including arteries HE vs. E, YS HE vs. E, and arteries HE vs. YS HE. Consistent with PCA analysis, we also observed many more differentially expressed genes for HE comparisons than HE vs. E. Additionally, we found that the difference between AGM HE and YS HE across time points is more constant than HE vs. E in terms of number of shared DEGs.

Next, we used the expression profile of 4,790 genes that were at least differentially expressed in one of the 6 pairwise comparisons described above. After running consensus clustering, we identified 9 gene clusters (Figure 21C). According to the expression pattern of each gene cluster, we classified them into 3 categories: 1) clusters whose member genes have higher expression in E than HE in both tissue types and time points; 2) clusters whose member genes have higher expression in HE than E in both tissue types and time points; 3) clusters whose gene members do not belong to the first 2 categories or have clear expression pattern (Figure 21C).

To understand the functions mediated by each gene clusters, we conducted Gene Ontology (GO) enrichment analysis (Figure 22). Clusters were grouped by functional category annotations, suggesting gene clusters with similar expression patterns are involved in similar biological functions. In particular, most gene clusters show over-representation of "cell migration", "cell morphogenesis", "regulation of apoptosis", "tissue development" and "Wnt receptor signaling pathway". We also found biological functions that are specific to each cluster category. For example, "Steroid metabolism", "phosphorylation" and "tissue homeostasis" are specific to cluster category 2. Steroid ablation induced lymphoid hematopoietic recovery by functionally enhancing both HSC self-renewal and propensity for lymphoid differentiation (Khong et al., 2015). By contrast, terms that are specific to cluster category 1 include "intracellular protein transport", "regulation of TGF beta signaling", "biopolymer glycosylation" and "protein catabolism", which suggests TGF beta signaling may have different activities in HE cells.

In addition to GO analysis, we utilized Gene Set Enrichment Analysis (GSEA) to systematically identify signal transduction pathways with significant activity change between two different conditions (Subramanian et al., 2005). We used the expression profile of all expressed genes for a given comparison to search against signal transduction pathways extracted from Reactome database (Croft et al., 2011). Pathways with significant activity change were summarized for all comparisons, and we found that identified pathways are generally unique to specific comparisons (Figure 23). For instance, identified pathways for HE vs. E comparison in AGM are include Hedgehog, Wnt and MAPK family signaling, whereas pathways for HE vs. E in YS include Notch, Wnt and TGF-beta signaling.

The Hedgehog signaling is highly conserved and plays an essential role in embryonic development and is also required for both primitive and definitive hematopoiesis (Lim and Matsui, 2010). In this study, we found that the Hedgehog signaling activity was significant different between HE and E. Sonic Hedgehog gene (Shh), a key gene in the Hedgehog signaling pathway, was highly up-regulated in HE. In addition, three important TFs, Gli1, Gli2 and Gli3 in the Hedgehog signaling pathway re also highly up-regulated in HE. This suggests that the Hedgehog signaling is very likely to play a role in the transition from endothelium to hemogenic endothelium.

3.2.2 Construction of condition-specific transcriptional regulatory networks

Existing computational methods are not designed for constructing condition-specific TRNs. To this end, we developed a novel method for constructing condition-specific transcriptional network based on sample elimination and network comparison. After building a pan-hematopoietic expression compendium with 138 samples, we applied our method to construct 6 TRNs that corresponds to 6 differential expression comparisons. On average, each TRN contains about 224,000 edges. With the constructed TRNs, we scored all edges using differential expression p-values in each comparison respectively. Next, we ranked all TFs in each TRN by computing the average shortest distance from a TF to all DEGs. Statistical significance was estimated by computing the distance from TFs to randomly selected genes, and combined p-values were also computed using Fisher's method for HE vs. E or HE comparisons (Fisher, 1925). In total, we identified 41 and 37 TFs for HE vs. E and HE comparisons respectively (Figure 24 B and C).

3.2.3 Key transcriptional factors that play a role in the development of hemogenic endothelium

Among top ranked TFs for HE vs. E comparison, we identified Spi1, Runx1 and Gfi1 which were previously reported as required factors to reprogram endothelial cells to hematopoietic cells (Sandler et al., 2014) (Figure 24 B). TEAD factors such as Tead3 and Tead1, identified as important factors in HE vs. E comparison, were also found to regulate hematopoietic specification in a recent study (Goode et al., 2016). For HE comparisons, we found a number of HOX family genes including Hoxa4, Hoxd8, Hoxa7, Hoxc8, Hoxb6 and Hoxa10. HOX family genes are known to be involved in hematopoiesis and hematopoietic stem cell function (Argiropoulos and Humphries, 2007). Therefore, the newly identified HOX genes potentially play a role in differentiating AGM HE from YS HE.

3.3 Discussion

In this chapter, I developed a computational framework to construct condition-specific TRNs when the number of samples for a condition of interest is limited. Existing computational methods all need a relatively large number of samples for a given condition to reach enough statistical power (Marbach et al., 2012). My method takes advantages of public gene expression data to address the issue of limited number of samples. A gene expression compendium is built by collecting data in related cell types from public databases. The method then constructs two regulatory networks, one using data for all conditions and one using data for all conditions minus the condition of interest. The method then compares the two networks to construct a condition-specific TRN. To validate the method, I collected gene expression data for different blood cell types from multiple data sources. For each cell type, I also extracted a set of ChIP-Seq

data for TFs in corresponding cell types from the GEO database. By comparing constructed condition-specific TRNs with corresponding ChIP-Seq data, we found our predicted TF-target gene interactions significantly overlap with the ChIP-Seq data, suggesting a good performance of my method.

I applied my method to construct condition-specific TRNs for HSC development data including 8 cell populations. To understand transcriptional regulation for the development of hemogenic endothelium, I identified transcriptional regulatory relationships that occur during the endothelial-to-hematopoietic transition. I also utilized the differential expression information to weight the regulatory network, so genes with large expression change are more likely to participate in the transition. To quantitatively measure the importance of a TF in the regulatory network, I developed a method to rank TFs using their average distance to differentially expressed genes. With this method, I prioritized top TFs that are likely to contribute to the cell fate transition. Among selected TFs, many of them have a previously reported role in HSC development including endothelial cell reprogramming factors, TEAD family factors and HOX family genes (Goode et al., 2016; Magli et al., 1997).

3.4 Materials and methods

3.4.1 Endothelial Tube Assay

Assessment of endothelial tubes was performed as previously described (Medvinsky et al., 2008). Briefly, sorted hematopoietic populations were cultured in aMEM, 10% FBS, 4mM glutamine, 0.1 mM 2 mercaptoethanol, 50 U/ml penicillin, 50 ug/mL streptomycin and 50 ng/mL VEGF (Peprotech, USA) for 4 days. Endothelial tubes

were labeled with anti-PECAM (BD Biosciences) and visualized with Vector Blue AP substrate kit (Vector Labs, Burlingame, CA).

3.4.2 RNA purification and sequencing

Cells were sorted into Trizol and RNA was purified (Qiagen). Total RNA was quantified with RNA HS Kit (Q32852) for Qubit fluorometer (Life Technologies) and analyzed for integrity using RNA 6000 Pico Kit (5067-1513) for 2100 Bioanalyzer (Agilent). mRNA was isolated from total RNA using NEBNext Poly(A) mRNA Magnetic Isolation Module (E7490, NEB). PolyA-selected mRNA was fragmented to average size of 300 nt, reverse transcribed to generate double-stranded cDNA, and converted to a paired end library using NEBNext Ultra RNA Library Prep Kit for Illumina (E7530, NEB) according to manufacturer's instructions including the optional double size selection procedure using Agencourt AMPure XP beads. Prepared libraries were quantified with dsDNA HS Kit (Q32851) for Qubit and the size distribution was assessed using High Sensitivity DNA Kit (5067-462) for Bioanalyzer. Libraries were sequenced on Illumina HiSeq2500 in paired-end mode with the read length of 75 nt.

RNA was prepared using RNeasy micro kit (Qiagen, Valencia, CA, USA). Samples were pooled by equivalent cell numbers onto a single column. DNase treatment was performed either on-column (Qiagen) or using DNA-*free* DNA Removal Kit (Life Technologies).

3.4.3 Transcriptome assembly and expression level estimate from read counts

Paired-end reads were mapped to the reference mouse genome (release mm9) using Tophat (Trapnell et al., 2009). Only uniquely mapped reads with fewer than 2 mismatches were used for downstream analyses. Transcripts were assembled using

Cufflinks (Trapnell et al., 2010) using mapped fragments outputted by Tophat. Ensemble (release 66) was used as the source of annotated genes and transcript isoforms. Normalized transcript abundance was computed using Cufflinks and expressed as FPKM (Fragments Per Kilobase of transcripts per Million mapped reads). Gene-level FPKM values were computed by summing up FPKM values of their corresponding transcripts (Trapnell et al., 2010). Following previous studies (Mortazavi et al., 2008), we used a FPKM value of one as the cutoff for expressed genes. It roughly represents 1 copy of gene per cell. RNA-Seq data reproducibility was assessed by computing spearman correlation of gene expression between a pair of biological replicates (Figure 27). Genes with zero read counts in all biological replicates were excluded from correlation calculation.

3.4.4 Identification of differentially expressed genes

FeatureCounts (Liao et al., 2014) was used to summarize read counts for genes. Ensemble (release 66) was used as the source of known transcripts for each gene. With normalized read counts for each gene, EBSeq (Leng et al., 2013) was used to detect significantly differentially expressed genes. We used a false discovery rate (FDR) cutoff of 0.05 and a fold change of 1.5 to call differentially expressed genes. To answer our research question, we conducted 3 pairwise comparisons that are AGM HE vs. AGM E, YS HE vs. YS E and AGM HE vs. YS HE at both E9.5 and E10.5 (6 comparisons in total).

3.4.5 Clustering of gene expression profiles

To identify gene clusters with similar expression patterns across different conditions, we applied consensus clustering which is considered as a robust clustering method (Stefano Monti, 2003). We only used the expression profile of genes that are

differentially expressed in at least one of the 6 comparisons to do the clustering. GO enrichment analysis was also conducted for each cluster to uncover related biological functions.

3.4.6 Identification of signal transduction pathways with significant change during endothelial-to-hematopoietic transition

To gain biological insights into the dynamics of signal transduction pathways between different conditions, we conducted GSEA analysis for 6 comparisons mentioned above using expression profiles of all expressed genes (Subramanian et al., 2005). We obtained annotations for signal transduction pathways from the Reactome pathway database as the input to GSEA (Croft et al., 2014).

3.4.7 Construction of condition-specific transcriptional regulatory networks using gene expression profiles

We compiled a set of 146 gene expression profiles for cells of the hematopoietic system from the following five studies, (McKinney-Freeman et al., 2012), (Li et al., 2014), (Bagger et al., 2012), and this study. Before combining datasets from different studies, we performed quantile normalization using the Robust Multi-array Average (RMA) algorithm and batch effect removal using the ComBat method (Leek et al., 2012; Wilson and Miller, 2005).

A number of methods have been developed to infer TRNs using gene expression profiles alone. These methods are collectively known as the reverse engineering methods. A recent study assessed 35 reverse engineering methods using gold-standard experimental data (Marbach et al., 2012). The study revealed that no single inference method performs optimally across all data sets. In contrast, integration of predictions from multiple inference methods shows robust and high performance across diverse data

sets. We thus sought to build a consensus transcriptional regulatory network by using five network inference methods including a method using Pearson correlation, the context likelihood of relatedness (CLR) method (Faith et al., 2007), Inferelator (Bonneau et al., 2006), trustful inference of gene regulation using stability selection (TIGRESS) (Haury et al., 2012) and gene network inference with ensemble of trees (GENIE3) (Huynh-Thu et al., 2010). These 5 methods ranked at the top of their corresponding category according to a recent benchmarking study (Marbach et al., 2012).

To infer conditional-specific TRNs, starting from the expression profile for all samples S_{all}, another expression profile $S_{all\text{-}condition}$ will be created by excluding samples we are interested in. Next, we will build consensus TRN G_{all} and $G_{all\text{-}condition}$ using the two expression profiles respectively. To obtain condition specific TRN $G_{condition}$, edges ($E_{all\text{-}condition}$) in $G_{all\text{-}condtiion}$ will be eliminated from G_{all} (Figure 24A).

3.4.8 Performance benchmarking

To determine if our TRN construction method can capture real interactions between transcription factors (TFs) and their targets, we used a series of published ChIP-Seq datasets for transcriptional factors as gold standard to evaluate the performance including B cell, nucleated erythrocytes, hematopoietic progenitor (HP), multipotent hematopoietic progenitor cell line 7 (HPC-7), macrophage, mast, nature killer and T cell (Table 2). For a given TF, we defined the true target genes as genes that have the nearest transcription start sites to TF peak binding sites. Next, we constructed TRNs for each cell type using gene expression data respectively. Compared to ChIP-Seq data, our predictions generally recovered a large set of true regulatory interactions. Additionally,

we found the most of identified interactions are specific by comparing our predictions across cell types (Figure 28).

3.4.9 Prioritization of key transcription factors in a TRN

We used the constructed TRNs to identify key transcriptional factors with a role in endothelial-to-hematopoietic transition. To this end, we assume that key TFs are closer to the set of differentially expressed genes in the TRN, either via direct or indirect connections. Based on this assumption, we computed a distance between two genes, i and j, in the TRN as following: $W(i,j) = 1 - \frac{\log(p_i)+\log(p_j)}{2\log(p_{min})}$, where p_i and p_j are the differential expression p-values for gene i and j, respectively. p_{min} is the minimum differential expression p-value among all genes in the TRN. With the distance-weighted TRNs, we calculated an average shortest distance between a given TF and all differentially expressed genes in the network. We computed the pairwise shortest path using dijkstra's algorithm. Statistical significance of an average shortest distance was computed using a null distribution computed based on the given TF and randomly selected genes (Figure 24A).

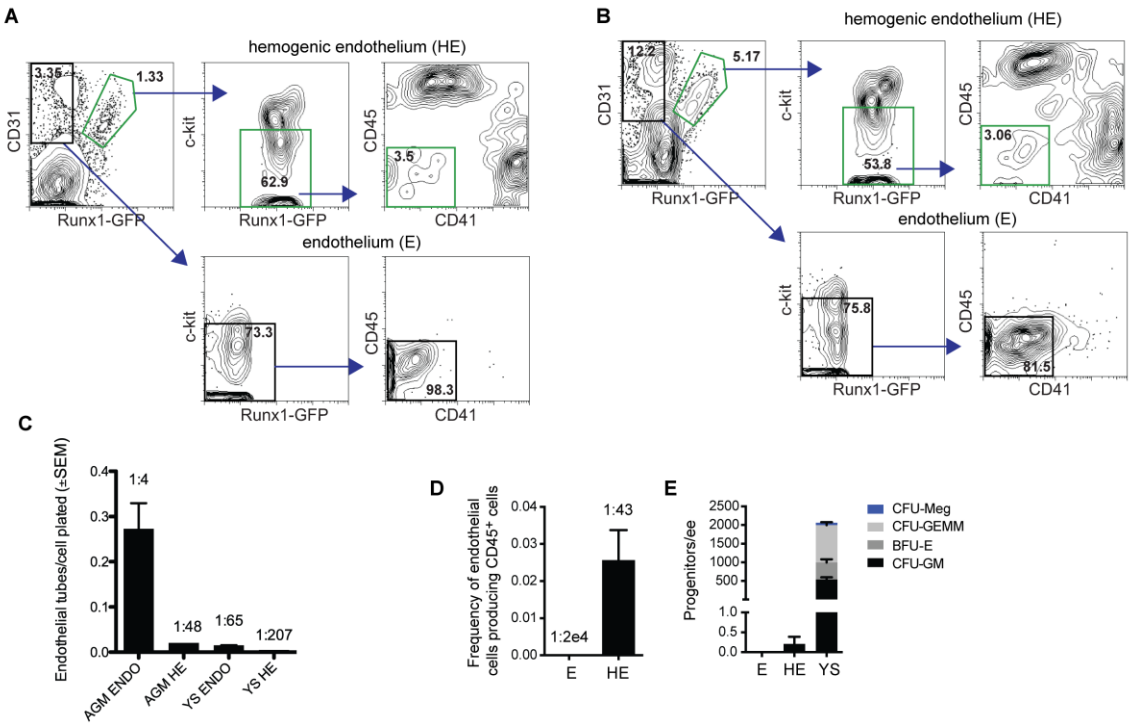

Figure 20. Functional characterization of hemogenic endothelium and endothelium. **A)** FACS gating strategy to purify hemogenic endothelial and **B)** endothelial cells. **C)** Tube formation assay for non-hemogenic endothelial cells. **D)** Hemogenic endothelial potential assay. **E)** Colony formatting unit potential for hemogenic endothelial cells.

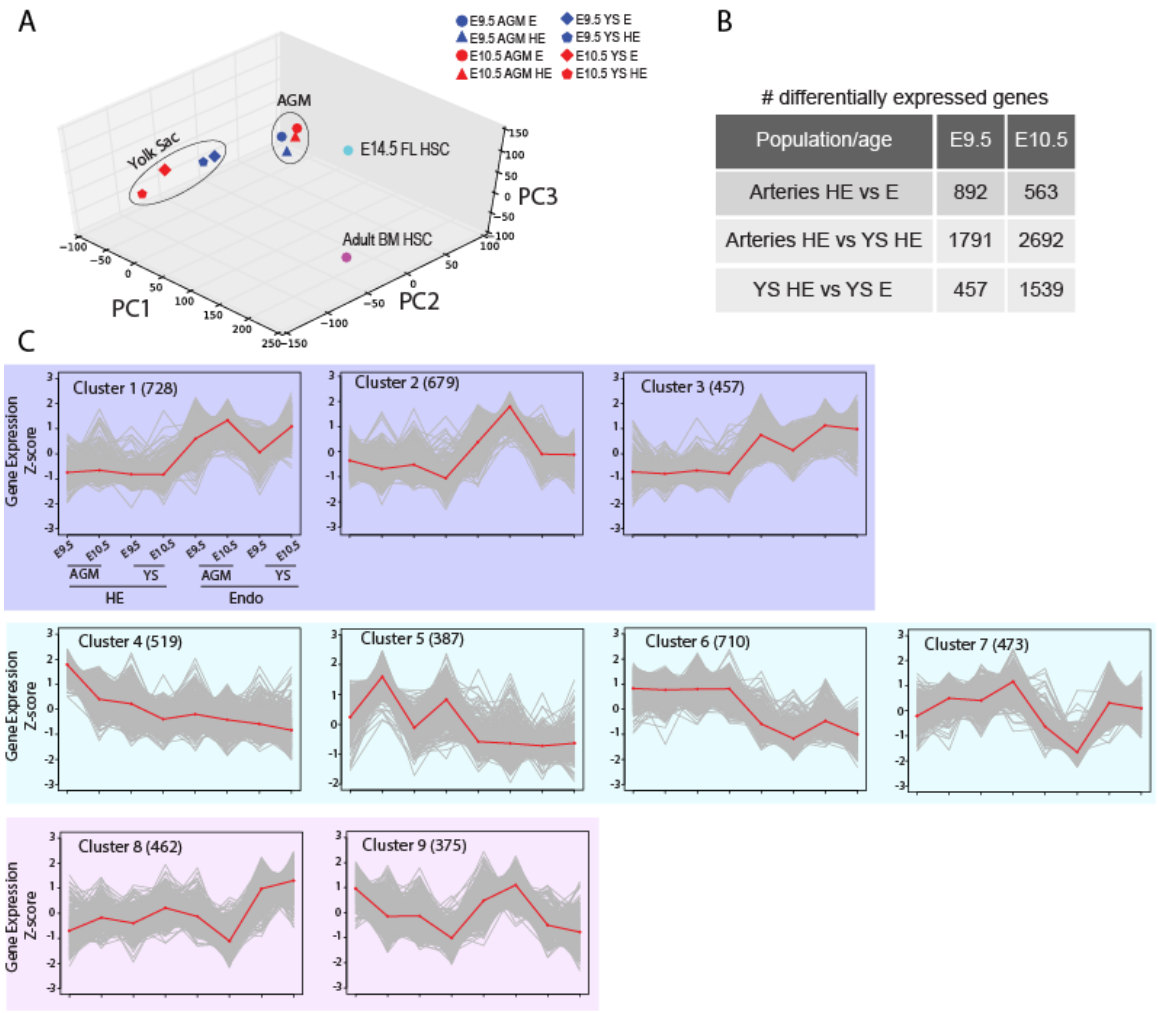

Figure 21. Global comparison of the transcriptomes of hemogenic endothelium and non-hemogenic endothelium. A) Principle component analysis of the transcriptome data. **B)** Number of differentially expressed genes based on pairwise comparisons. C**)** According to the expression pattern of each gene cluster, we classified them into 3 categories: 1) clusters whose member genes have higher expression in E than HE in both tissue types and time points; 2) clusters whose member genes have higher expression in HE than E in both tissue types and time points; 3) clusters whose gene members do not belong to the first 2 categories or have clear expression pattern.

Figure 22. Heat map of enriched GO terms for each gene cluster identified by consensus clustering. GO enrichment analysis was performed for each cluster, and clusters are grouped by category. Gene clusters with similar expression patterns are involved in similar biological functions. Color intensity indicates the minus logarithm of GO enrichment p-value.

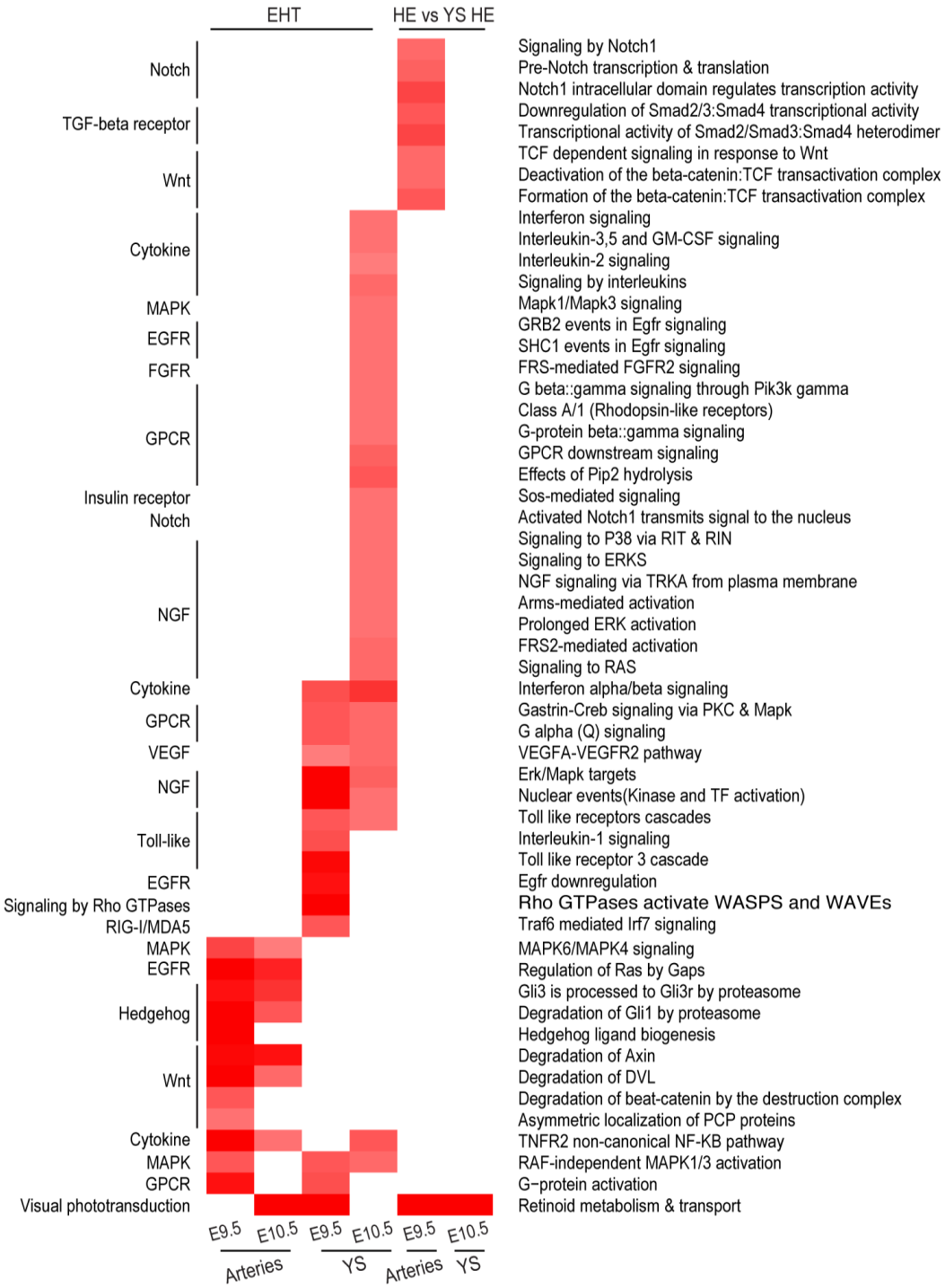

Figure 23. Difference in signal transduction pathways between hemogenic endothelium and non-hemogenic endothelium. Identified pathways for HE vs. E comparison in AGM contain Hedgehog, Wnt and MAPK family signaling, whereas pathways for HE vs. E in YS include Notch, Wnt and TGF-beta signaling. Color intensity indicates the minus logarithm of pathway enrichment p-value.

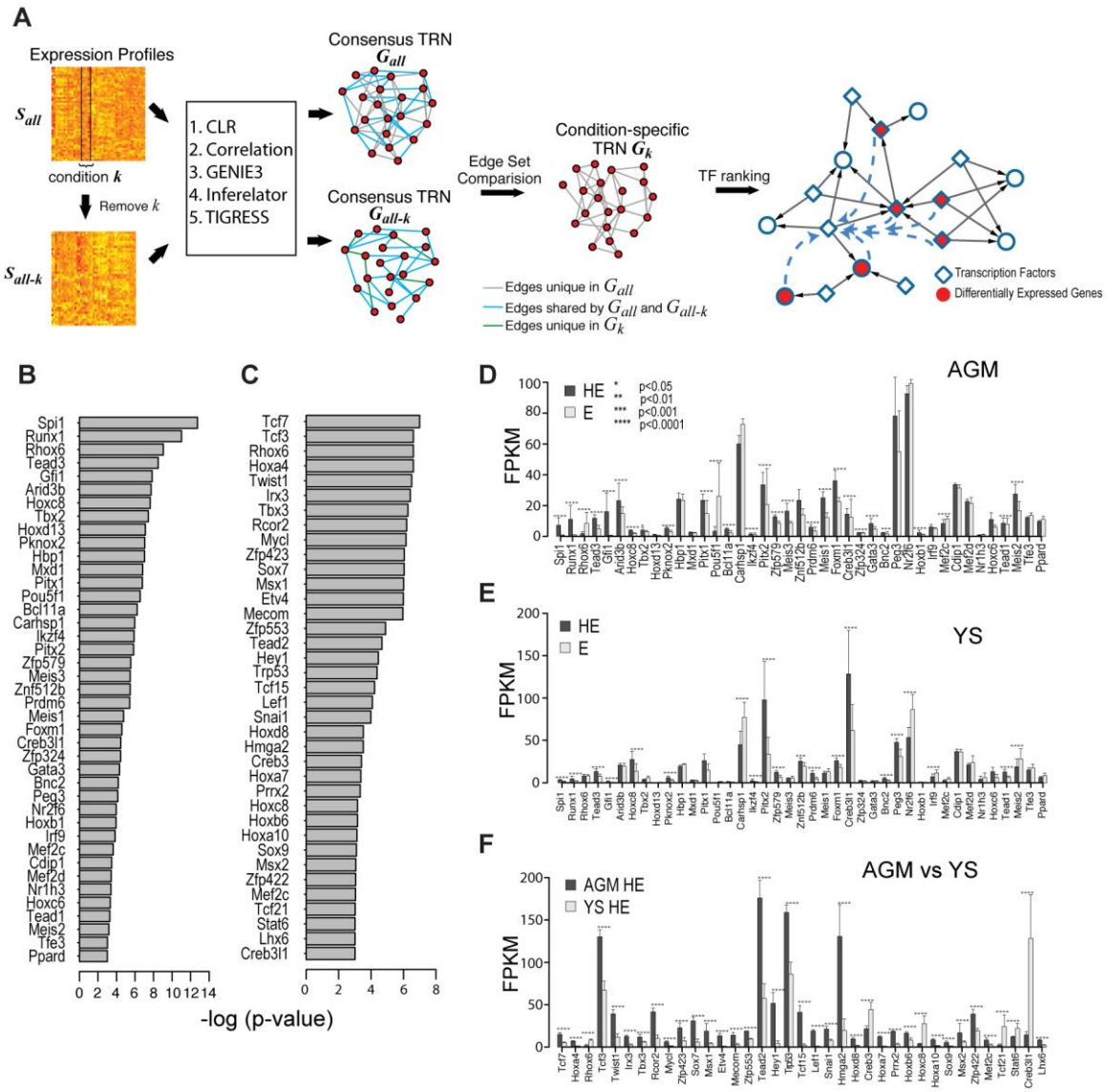

Figure 24. Transcriptional factors controlling endothelial-to-hematopoietic transition. A) The computation framework for constructing condition-specific TRN and TF ranking. **B)** Identified key TFs for endothelial-to-hematopoietic transition. **C)** Identified key TFs for hemogenic endothelial comparison between AGM and YS. **D-F)** Expression levels of identified key TFs for HE vs. E in the same tissue and HE comparisons between AGM and YS.

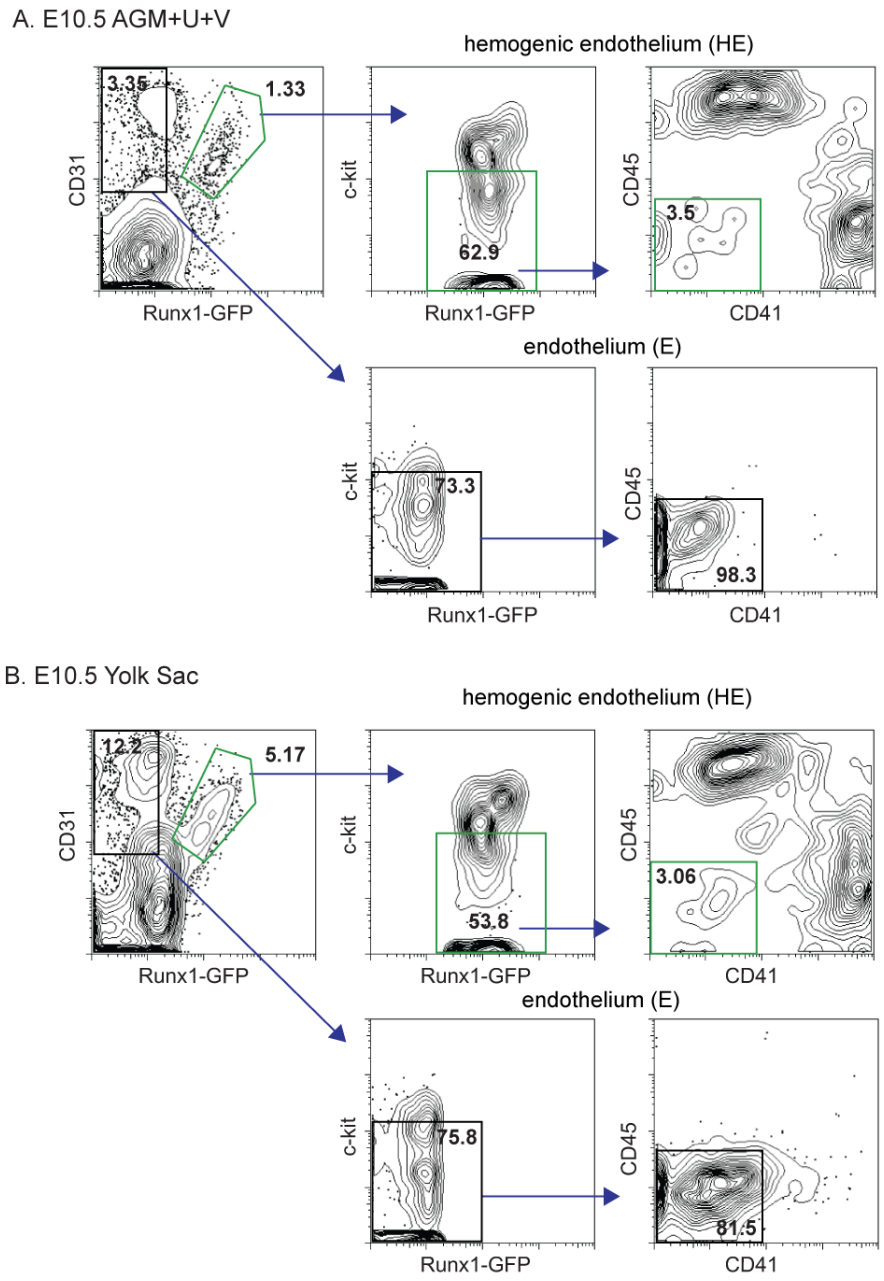

Figure 25. Isolation of hemogenic endothelial and endothelial cells from mouse embryo by FACS.

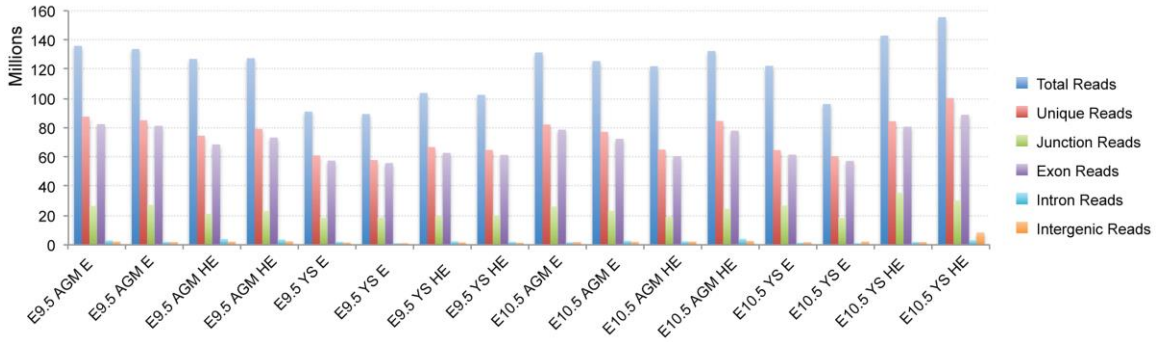

Figure 26. RNA-Seq read mapping statistics.

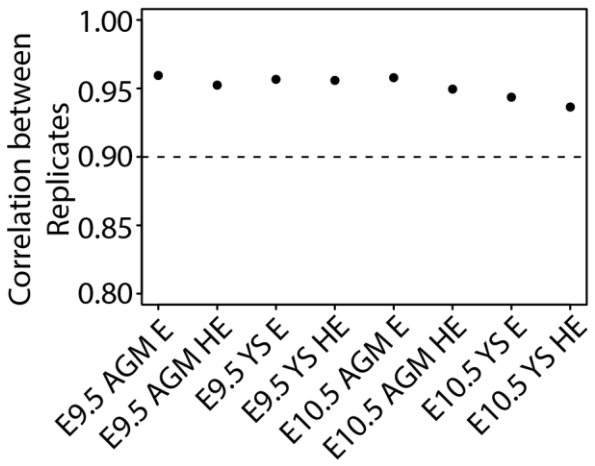

Figure 27. Correlation RNA-Seq data of biological replicate samples.

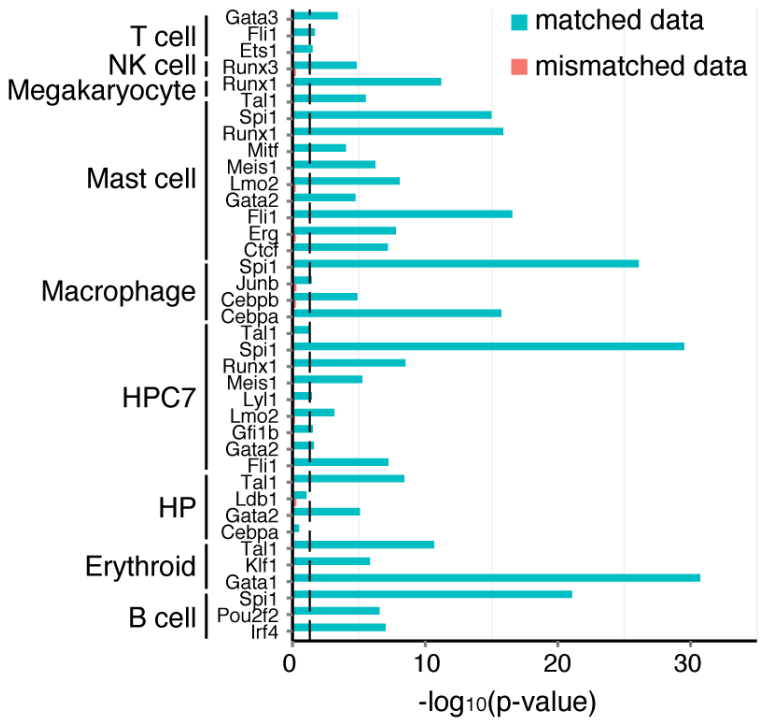

Figure 28. Performance benchmarking of our algorithm for inferring condition-specific TF-target interactions. ChIP-Seq data for a given TF was used as the gold standard. The gene closest to a TF ChIP-Seq peak was assigned as the target of the TF. Blue bars represent the overlap p-value between our predictions and gold standard TF targets in the same cell type (i.e. matched data). As a control, red bars represent the average overlap p-values between our predictions in a given cell type and ChIP-Seq targets of the same TF in other cell types (i.e. mismatched data). Overlap p-value was computed using the hypergeometric distribution.

Table 2. Gene expression data set used to construct the hematopoietic cell specific co-expression networks. Total number of samples was 146. HE, hemogenic endothelium; E, vasculature endothelium; YS, yolk sac; WBM, whole bone marrow; NK, natural kill cell; CD4, CD4+ T cell; CMP, common myeloid progenitor; GMP, granulocyte macrophage progenitor cell; CLP, common lymphoid progenitor; LT-HSC, long-term HSC; ST-HSC, short-term HSC; MEP, megakaryocyte-erythroid progenitor cell; CFU-E, colony-forming unit erythroid cell; IgM+SP, IgM positive spleen cell; LMPP, lymphoid-primed multipotential progenitor cell; MkP, megakaryocyte precursor cell; MkE, megakaryocyte erythroid cell; preB, pre-B cell; PreCFU-E, pre-colony-forming unit erythroid cell; ProB, B-cell progenitor cell; ProE, erythroid progenitor cell;

Reference	Accession number	Cell types
This study	N/A	E9.5 AGM HE (2), E9.5 AGM E (2), E10.5 AGM HE (2), E10.5 AGM E (2), E9.5 YS HE (2), E9.5 YS E (2), E10.5 YS HE (2), E10.5 YS E (2)
McKinney-Freeman S et al. (47)	GSE37000	E 9 YS HSPC (5), E11.5 AGM HSPC (6), E12.5 FL HSC, E13.5 FL HSC (), E14.5 HSC (), WBM HSC (5), ESC-HSC (3), EB (6), E12.5 Placenta HSPC (6)
Li Y et al. (8)	GSE55493	E11.5 endo_GFP_pos (2), E11.5 endo_GFP_neg (2), E11.5 HE_GFP_pos (2), E11.5 HE_GFP_neg (2)
Chambers SM et al. (20)	GSE6505	LT-HSC (2), NK (2), naïve T (4), activated T (4), B (2), monocyte (2), granulocyte (2), nucleated erythrocyte (2)
Di Tullio A et al. (47)	GSE14833	CD4 (2), CFU-E (3), CLP (2), ETP (3), GM (2), GMP (3), IgM+SP (2), LMPP (5), LT-HSC (2), MkE (4), MkP (3), Nkmature(3), PreB (4), PreCFUE (3), ProB (2), ProE (2), ST-HSC (2)

CHAPTER 4: A COMPUTATIONAL METHOD TO UNCOVER CAUSAL NON-CODING VARIANTS FOR DISEASES

Abstract

Technology development has revealed thousands of common and rare DNA sequence variants, the vast majority of which are located in noncoding regions. Interpretation of these variants remains a daunting task. Molecular networks have been used extensively to improve the inference accuracy of causal coding variants. This potential has not been investigated to the same extent for noncoding variants. We hypothesize that disease-relevant gene regulatory network can significantly improve the inference accuracy of noncoding risk variants. Here we present Annotation of Regulatory Variants using Integrated Networks (ARVIN), a general computational framework for identifying causal noncoding variants that affect a specific trait. We first developed a strategy to construct disease-relevant gene regulatory networks. We then developed and characterized multiple network-based features which are more discriminative than existing genomic and epigenomics features. We applied our method to 233 regulatory promoter variants annotated in the Human Gene Mutation Database (HGMD) for 20 diseases and 15 enhancer variants for 10 diseases from published literature. In total, there are 22 unique diseases. ARVIN outperforms state-of-the-art methods using genomic and epigenomic features alone. We identified causal SNP candidates and pathways which are likely to play roles in seven autoimmune diseases. Detailed analysis of predicted causal SNPs revealed that they tend to target disease associated genes. Further, enhancers harboring causal SNPs potentially have a combinatorial regulatory role in disease pathogenesis.

4.1 Introduction

Genome wide association studies (GWASs) and whole-genome sequencing have revealed thousands of DNA sequence variants associated with different traits/diseases (Consortium et al., 2015; Hindorff et al., 2009; Kandoth et al., 2013). The vast majority of identified variants are located outside of known protein-coding sequences, making direct interpretation of their functional effects challenging. In the majority of cases, causal noncoding variants have been shown to perturb binding sites of transcription factors, local chromatin structure or co-factor recruitment, ultimately resulting in changes of transcriptional output of the target gene(s).(Chorley et al., 2008; Freedman et al., 2011; Noonan and McCallion, 2010).

Among the different classes of noncoding regulatory sequences, transcriptional enhancers represent the primary basis for differential gene expression, with many human diseases resulting from altered enhancer action (Noonan and McCallion, 2010); (Epstein, 2009; Noonan and McCallion, 2010; Visel et al., 2009). Numerous recent studies (both large- and small-scale) have uncovered tens of thousands of putative enhancers in a diverse array of human cells and tissues (Andersson et al., 2014; Consortium, 2004; Firpi et al., 2010). Overlapping catalog of variants and enhancers has revealed an enrichment of disease-associated variants in tissue-specific enhancers, emphasizing the importance of knowledge about tissue-specific cis-regulatory sequences for identifying causal variants. Hereby, we term SNPs located in enhancer *eSNPs*. A number of computational methods have been developed to nominate causal noncoding variants. Conceptually, these methods operate by annotating genetic variants using a catalog of cis-regulatory

sequences (based on chromatin accessibility, transcription factor binding, epigenetic modification) (Khurana et al., 2013; Kircher et al., 2014; Ritchie et al., 2014). Although biologically intuitive, such an approach does not prove molecular function and causality. Current approaches do not take into account the complex interactions among the genes affected by regulatory variants. Molecular networks have been used to improve the inference accuracy of causal coding variants (Jia et al., 2011; Lee et al., 2011; Linghu et al., 2009; Moreau and Tranchevent, 2012). This potential has not been examined for noncoding variants. One proposed algorithm, GWAVA (Ritchie et al., 2014) prioritizes variants using 175 genomic and epigenomic features, but does not employ any network related information. Another state of the art method, FunSeq (Khurana et al., 2013), employs only a single binary feature denoting whether the target gene of the variant is a hub and it does not employ the network related information sufficiently for ranking variants. To address these shortcomings, we postulate that 1) the impact of causal eSNPs on gene expression is transmitted through the Gene Regulatory Networks (GRNs) of the cell/tissue type(s) that are relevant to studied trait; and 2) the genes affected by the full set of causal eSNPs for a trait are organized in a limited number of pathways. We explored this hypothesis by developing a general computational framework for identifying causal noncoding variants that affect a specific disease/trait.

Linkage disequilibrium presents another challenge for finding causal noncoding variants in human diseases/traits. By casting the causal inference problem into a subnetwork identification problem, our method evaluates both index and linked SNPs simultaneously, thus increasing the power of the inference. Further, our network-based approach naturally provides a pathway content for the resulting causal eSNPs. We first

characterized the performance of our method using known promoter mutations in 20 diseases and enhancer mutations in 10 diseases. We also applied our method to uncover novel causal variants associated with seven autoimmune diseases.

4.2 Results

4.2.1 Construction of an integrative and disease-relevant gene regulatory network

A number of previous studies have reported enrichment of GWAS SNPs in regulatory DNA sequences specific to the disease-relevant tissues or cell types (Farh et al., 2015; Maurano et al., 2012), emphasizing the importance of knowledge about tissue-specific regulatory sequences for identifying risk variants. Additionally, gene-gene and protein-protein interaction networks have been used to identify causal coding variants (Hofree et al., 2013; Lee et al., 2011; Zhang et al., 2013a). Because the effects of non-coding variants are transcriptionally integrated, a network-based approach should be an effective strategy to identify causal noncoding variants. To date, tissue-relevant gene regulatory network (GRN) has not been used explicitly to prioritize noncoding variants. As a first step towards this goal, we sought to construct an integrative regulatory network for each disease-relevant cell/tissue type. We integrate epigenomic, transcriptomic and functional gene-gene interactions to construct the network. A challenge in constructing gene regulatory network is linking transcriptional enhancers with their target promoters. By using our recently developed algorithm, IM-PET (Figure 29A) (He et al., 2014), we constructed 23 cell/tissue-specific enhancer-promoter (EP) networks that are relevant to the set of 10 diseases for risk enhancer SNPs in this study (Table 4). We evaluated our EP network construction method using a compendium of Hi-C and ChIA-PET chromatin interaction data from nine cell types (GM12878, K562, IMR90, HMEC, NHEK,

HUVEC, Hela, CD34+ cells, and CD4+ T cells). The overall Area Under the Precision and Recall Curve (auPRC) curve were 0.89 and 0.84 using Hi-C and ChIA-PET interactions as the gold standard, respectively (Figure 29B), suggesting high-quality EP predictions. Next, the enhancer-promoter networks were combined with a probabilistic functional gene interaction network inferred by integrating multiple lines of evidence (Lee et al., 2011). The resulting integrative GRN contains two types of edges, EP edge representing enhancer-promoter interaction and FI edge representing functional gene-gene interaction (Figure 29C). To add disease-specific information to an integrative network, we map genetic variants to enhancer sequences and differential gene expression between cases and controls to genes in the network. The final product is an edge- and node-weighted, disease-relevant gene regulatory network, which is used for predicting risk genetic variants. See Methods for additional details of the network construction procedure.

4.2.2 ARVIN combines sequence-based and network-based features to predict risk eSNPs

We hypothesize that disease-relevant regulatory network can improve the inference accuracy of noncoding risk variants. To this end, we examined a number of network-based features to see if they can discriminate the true risk SNPs from the negative control SNPs. We obtained 232 gold standard risk SNPs located in gene promoters from the Human Gene Mutation Database (HGMD). This set of promoter SNPs is associated with 20 different diseases (Table 3). We used the disease-relevant GRNs described above to extract the following network-based features: module score, weighted node degree, betweenness centrality, closeness centrality, and page rank centrality (see Methods for details). These features are designed to evaluate topological

importance of the direct target gene of an eSNP and the local network neighborhood of the target gene. Our hypothesis is that target genes with large topological importance in the GRN might be rate-limiting genes for a disease. We found that the set of network features can distinguish true risk SNPs from control SNPs (Figure 30A). To further test the network-based features, we built a random forest classifier using these features as well as sequence-based features used by two state-of-the-art methods, Genome-wide annotation of variants (GWAVA) (Ritchie et al., 2014) and FunSeq (Khurana et al., 2013). We first sought to evaluate the relative importance of all features (six from this study and 181 from GWAVA and FunSeq combined) by using recursive feature elimination. Applying the RFE procedure yielded a set of 35 most discriminative features based on classification error (Figure 35). Among them, all but one (closeness centrality) network features were top-ranked as compared to features of GWAVA and FunSeq (Table 10), suggesting the network-based features are independently discriminative from the sequence-based features. On the other hand, this analysis also suggests that network-based feature and sequence-based feature are complementary to each other. We thus developed the Annotation of Regulatory Variants using Integrated Networks (ARVIN) algorithm by combining network features with sequence features. We evaluated the classification accuracy using 5-fold cross validation and gold-standard risk SNPs in gene promoters. ARVIN achieved an area under the ROC curve (auROC) of 0.96. Compared to GWAVA and FunSeq, ARVIN achieved significant improvement over GWAVA ($p=1.7x10^{-12}$) and FunSeq ($p=2.2x10^{-16}$) (Figure 30C).

Many genes are regulated by distal enhancers. Compared to promoter variants, risk variants located in enhancers are more challenging to study due to the difficulty of

assigning enhancer targets and existence of multiple enhancers targeting the same gene. We further tested the performance of ARVIN using risk SNPs located in enhancers. To this end, we curated a set of 15 experimentally validated risk enhancer SNPs associated with 10 complex diseases, including autoimmune, heart, lung, psychiatric diseases, obesity and cancer (Table 4). Compared to promoter variants, the gold-standard set for enhancer variants is too small for ROC curve analysis to be meaningful. Therefore, for each risk SNP, we asked how it is ranked by a method among all SNPs in the same linkage disequilibrium block as the risk SNP. The number of linked SNPs ranges from 1 to 168 with an average of 28 (Table 4). Overall, both ARVIN and ARVIN with network feature alone (ARVIN-N) outperformed GWAVA and FunSeq. The average percentile ranking of the set of known risk SNPs were 14%, 29%, 48%, and 41% for ARVIN, ARVIN-N, FunSeq, and GWAVA, respectively (vertical lines, Figure 31).

In summary, using gold-standard risk SNPs in both promoters and enhancers, we demonstrate that incorporation of network features can significantly improve the accuracy of finding risk noncoding mutations.

4.2.3 Application of ARVIN to autoimmune diseases

We applied ARVIN to identify risk eSNPs associated with seven autoimmune diseases (systemic lupus erythematosus, psoriasis, rheumatoid arthritis, type 1 diabetes, Crohn's Disease, ulcerative colitis, and multiple sclerosis). We first obtained lead SNPs associated with those diseases from the National Human Genome Research Institute (NHGRI) GWAS Catalog (Welter et al., 2014). On average, 144 lead SNPs are associated with each disease. As candidate SNPs, we considered both lead SNPs and SNPs that are in the same linkage disequilibrium block with the tag SNPs. By

overlapping SNPs with enhancers from disease-relevant cell/tissue types, we obtained the list of eSNPs as the final input to ARVIN. On average, 123 GWAS lead SNPs are identified for each disease, yielding 66 eSNPs for each disease-associated loci tagged by a lead GWAS SNP.

We devised a strategy for choosing ARVIN prediction threshold by considering whether a candidate eSNP is located in a disease-associated region (see Methods for details). On average, we predicted 160 causal eSNPs for each autoimmune disease. We evaluated the predictions using the following orthogonal lines of evidence, including 1) eQTLs identified in disease-relevant tissues by the GTEx consortium and by Westra et al. (Carithers and Moore, 2015; Westra et al., 2013) (Table 7); 2) genes (n = 573) associated with autoimmune diseases that are documented in the ImmunoBase database (https://www.immunobase.org); 3) genes whose protein products are known drug targets (n = 227) of autoimmune diseases according to the Therapeutic Target Database (TTD) database (Yang et al., 2016); and 4) genes (n=776) whose promoters physically interact with cis-regulatory elements that harbor autoimmune disease-associated SNPs (Javierre et al., 2016; Martin et al., 2015). For six out of seven autoimmune diseases, the set of risk eSNPs predicted by ARVIN has significant overlap with the set of supporting evidence. By contrast, in only one disease, the set of predictions by GWAVA (T1D) and FunSeq (ULC) significantly overlap with supporting evidence (Figure 32A).

The mechanism of how causal variants exert their effect on enhancer activity and gene expression is investigated in multiple studies. In the studies that aim to reveal individual causal variants for some specific disease, it has been shown that the causal variants colocalize with transcription factor binding site and change the binding affinity

by altering the TF motif (Cowper-Sal lari et al., 2012; Oldridge et al., 2015; Tuupanen et al., 2009). However, in larger scale analysis of multiple variants, not in contradiction with those findings, it has been reported that variants can change TF binding remotely without changing the binding motif and only 10-20 percent of causal variants directly alter recognizable motifs (Farh et al., 2015; Kilpinen et al., 2013; Patwardhan et al., 2012). In parallel, for our study regarding to seven autoimmune diseases, only 173 out of 1120 predicted causal variants (15%) significantly disrupt the overlapping motif (multiple testing corrected p-value < 0.1), although 929 causal variants (83%) overlap with some known TF motif.

Increasing evidence suggest that many genes are regulated by multiple enhancers during normal and disease development (Chatterjee et al., 2016; He et al., 2014; Hong et al., 2008; Li et al., 2012; Sanyal et al., 2012). This phenomenon suggests that mutations in multiple enhancers of the same gene could collectively contribute to the deregulation of the gene during pathogenesis. Consistent with this hypothesis, we found that 32% of genes are targeted by multiple eSNPs which in turn are located in multiple enhancers (Figure 32B).

We found several unique features that distinguish genes targeted by multiple risk eSNPs from those targeted by a single risk eSNP. First, they tend to have higher network centrality measures (Figure 32C). Second, their expression levels are more perturbed in disease samples compared to control samples. The regulating risk eSNPs also have higher overlap with eQTLs (Figure 32D). Finally, they are enriched for more GO terms for immune responses (Figure 32E). Taken together, these unique properties of multi-targeted genes suggest they might be rate-limiting genes in disease pathogenesis.

Figure 32F shows two example genes that are targeted by multiple risk eSNPs. *IRF1* plays a critical role in regulatory T cell function and autoimmunity (Karwacz et al., 2017) (Figure 32F). It is targeted by two enhancers based on both IM-PET prediction and experimental Capture-Hi-C data in CD4+ T cells. Each risk eSNP (rs4143335, rs2706356) significantly disrupt the binding of HNF4A and E2F1, respectively. Both E2F1 (Lissy et al., 2000) and POU2F1 (Shakya et al., 2015) have been shown to be important transcriptional regulators of CD4+ T cell function. Another example involves the gene *PFKFB3* that encodes a rate-limiting glycolytic enzyme. Deficiency of PFKFB3 has been linked to reprogrammed metabolism in T cells from rheumatoid arthritis patients (Yang et al., 2013; Yang et al., 2015). Each targeting eSNP (rs77950884, rs17153333) significantly disrupt the binding of HNF4A and E2F1, respectively. Interestingly, in both cases, the lead GWAS SNPs are not predicted to be the risk SNPs, emphasizing the challenge of finding risk SNPs in the presence of genetic linkage.

4.2.4 Subnetwork comprising risk eSNPs and their target pathways

It has been suggested that the effects of multiple low-penetrance enhancer variants can be amplified through coordinated dysregulation of the entire gene regulatory network of a key disease gene, as illustrated in an elegant study by Chatterjee and colleagues (Chatterjee et al., 2016). To obtain a system-level view of the pathways collectively affected by all risk eSNPs implicated in a disease, we used the Prize Collection Steiner Tree (PCST) algorithm to identify a connected subnetwork composed of all risk eSNPs and genes bridging the risk eSNPs. By design, the resulting subnetwork is maximized for nodes and edges with large scores. In other words, these are genes (downstream of risk eSNPs) that have high levels of differential expression and

functional interactions. Therefore, the effects of the risk eSNPs are likely propagated via such a subnetwork.

For each disease, we identified the subnetwork downstream of risk eSNPs predicted by ARVIN, which we compared to subnetworks downstream of risk eSNPs predicted by GWAVA, and FunSeq, respectively. We found that subnetworks predicted by ARVIN have more enriched Gene Ontology (GO) terms related to immune cell functions, such as "cytokine-mediated signaling", "innate immune response" and "T cell activation". Thus, our identified SNPs may involve in the failure of self-tolerance immune mechanism, which is a critical part of autoimmunity.

Figure 33B shows an example subnetwork for rheumatoid arthritis. Such a network view reveals two interesting features of the perturbations caused by risk eSNPs. First, we found that multiple members of a pathway can be targeted by distinct risk eSNPs. For instance, the subnetwork contains ten genes that are involved in the *RhoA*-mediated small GTPase signaling. Six of the ten genes are individually targeted by risk eSNPs. Second, we found that many genes targeted by risk eSNPs are not located in disease-associated loci. This is consistent with the notion of long-range interaction between enhancers and their target genes.

4.3 Discussion

We presented a computational method for prioritizing candidate causal SNPs that reside in gene-distal enhancers. It is also expected that what is learned will be applicable to other classes of noncoding elements, such as regulatory sequences for alternative splicing. Although there are several computational methodologies proposed for predicting causal SNPs, to our knowledge, none of them are capable of inferring causal SNPs using

disease specific information. Those methodologies can theoretically identify causal SNPs, but do not provide disease information. It is obvious that many causal SNPs for one type of disease may be neutral for others. Our study proposes a novel methodology that prioritizes variants according to specific disease types.

In this study, we have demonstrated the application of our methodology for autoimmune diseases. The proposed method can be applied to any kind of disease for which histone modification data and gene expression data is available. We believe applying this method for somatic mutations and rare variants in cancer types can generate valuable findings. The computational framework is applicable to somatic mutations in cancer and rare variants**.** Because risk eSNPs function by perturbing GRNs, network-based view of risk eSNPs is the route for identifying components and mechanisms of pathogenesis.

Recent advances in technological tools for massively parallel, high-throughput sequencing of DNA such as whole-genome sequencing (WGS) have allowed the comprehensive characterization of somatic mutations in a large number of tumor samples. WGS for cancer provides a base-by-base view of the unique mutations present in cancer tissue (Watson et al., 2013). It allows discovery of novel cancer-associated variants, including single nucleotide variants (SNVs), copy number variations (CNVs), and structural variants (SVs). Through comparing tumor and normal DNA, WGS is able to provide a comprehensive view of alterations in specific tumor sample. Therefore, we can identify driver mutations in specific cancer using ARVIN algorithm. Regulatory mutations can be found by overlapping with enhancers predicted using corresponding histone modification marks. In addition, regulatory mutations will be linked to target

genes using computational predictions or chromatin interaction data. In this way, we can construct cancer specific gene regulatory network and run ARVIN to identify driver mutations. On the other hand, gene targets of causal eSNPs identified by ARVIN show very interesting topological features. Many of genes are also overlapped with known drug targets genes. Hence, target genes of risk eSNP are also ideal candidates for drug development.

4.4 Materials and methods

4.4.1 ARVIN framework

Key components of the computational framework are described in the following sections, including: construction of disease-relevant regulatory network, network-based features associated with candidate eSNPs, and classification strategy for risk eSNPs using genomic, epigenomic and network-based features.

4.4.2 Construction of weighted and disease-relevant regulatory network

Network construction starts with identifying eSNPs. For each lead GWAS SNP, we identify the linkage disequilibrium (LD) block to which it belongs. We then intersect the set of SNPs in the LD block with the set of enhancers from cell/tissue types relevant to the disease. This gives us a set of enhancer SNPs (eSNPs) in a given LD block identified by the lead GWAS SNP.

The gene regulatory network consists of two types of nodes, representing eSNPs and genes, and two types of edges, those between eSNPs and gene(s) (denoted as EP edges) and those between genes (denoted as FI edges) (Figure 29C). EP edges represent regulatory relationship between an enhancer and its target(s). FI edges represent functional interactions between genes. EP edges are based on enhancer-promoter

interactions predicted by the IM-PET algorithm (see below for details). FI edges are taken from HumanNet, which is a probabilistic functional gene network of 18,714 protein-encoding genes in humans (Lee et al., 2011). Each interaction in HumanNet has an associated probability representing a true functional linkage between two genes. It is constructed by a Bayesian integration of 21 types of 'omics' data including physical interactions, genetic interactions, gene co-expression, literature evidence, homologous interactions in other species, etc. HumanNet has proven to be very useful for improving inference accuracy of coding variants.

Nodes and edges in the network were weighted to 1) take into account the noise in the data; 2) to represent the relative importance of different genes and interactions. Weights for eSNPs, $\boldsymbol{w^{eSNP}}$, are based on the p-value of disruption of putative transcription factor binding site due to the SNP (see below for details). Weights for genes, $\boldsymbol{w^{DE}}$, are based on the p-values of differential gene expression between diseased and healthy conditions. Weights for EP edges, $\boldsymbol{w^{EP}}$, are based on the probability for enhancer-promoter interaction outputted by the IM-PET algorithm. Weights for FI edges, $\boldsymbol{w^{FI}}$, are taken from HumanNet. To make the values of each type of weights comparable, we conducted min-max normalization for each type of weights.

4.4.3 Network-based features associated with candidate eSNPs

1) Gene modules downstream of an eSNP. Our overall hypothesis is that a causal eSNP contributes to disease risk by directly causing expression changes in genes of disease-relevant pathways. Thus, in addition to the direct target gene of the eSNP, other genes in the same pathway can also provide discriminative information. With the weighted regulatory network, our goal is to identify “heavy” gene modules in the network

that connects a given eSNP to a set of genes (encircled modules in Figure 29C), hereby termed eSNP module. On the other hand, non-causal eSNPs are expected to be associated with "light" modules, i.e. having marginal impact on pathway gene expression (e.g. eSNP3 in Figure 29C). To score a candidate module, we used the following additive scoring scheme by summing up all node and edge weights divided by the number of nodes (*N*) in the candidate module.

$$\boldsymbol{S = (W^{eSNP} + W^{DE} + W^{EP} + W^{FI})/N}$$

We conducted module search from all eSNPs in the weighted network. It is a NP-hard to obtain a global optimal solution consisting of all heavy subnetworks. We thus used a greedy local search strategy. Starting with each eSNP, our algorithm considers all genes connected to the current eSNP-module and add the node whose addition leads to the maximal increase of the scoring function. This procedure repeats until there is no node whose addition can improve the module score. Several recent studies have reported that multiple enhancer elements could be present at a single GWAS locus (Corradin et al., 2014; Yang et al., 2012). Our network-based framework can naturally handles such cases because we consider all eSNPs simultaneously during module search. We assess the statistical significance of candidate modules using randomized networks. Specifically, for edges, we randomized them by edge-preserved shuffling. For nodes, we randomly shuffled their values within each type (i.e. among genes or among eSNPs). The empirical p-values are computed based on the null score distribution from the randomized networks.

2) Weighted degree of a node $\boldsymbol{v}$ directly downstream of an eSNP. It is defined as $\sum_{(u,v)\in E} \boldsymbol{W}(\boldsymbol{u},\boldsymbol{v})$, where $\boldsymbol{W}(\boldsymbol{u},\boldsymbol{v})$ is the edge weight for the edge connecting node $\boldsymbol{u}$ and $\boldsymbol{v}$.

3) Betweenness centrality of a node $\boldsymbol{v}$ directly downstream of an eSNP. Betweenness centrality quantifies the number of times a node acts as a bridge along the shortest path between two other nodes. The raw betweenness centrality is defined as $\boldsymbol{C_B}(\boldsymbol{v}) = \sum_{v\neq t\neq u} \frac{\sigma_{tu}(v)}{\sigma_{tu}}$, where $\boldsymbol{\sigma_{tu}}$ is the total number of shortest paths between node *t* and node $\boldsymbol{u}$. $\boldsymbol{\sigma_{tu}}(\boldsymbol{v})$ is the subset of $\boldsymbol{\sigma_{tu}}$ that go through *v*. The normalized betweenness centrality is defined as $\boldsymbol{C'_B}(\boldsymbol{v}) = \boldsymbol{C_B}(\boldsymbol{v}) * \boldsymbol{N}$, where *N* is the total number of nodes in the network.

4) Closeness centrality of a node $\boldsymbol{v}$ directly downstream of an eSNP. Closeness centrality is the inverse of the sum of shortest paths between a node and all nodes in a network. It can be regarded as a measure of how long it will take to spread information from a given node to all other nodes in the network. The raw closeness centrality is $\boldsymbol{C_C}(\boldsymbol{v}) = \frac{1}{\sum_{u\neq v} \boldsymbol{d}(\boldsymbol{u},\boldsymbol{v})}$, where $\boldsymbol{d}(\boldsymbol{u},\boldsymbol{v})$ indicates the length of the shortest path between *u* and *v*. The normalized closeness centrality is defined as $\boldsymbol{C'_C}(\boldsymbol{v}) = \boldsymbol{C_C}(\boldsymbol{v}) * \boldsymbol{N}$, where *N* is the total number of nodes in the network.

5) Page rank centrality of a node $\boldsymbol{v}$ directly downstream of an eSNP. Page rank centrality is a measure of the influence of a node in a network. It is based on the concept that connections to high-scoring nodes contribute more to the score of the node in question than connections to low-scoring nodes. Page rank centrality of a node $\boldsymbol{v}$ is defined as $\boldsymbol{C_P}(\boldsymbol{v}) = \frac{\boldsymbol{1-d}}{\boldsymbol{N}} + \sum_{v\in V(v)} \frac{\boldsymbol{C_P}(v)}{\boldsymbol{L}(v)}$, where $\boldsymbol{V}(\boldsymbol{v})$ is first neighbors of node *v* and $\boldsymbol{L}(\boldsymbol{v})$ is the set of edges incident on node *v*. *d* denotes a damping factor adjusting the derived value

downward and *N* is the total number of nodes in the network. The normalized page rank centrality is defined as $\boldsymbol{C'_P(v) = C_P(v) * N}$.

4.4.4 FunSeq and GWAVA features

FunSeq (Khurana et al., 2013) uses 6 binary features to determine if a variant is deleterious, including: 1) overlap with ENCODE annotation of cis-regulatory elements such as enhancer, promoter or DHS; 2) overlap with sensitive region (i.e. high level of negative selection); 3) overlap with ultrasensitive region; 4) disruption of a TF binding site; 5) target gene of the variant is known; and 6) target gene of the variant is a hub in a protein-protein interaction network. FunSeq feature values for candidate SNPs were obtained by SNP coordinates to FunSeq web portal.

GWAVA uses (Ritchie et al., 2014) 175 genomic and epigenomic features including overlap with histone modification and Transcription Factor ChIP-Seq peaks. We obtained GWAVA feature values for candidate SNPs using the various annotation data sources and Python script (gwava_annotate.py) provided in the GWAVA supplementary portal.

4.4.5 Predict risk variants using a random forest classifier with recursive feature elimination (RFE)

To classify risk eSNPs, we trained a random forest (RF) classifier using the combined feature set that consists of 6 network-based features, 6 binary features from FunSeq and 175 features from GWAVA. The classifier contained 500 decision trees. Each decision tree was built using ~20% of randomly selected training data (100 out of 464) and $\sqrt{\mathbf{187}} = \mathbf{14}$ randomly selected features. Classification error was measured with data not used for training (i.e. out of bag data). To compute feature importance, for each

decision tree, the classification error was computed using permuted and non-permuted feature values. The difference between the two classification errors were then averaged over all trees and used as feature importance.

To select most predictive features, we used a recursive feature elimination (RFE) strategy (Kuhn, 2008). At each iteration of feature selection, the top *S* most important features (based on feature importance) were selected. The random forest model was refit and corresponding performance was evaluated. To access the variance in performance at each iteration of feature selection, we did 5-fold cross validation. After all iterations, the optimal set of features was determined using the subset with best average performance across 5-fold cross validation. Receiver Operating Characteristic (auROC) curve is used to denote the prediction performance. To compare the performance between different classification tasks, a bootstrap strategy is used to do the ROC curve comparison (Pepe et al., 2009).

Based on the optimal set of features, we built a random forest classifier. Given a genetic variant along with its feature values, the classifier outputs a prediction probability indicating how likely this genetic variant is a risk variant in a given disease.

4.4.6 Identification of linkage equilibrium blocks

We used data from the 1000 Genomes project (phase 3 release) to identify SNPs in the same linkage disequilibrium (LD) with experimentally validated enhancer SNPs and GWAS catalog lead SNPs. PLINK (Purcell et al., 2007) was used to identify linked SNPs with D' > 0.9 and within 1Mb from either validated enhancer SNPs or GWAS lead SNPs.

4.4.7 Predictions of enhancers and enhancer-promoter interactions

Enhancers were predicted using the Chromatin Signature Inference by Artificial Neural Network CSI-ANN algorithm (Firpi et al., 2010). The input to the algorithm is the normalized ChIP-Seq signals of three histone marks (H3K4me1, H3K4me3 and H3K27ac). The algorithm combines signals of all histone marks and uses an artificial neural network-based classifier to make predictions of active enhancers with the histone modification signature "H3K4me1^{hi} + H3K4me3$^{neg/lo}$ + H3K27ac^{hi} ". The training set for the classifier was prepared using ENCODE data of mouse ES-Bruce4, MEL and CH12 cell lines. To create the training set for active enhancers, we first selected a set of promoter-distal p300 binding sites (2.5 kb away from Refseq TSS), and overlapped them with the histone modification peaks. The top 300 distal p300 sites that overlapped with H3K4me1 and H3K27ac peaks, but not H3K4me3 peaks, were selected as the positive set. One thousand randomly selected genomic regions and 500 active promoter regions were used as the negative set. Enhancers were predicted using a False Discovery Rate (FDR) cutoff of 0.05. Predicted enhancers that overlapped by at least 500 bp were merged by selecting the enhancer with the highest CSI-ANN score. We obtained histone modification ChIP-Seq data from the NCBI Epigenome Atlas, Roadmap Epigenomics Project, Encyclopedia of DNA Elements (ENCODE), International Human Epigenome Consortium and the GEO database (Table 6).

Target promoter(s) of an enhancer were predicted using the IM-PET (He et al., 2014) algorithm. It predicts enhancer-promoter interactions by integrating four features derived from transcriptome, epigenome, and genome sequence data, including: 1) enhancer-promoter activity correlation, 2) transcription factor-promoter co-expression, 3)

enhancer-promoter co-evolution, and 4) enhancer-promoter distance. Here, we used tissue/cell type specific histone modification ChIP-Seq and RNA-Seq data (Table 6) to compute values of features 1, 3 and 4 for the given tissue/cell type. Values of feature 3 were based on sequence conservation across 15 mammalian species (human, chimp, gorilla, orangutan, gibbon, rhesus, baboon, marmoset, tarsier, mouse lemur, tree shrew, mouse, rat, rabbit, and guinea pig). We used FDR cutoff 0.05 as the threshold for making predictions.

4.4.8 P-value for eSNPs that disrupt transcription factor binding sites

For each eSNP, we first scan sequences containing the eSNP using TF binding motifs from Cis-BP database (Weirauch et al., 2014) and calculate the log-odds ratio score for the SNP-containing sequence. If at least one allele for the SNP has a score greater than the threshold that corresponds to a p value $4x10^{-7}$, which is computed using TFM-Pvalue method (Touzet and Varre, 2007) for each motif separately, the sequence is considered as a TF binding site.

Next, the difference in the motif score between the two alleles is computed and compared to a null distribution of motif score differences using 1 million randomly selected SNPs reported by the 1000 Genomes project. Raw p-value is corrected for multiple testing using the Benjamini-Hochberg method. The motif disruption score for a given eSNP is the negative logarithm of the most significant motif disruption p-value among all TF motifs having a binding site overlapping with the eSNP.

4.4.9 Processing of Gene Expression Profiling Data

Gene expression microarray data were analyzed using the limma package (Ritchie et al., 2015). Raw microarray data was background corrected and quantile normalized. Linear model was fit to the data using the lmFit function of limma. Differential expression was assessed at probe level using the empirical Bayes (eBayes) method. To summarize differential expression at gene level, we selected the minimum p value across the probes that match to a gene. The list of datasets used in this study is provided in Table 8.

4.4.10 Gold standard risk variants located in gene promoters

The Human Gene Mutation Database (HGMD, version 2014 r1) (Stenson et al., 2009) was used to select regulatory variants located in promoter region which was defined as 2 kb upstream and 0.5 kb downstream of TSS. Transcript annotation was based on GENCODE v19 (GRCh37). Only transcripts with high confidence were used (level <3). We selected all diseases and their associated SNPs in HGMD that satisfied the following three criteria: 1) SNPs have annotation of "DP" (disease-associated polymorphism) or "DFP" (disease-associated polymorphism with additional supporting functional evidence) in HGMD. 2) case and control gene expression data was available for the disease; 3) genes of the reported promoter were present in the HumanNet connected network. For negative control SNPs, we first randomly selected common (minor allele frequency ≥1%) SNPs from the 1000 Genomes Project. Because HGMD variants are not distributed randomly across the genome and 75% lie within a 2 kb window around an annotated transcription start sites, to control for this bias, our second negative control set was selected such that distance distribution to the nearest TSS

matches that of the positive set. The lists of positive (HGMD) and control variants are provided in Table 3.

4.4.11 Gold standard risk variants located in enhancers

We curated a set of experimentally validated eSNPs from multiple resources, including HGMD, Open Regulatory Annotation Database (OregAnno) (Griffith et al., 2008), and manual search of PubMed literature. We accepted an eSNP as being validated only if it satisfies the following criteria: 1) significance association of the eSNP with the disease; 2) there is direct experimental evidence that the GWAS SNP causes differential TF binding and gene expression change; and 3) the enhancer is located more than 5Kbp away from the affected gene promoter. The list of experimentally validated eSNPs is provided in Table 4.

4.4.12 SNPs associated with autoimmune diseases

We obtained SNPs associated with seven autoimmune diseases (SLE, PSO, RA, T1D, CRH, MS, ULC) from the GWAS Catalog (Welter et al., 2014). All SNPs have a genome-wide association p-value of $5x10^{-8}$ or less. We identified SNPs in the same linkage disequilibrium with the GWAS catalog SNPs.

4.4.13 Identification of optimal set of candidate eSNPs in a disease

ARVIN computes a probability score for each candidate eSNP. In order to choose a cutoff for final predictions, we developed the following procedure based on the assumption that a true risk eSNP should either be a lead or linked to a lead GWAS SNP. We first rank all eSNPs in descending order of their ARVIN scores. Next, we compute a cumulative enrichment score as following:

$$S = \sum_{i=1}^{n} \begin{cases} d * p_i \\ d * (1 - p_i) \end{cases} d = 1 \; if \; i \; in \; disease \; associate \; region - 1 \; otherwise$$

where p_i is the ARVIN score for eSNP i and d is an indicator function whose value depends on whether the SNP is located in a disease associated region, which is defined as the LD block anchored by a GWAS or ImmunoChIP (Trynka et al., 2011) lead SNP with an association p-value < $5x10^{-8}$. Based on this scoring scheme, eSNPs located outside of disease-associated regions contributes negative value to the enrichment score. When S reaches the maximum value, we use the index i as the optimal number of eSNPs for a given disease.

4.4.14 Evaluation of enhancer-promoter predictions using Hi-C and ChIA-PET data

We searched for large-scale chromatin interaction data measured using either Hi-C or ChIA-PET protocol. We used the reported enhancer-promoter (EP) interactions in these studies as the gold standard to assess the quality of our predicted enhancer-promoter pairs. We first identified EP pairs in which the enhancers overlap with the interacting fragments reported by Hi-C or ChIA-PET studies. Those EP pairs are regarded as eligible for comparison with Hi-C or ChIA-PET data. We then computed the Receiver Operating Characteristic curves using either Hi-C or ChIA-PET data.

4.4.15 Identifying the subnetwork affected by a set of risked eSNPs using the Prize Collecting Steiner Tree algorithm

To identify the subnetwork collectively affected by a set of risk eSNPs in a disease, we use the Prize Collecting Steiner Tree (PCST) algorithm. Given an undirected graph $G = (V, E, c, p)$ where vertices V are associated with non-negative profits p and edges E are associated with non-negative costs c. The PCST algorithm finds a connected

subgraph $\boldsymbol{G'} = (\boldsymbol{V'}, \boldsymbol{E'})$ of $\boldsymbol{G}$ that maximizes the net profit which is defined as the sum of all node-associated profits minus all edge-associated costs (I. Ljubić, 2006). The algorithm takes as the input the disease-relevant regulatory network and all risk eSNPs implicated in a given disease. Every input eSNP is considered as a possible root node of the Steiner tree but the one resulting in a Steiner tree with the largest profit chosen as the final root node. To identify the optimal solution, the algorithm will link every input eSNP to the selected root node maximizing the net profit. This can be solved using message-passing technique (Bailly-Bechet et al., 2011). We convert our edge score into edge cost by $\mathbf{1} - \boldsymbol{S}(\boldsymbol{i}, \boldsymbol{j})$, where $\boldsymbol{S}(\boldsymbol{i}, \boldsymbol{j})$ is the edge score. The final output of the algorithm is a tree composed of all risk eSNPs and genes that are targeted by the eSNPs. The eSNPs are connected via interactions among the target genes.

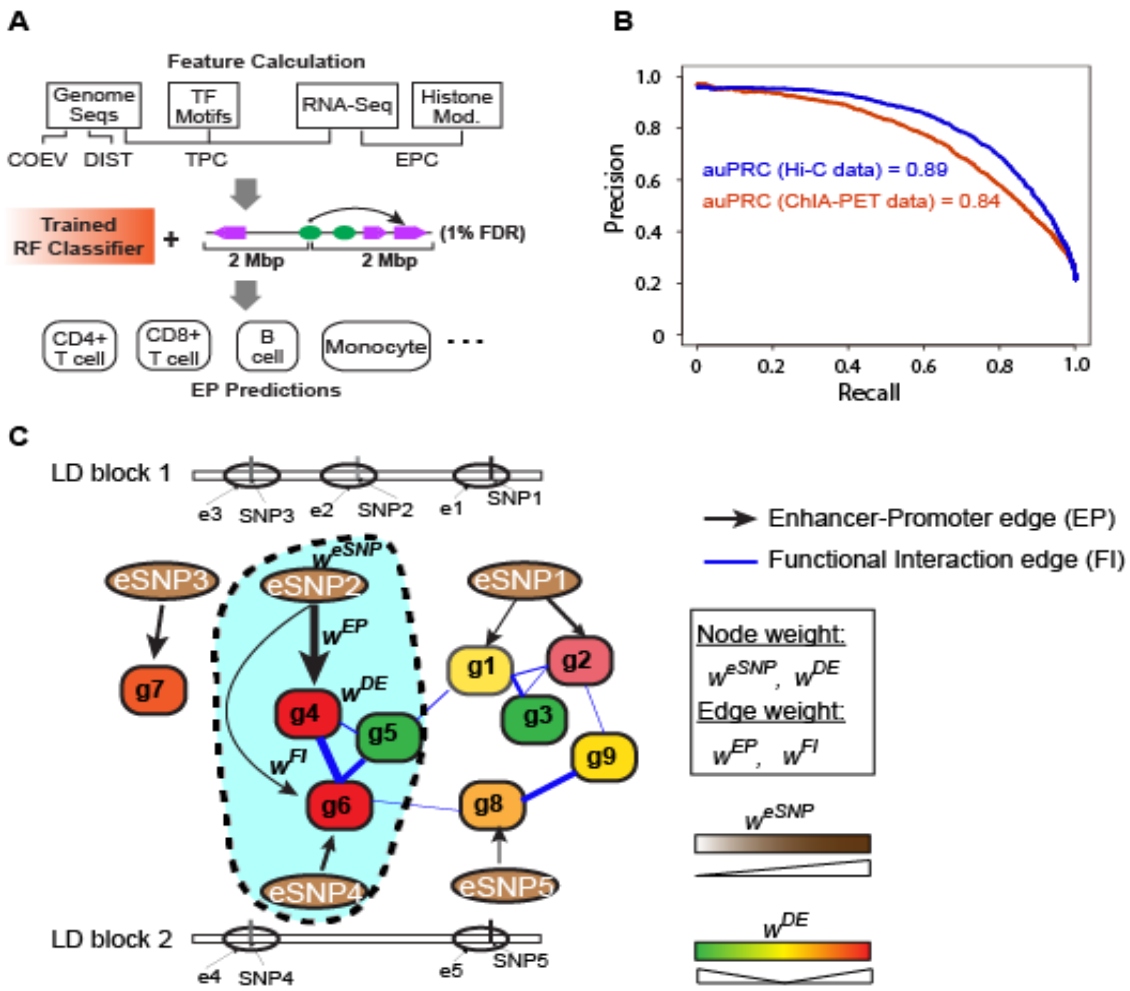

Figure 29. Construction of weighted and disease-relevant regulatory network for prioritizing risk SNPs located in regulatory DNA sequences. A) Schematic for an integrated, disease-relevant regulatory network. The network involves SNP-containing enhancers and their target genes and functional interactions among the target genes. Such a network can be constructed using transcriptomic and epigenomic data on cells/tissues relevant to the disease under study. The encircled subnetwork (highlighted in cyan) represents pathways targeted by candidate risk eSNPs. LD, linkage disequilibrium; e, enhancer; g, gene. EP, enhancer-promoter interaction; FI, functional gene interaction; eSNP, enhancer SNP; W^{eSNP}, weights for eSNPs; W^{DE}, weights for differential gene expression; W^{EP}, weights for EP edges; W^{PPI}, weights for FI edges. B) Schematic for the enhancer target prediction algorithm, IM-PET. Features used by the random forest (RF) classifier are: COEV, coevolution of enhancer and target promoter; DIST, distance constraint between enhancer and target promoter; TPC, transcription factor and target promoter correlation; EPC, enhancer and target promoter profile correlation. FDR, false discovery rate. C) Performance evaluation of IM-PET using Hi-C and ChIA-PET data. Hi-C data are from GM12878, K562, IMR90, HMEC, NHEK, HUVEC, and HELA. ChIA-PET data are from GM12878, K562, MCF7, HELA, $CD34^+$ HSPCs, and $CD4^+$ T cells.

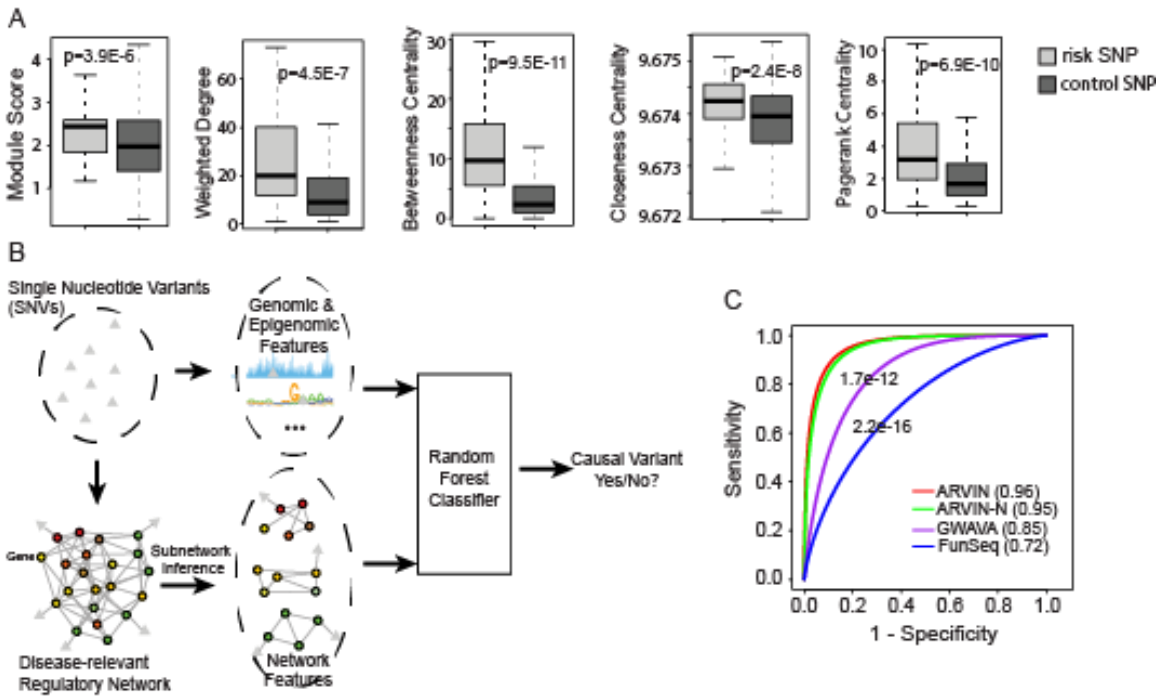

Figure 30. ARVIN combines both genomic and network features to prioritize risk SNPs. A) Topological features extracted from an integrated regulatory network are discriminative. P-values are based on t-test. B) Overview of the ARVIN method. C) Receiver Operating Characteristic (ROC) curves using known risk SNPs located in gene promoters. Values in parenthesis are area under the ROC curve. P-values are based on Z-test. ARVIN-N, ARVIN using only network-based features.

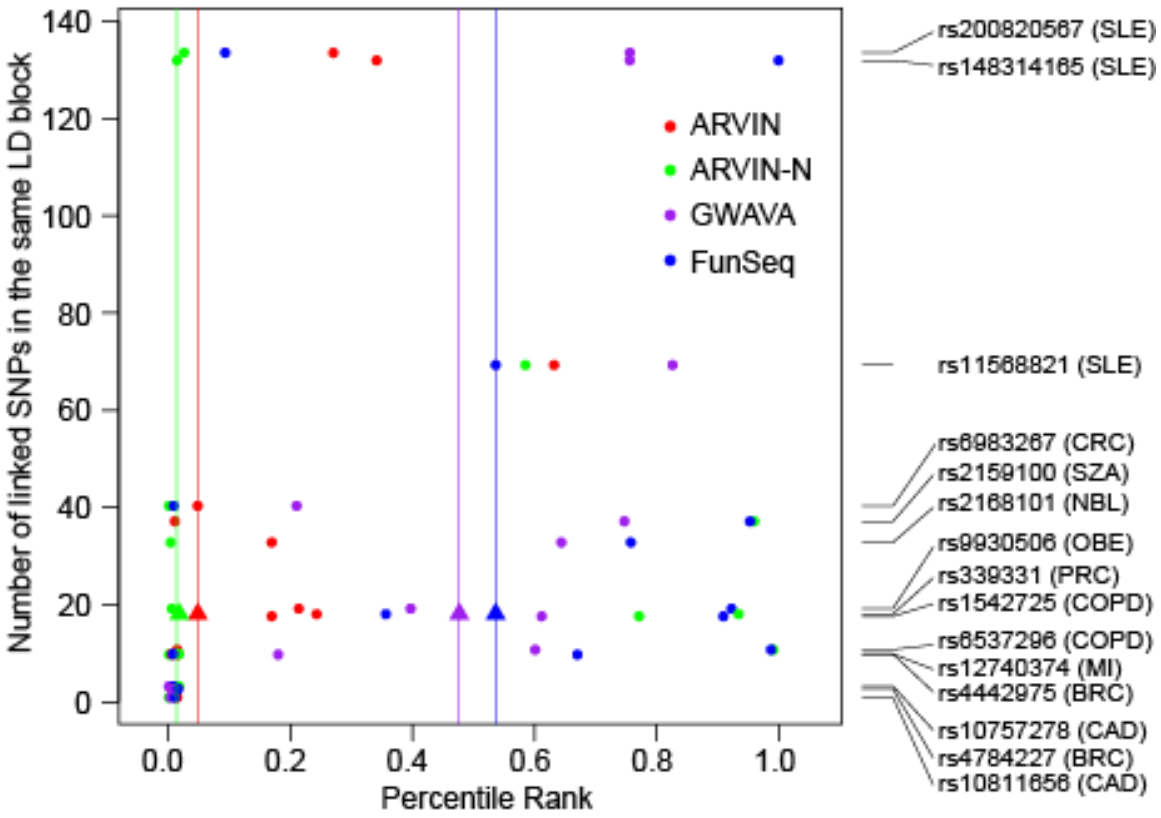

Figure 31. Performance benchmarking using known risk SNPs located in enhancers. References for validated risk enhancer SNPs are provided in Table 4. Each gold standard risk SNP was ranked against all other SNPs in the same linkage equilibrium block as the gold standard SNP. Percentile rank is shown on the X-axis with 1 being the last ranked. Filled circle, rank of an individual SNP by a given method. Filled triangle, medium rank of the full set of gold standard SNPs by a given method. SNP IDs and associated diseases are shown on the right. SLE, systemic lupus erythematosus; PSO, psoriasis; CRC, colorectal cancer; PRC, prostate cancer; RA, rheumatoid arthritis; OBE, obesity; MI, myocardial infarction; BRC, breast cancer; COPD, chronic obstructive pulmonary disease; SZA, schizophrenia; CAD, coronary artery disease; NBL, neuroblastoma.

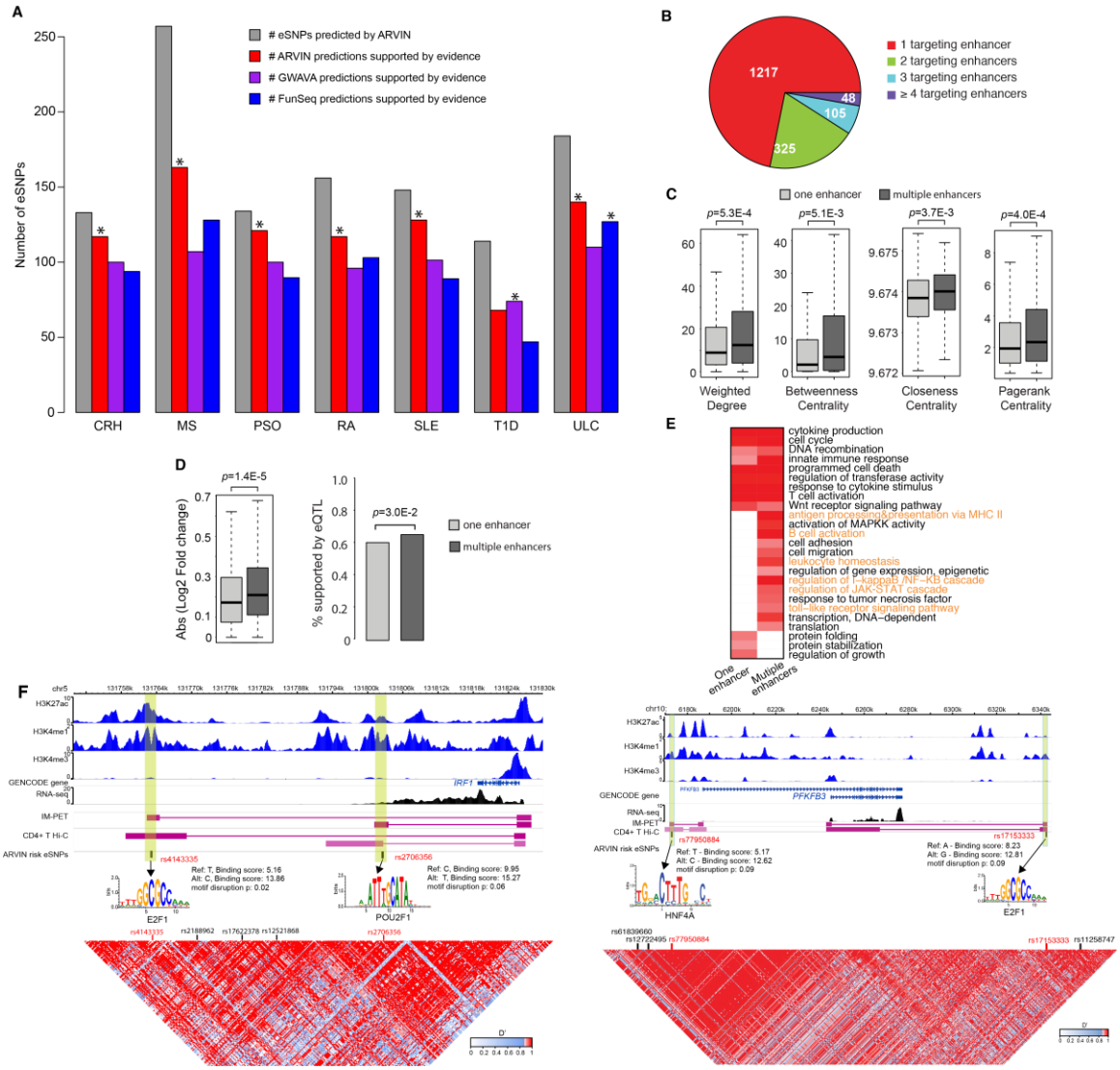

Figure 32. Predicted risk enhancer SNPs associated with seven autoimmune diseases. A) Number of predicted risk eSNPs in each disease and overlap with supporting evidence. For comparison purpose, the prediction thresholds of GWAVA and FunSeq were set to give the same number of predictions as ARVIN. Statistical significance of overlap between predictions and supporting evidence was computed using hypergeometric distribution. *, p-value < 0.05. SLE, systemic lupus erythematosus; PSO, psoriasis; RA, rheumatoid arthritis; T1D, type 1 diabetes; CRH, Crohn's disease; ULC, ulcerative colitis; MS, multiple sclerosis. B) Number of genes that are targeted by different numbers of eSNP-containing enhancers. A considerable fraction of genes is targeted by multiple enhancers, suggesting combinatorial regulation of affected genes by multiple risk eSNPs. Genes targeted by multiple risk eSNPs have higher values of network topological features (C), higher expression fold changes between case and control samples (D), higher overlap with eQTLs (D), and more enriched GO terms for immune responses (highlighted in orange) (E). F) Examples of genes targeted by two risk eSNPs. Left, *IRF1*. Right, *PFKFB3*. IM-PET, enhancer-promoter interactions predicted

by IM-PET. CD4+ T Hi-C, enhancer-promoter interactions detected by Capture Hi-C data. Annotation for autoimmune disease-associated loci is based on ImmunoBase.

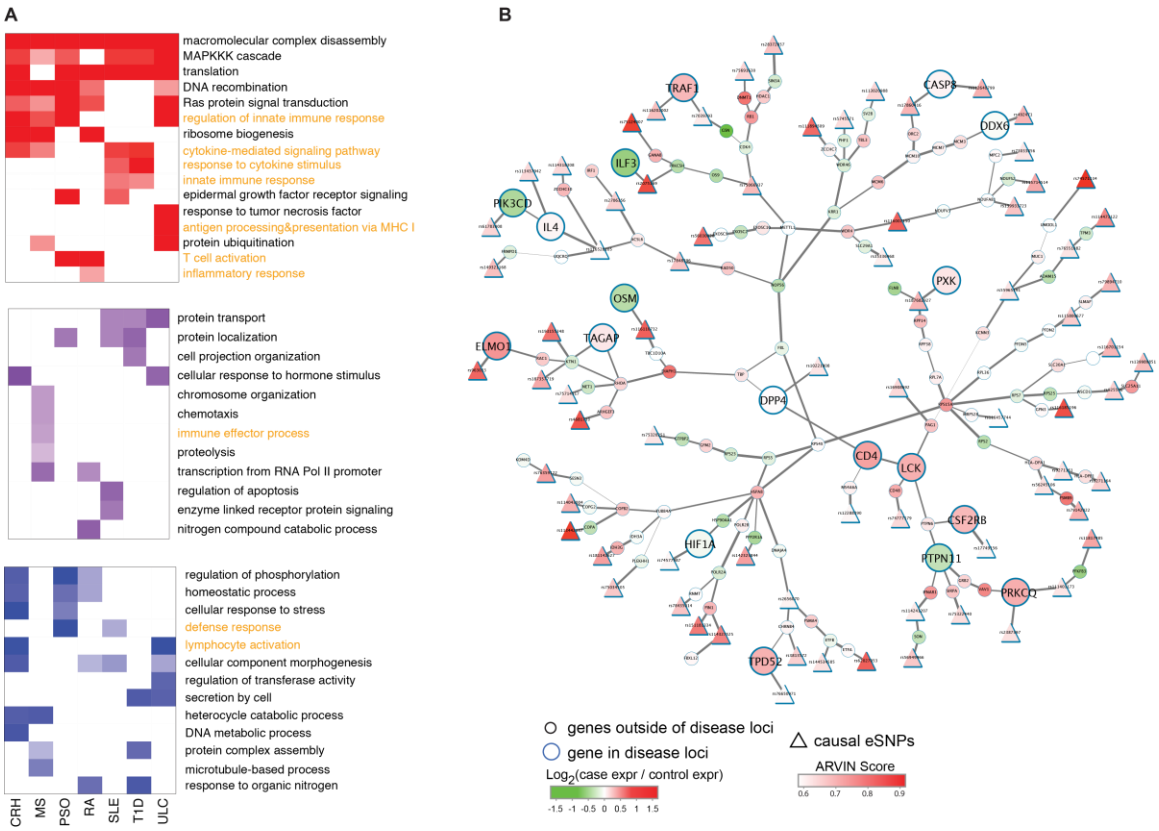

Figure 33. Gene subnetwork collectively perturbed by all risk eSNPs in a disease. A) Uniquely enriched GO terms among genes of perturbed subnetworks predicted by ARVIN (red), GWAVA (purple) and FunSeq (blue), respectively. GO terms for immune responses are highlighted in orange. B) An example of perturbed subnetwork by all risk eSNPs in rheumatoid arthritis. Circle, genes. Node size represents location of a gene relative to disease-associated loci; bigger node, within a disease-associated locus, smaller node, outside a disease-associated locus; Node color represents differential gene expression between case and control samples. Triangle, predicted risk eSNPs.

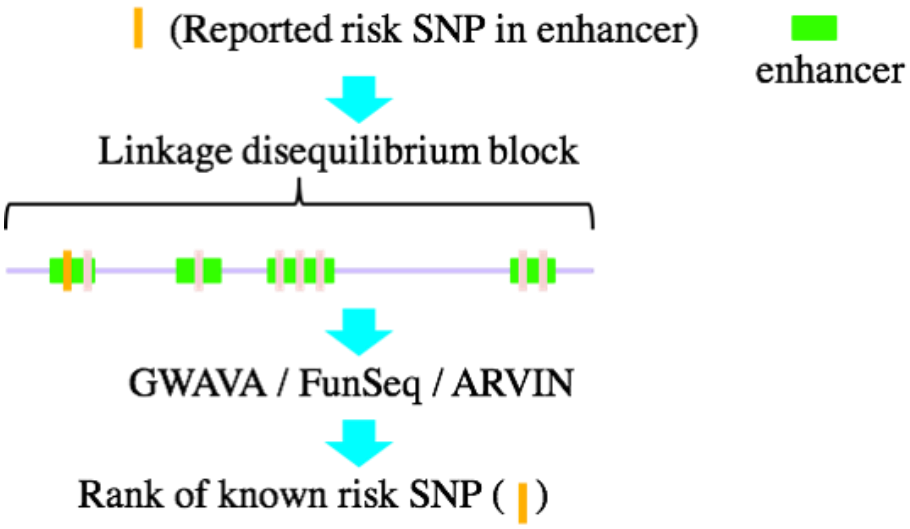

Figure 34. Workflow for evaluating risk eSNPs.

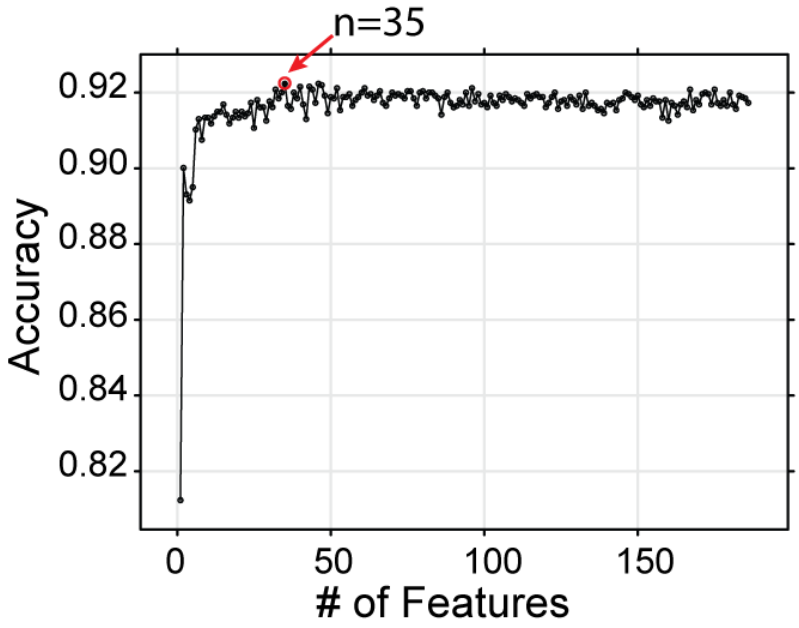

Figure 35. Recursive feature elimination. The optimal set contains 35 features.

Table 3. List of gold standard risk SNPs located in gene promoters. AD, Alzheimer's disease; ASD, Autism spectrum disorder; AST, asthma; BLC, bladder cancer; CAD, coronary artery disease; CF, cystic fibrosis; COPD, chronic obstructive pulmonary disease; CRC, colorectal cancer; HC, hypercholesterolaemia; MI, myocardial infarction; OBE, obesity; PD, Parkinson's disease; PRC, prostate cancer; PSO, psoriasis; RA, rheumatoid arthritis; SLE, systemic lupus erythematosus; SZA, schizophrenia; T1D, Type 1 diabetes; T2D, Type 2 diabetes; TSB, thalassemia beta.

chrom	HGMD_ACC	RS_id	Alt	Ref	HGNC symbol	Disease	Nearest_transcript	Distance_to_nearest_TSS	Mutation_type
chr6	CR080767	rs3748079	T	C	ITPR3	SLE	ENST00000374316	4	DP
chr1	CR083996	rs3093061	G	A	CRP	SLE	ENST00000255030	602	DP
chr2	CR0911347	rs13385731	T	C	RASGRP3	SLE	ENST00000494927	157	DP
chr12	CR0911356	rs1385374	G	A	SLC15A4	SLE	ENST00000366292	220	DP
chr6	CR095246	rs762624	C	A	CDKN1A	SLE	ENST00000448526	1	DFP
chr11	CR095443	rs360719	T	C	IL18	SLE	ENST00000280357	1308	DFP
chr7	CR109943	N/A	A	C	IL6	SLE	ENST00000426291	225	DFP
chr15	CR1210014	rs34933034	G	A	CSK	SLE	ENST00000567571	1412	DP
chr17	CR1110626	N/A	G	A	ZNF750	PSO	ENST00000572562	179	DM
chr17	CR1110627	N/A	G	A	ZNF750	PSO	ENST00000572562	18	DM
chr17	CR1110628	N/A	C	T	ZNF750	PSO	ENST00000572562	46	DM
chr17	CR1110629	N/A	G	A	ZNF750	PSO	ENST00000572562	47	DM
chr20	CR123852	N/A	C	A	RNF114	PSO	ENST00000244061	68	DM
chr20	CR123853	N/A	C	A	RNF114	PSO	ENST00000244061	70	DM
chr20	CR123854	N/A	C	T	RNF114	PSO	ENST00000244061	45	DM
chr20	CR123855	N/A	A	C	RNF114	PSO	ENST00000244061	13	DM
chr1	CR053508	rs3811021	C	T	PTPN22	RA	ENST00000261441	1564	DP
chr5	CR057602	rs7702919	A	G	HAVCR1	RA	ENST00000522693	460	DP
chr19	CR087182	rs251864	A	G	ZFP36	RA	ENST00000597629	161	DFP
chr13	CR106322	rs7984870	G	C	TNFSF11	RA	ENST00000239849	1808	DFP
chr6	CR1111594	rs805297	C	A	APOM	RA	ENST00000375916	643	DFP
chr19	CR045993	rs2241712	A	G	TGFB1	COPD	ENST00000413014	206	DP
chr14	CR061338	rs17751769	C	T	SERPINA1	COPD	ENST00000553327	295	DP
chr4	CR002153	rs1799723	A	G	CCKAR	SZA	ENST00000295589	90	DP
chr6	CR004775	N/A	G	A	TFAP2A	SZA	ENST00000498450	73	DP
chr22	CR014437	rs165596	A	G	SNAP29	SZA	ENST00000572273	595	DP
chr1	CR070418	rs6691378	T	C	CHI3L1	SZA	ENST00000255409	1244	DP
chr17	CR077670	rs408067	G	C	SRR	SZA	ENST00000575840	9	DFP
chr2	CR078170	rs3749034	A	G	GAD1	SZA	ENST00000344257	52	DP
chr2	CR085115	rs3791878	T	G	GAD1	SZA	ENST00000445006	434	DP
chr8	CR098746	rs7825588	A	G	nrg1smdf	SZA	ENST00000520502	837	DFP
chr2	CR102886	N/A	G	T	NRXN1	SZA	ENST00000405472	159	N/A

Table 3 - Continued

chr2	CR102887	rs200865985	T	C	NRXN1	SZA	ENST00000405472	59	N/A
chr8	CR109398	rs208747	T	A	PCM1	SZA	ENST00000325083	1485	DP
chr3	CR1111322	rs3755557	A	T	GSK3B	SZA	ENST00000264235	1692	DFP
chr3	CR117967	rs2239547	A	G	ITIH4	SZA	ENST00000471505	148	DP
chr3	CR117970	rs736408	C	T	ITIH3	SZA	ENST00000465314	867	DP
chr6	CR1212004	rs4141761	C	T	LYRM4	SZA	ENST00000463032	67	DFP
chr6	CR1212005	rs7752203	G	C	LYRM4	SZA	ENST00000463032	71	DFP
chr11	CR129186	N/A	C	G	NRGN	SZA	ENST00000412681	35	DM
chr11	CR129190	N/A	G	A	NRGN	SZA	ENST00000412681	2	DM
chr11	CR129201	N/A	G	A	NRGN	SZA	ENST00000284292	106	DM
chr11	CR129203	N/A	C	G	NRGN	SZA	ENST00000284292	340	DM
chr11	CR129204	N/A	A	G	NRGN	SZA	ENST00000284292	382	DM
chr1	CR025422	rs41303970	C	T	GCLM	MI	ENST00000370238	342	DFP
chr6	CR073540	rs2781666	G	T	ARG1	MI	ENST00000368087	726	DP
chr1	CR083997	rs1325920	C	T	ENO1	MI	ENST00000489867	533	DP
chr12	CR093431	rs11066001	A	G	BRAP	MI	ENST00000539060	337	DFP
chr10	CR128695	N/A	G	C	SIRT1	MI	ENST00000212015	525	DM
chr10	CR128697	N/A	G	C	SIRT1	MI	ENST00000212015	193	DM
chr10	CR128698	N/A	G	T	SIRT1	MI	ENST00000212015	75	DM
chr10	CR128699	rs35706870	A	C	SIRT1	MI	ENST00000212015	811	DM
chr20	CR973336	N/A	C	A	THBD	MI	ENST00000377103	64	DM
chr9	CR020829	rs1800976	G	C	ABCA1	CAD	ENST00000374736	190	DP
chr15	CR025894	N/A	T	C	LIPC	CAD	ENST00000299022	395	DFP
chr21	CR096274	rs1378577	T	G	ABCG1	CAD	ENST00000398457	135	DFP
chr22	CR097872	rs72558180	C	T	MKL1	CAD	ENST00000466278	167	DFP
chr2	CR139457	rs11685424	A	G	IL1RL1	CAD	ENST00000393393	982	DFP
chr9	CR139459	rs7025417	C	T	IL33	CAD	ENST00000381434	1599	DFP
chr11	HR971651	N/A	G	A	APOA1	CAD	ENST00000375320	4	DM
chr16	CR014433	rs5030981	T	C	AGRP	OBE	ENST00000290953	37	DFP
chr2	CR051277	rs6546511	G	A	GFPT1	OBE	ENST00000357308	912	DP
chr22	CR054252	rs133068	C	G	MCHR1	OBE	ENST00000249016	348	DP
chr4	CR065648	rs6857530	A	G	NPY2R	OBE	ENST00000329476	628	DP
chr4	CR071282	N/A	C	T	NPY2R	OBE	ENST00000329476	331	DM

Table 3 - Continued

chr15	CR071287	rs16964465	C	A	SCG3	OBE	ENST00000220478	1204	DFP
chr9	CR073539	rs2989924	A	G	AQP7	OBE	ENST00000541274	826	DFP
chr5	CR993027	rs1801704	T	C	ADRB2	OBE	ENST00000305988	218	DP
chr7	CR015138	rs185028612	A	G	CFTR	CF	ENST00000446805	329	DM
chr7	CR015139	N/A	C	T	CFTR	CF	ENST00000446805	295	DM
chr7	CR120402	N/A	C	G	CFTR	CF	ENST00000003084	46	DM
chr7	CR1213098	N/A	G	C	CFTR	CF	ENST00000003084	118	DM
chr7	CR132467	N/A	T	C	CFTR	CF	ENST00000446805	116	DM
chr7	CR132468	N/A	A	G	CFTR	CF	ENST00000446805	518	DM
chr7	CR136679	N/A	T	C	CFTR	CF	ENST00000446805	47	DM
chr7	CR136680	N/A	A	T	CFTR	CF	ENST00000446805	188	DM
chr7	CR136681	rs139688774	G	C	CFTR	CF	ENST00000003084	157	DM
chr7	CR136682	N/A	G	C	CFTR	CF	ENST00000003084	21	DM
chr7	CR962504	N/A	C	G	CFTR	CF	ENST00000426809	2	DM
chr7	CR040864	rs1851426	C	T	CYP3A4	PRC	ENST00000336411	1047	DP
chr19	CR073549	rs266882	G	A	KLK3	PRC	ENST00000601503	159	DP
chr8	CR106743	rs11781886	A	G	NKX3-1	PRC	ENST00000523261	14	DFP
chr19	CR119475	rs3787016	C	T	POLR2E	PRC	ENST00000586215	873	DP
chr19	CR119966	rs2659056	A	G	KLK15	PRC	ENST00000598673	501	DP
chrX	CR973135	N/A	G	T	AR	PRC	ENST00000374690	602	DM
chrX	CR973136	N/A	C	A	AR	PRC	ENST00000374690	390	DM
chr4	CR045670	rs3755910	A	C	TDO2	ASD	ENST00000509738	246	DP
chrX	CR084749	rs5989681	C	G	ASMT	ASD	ENST00000381241	3	DFP
chrX	CR094444	rs73457060	G	A	NLGN4X	ASD	ENST00000381093	334	DM
chr3	CR015013	rs56198082	A	G	MLH1	CRC	ENST00000457004	3	DM
chr2	CR020339	rs138068023	G	C	MSH2	CRC	ENST00000233146	3	DM
chr3	CR064470	rs35032294	C	G	MLH1	CRC	ENST00000322716	26	DM
chr2	CR076695	rs62626348	G	T	ERBB4	CRC	ENST00000342788	470	DFP
chr3	CR078287	N/A	C	G	MLH1	CRC	ENST00000457004	76	DM
chr3	CR114539	N/A	G	T	MLH1	CRC	ENST00000457004	33	DM
chr3	CR116619	N/A	C	A	MLH1	CRC	ENST00000457004	4	DM
chr11	CR136167	N/A	G	C	PICALM	CRC	ENST00000534412	59	DM
chr11	CR004576	N/A	C	T	HBB	TSB	ENST00000335295	8	DM

Table 3 - Continued

chr11	CR034843	N/A	C	T	HBB	TSB	ENST00000485743	84	DM
chr11	CR040152	N/A	G	C	HBB	TSB	ENST00000335295	45	DM
chr11	CR075246	N/A	C	A	HBB	TSB	ENST00000485743	100	DM
chr11	CR075247	N/A	G	A	HBD	TSB	ENST00000292901	79	DM
chr11	CR076704	N/A	A	T	HBB	TSB	ENST00000485743	71	DM
chr11	CR076705	rs34500389	C	T	HBB	TSB	ENST00000485743	30	DM
chr11	CR076706	rs63750400	G	C	HBB	TSB	ENST00000485743	23	DM
chr11	CR077840	N/A	A	G	HBB	TSB	ENST00000335295	1	DM
chr11	CR082014	rs63750681	G	C	HBB	TSB	ENST00000485743	54	DM
chr11	CR082015	N/A	G	A	HBB	TSB	ENST00000485743	188	DM
chr11	CR096076	rs72561473	G	A	HBB	TSB	ENST00000485743	81	DM
chr11	CR097470	rs281864524	G	A	HBB	TSB	ENST00000485743	48	DM
chr11	CR100816	N/A	A	C	HBB	TSB	ENST00000485743	39	DM
chr11	CR102102	rs113115948	C	T	HBB	TSB	ENST00000335295	39	DM
chr11	CR106856	N/A	C	G	HBB	TSB	ENST00000485743	40	DM
chr11	CR112101	rs281864525	A	C	HBB	TSB	ENST00000485743	24	DM
chr11	CR118399	rs281864518	C	T	HBB	TSB	ENST00000485743	69	DM
chr11	CR119545	rs33981098	A	C	HBB	TSB	ENST00000485743	29	DM
chr11	CR125830	rs63751043	C	G	HBB	TSB	ENST00000485743	91	DM
chr11	CR138585	N/A	A	T	HBB	TSB	ENST00000335295	1	DM
chr11	CR139343	N/A	A	C	HBB	TSB	ENST00000485743	9	DM
chr11	CR139346	N/A	A	G	HBB	TSB	ENST00000485743	9	DM
chr11	CR139347	N/A	G	T	HBB	TSB	ENST00000485743	53	DM
chr11	CR830008	rs33931746	A	C	HBB	TSB	ENST00000485743	26	DM
chr11	CR860022	rs33981098	A	G	HBB	TSB	ENST00000485743	29	DM
chr11	CR890140	rs63751208	C	T	HBB	TSB	ENST00000485743	99	DM
chr11	CR890183	rs33980857	T	C	HBB	TSB	ENST00000485743	28	DM
chr11	CR910468	rs33941377	C	T	HBB	TSB	ENST00000485743	85	DM
chr11	CR910597	rs34704828	G	A	HBB	TSB	ENST00000335295	22	DM
chr11	CR920786	rs34500389	C	A	HBB	TSB	ENST00000485743	30	DM
chr11	CR920787	rs33994806	C	A	HBB	TSB	ENST00000485743	84	DM
chr11	CR920788	rs33994806	C	G	HBB	TSB	ENST00000485743	84	DM
chr11	CR920789	rs33941377	C	A	HBB	TSB	ENST00000485743	85	DM

Table 3 - Continued

chr11	CR920790	rs33944208	C	A	HBB	TSB	ENST00000485743	86	DM
chr11	CR920791	rs34999973	C	T	HBB	TSB	ENST00000485743	88	DM
chr11	CR930875	rs34883338	C	T	HBB	TSB	ENST00000485743	90	DM
chr11	CR961734	rs34135787	C	G	HBB	TSB	ENST00000335295	33	DM
chr11	CR994362	N/A	A	G	HBB	TSB	ENST00000485743	25	DM
chr11	CR994659	N/A	A	T	HBB	TSB	ENST00000485743	25	DM
chr11	HR0602	N/A	C	A	HBB	TSB	ENST00000485743	74	DM
chr19	CR021774	N/A	C	T	LDLR	HC	ENST00000558518	44	DM
chr19	CR042572	N/A	A	C	LDLR	HC	ENST00000558013	47	DM
chr19	CR042573	N/A	C	T	LDLR	HC	ENST00000557933	34	DM
chr19	CR045713	N/A	A	C	LDLR	HC	ENST00000558518	25	DM
chr19	CR045714	N/A	C	T	LDLR	HC	ENST00000558518	30	DM
chr19	CR055624	N/A	C	T	LDLR	HC	ENST00000558518	2	DM
chr19	CR055625	N/A	C	T	LDLR	HC	ENST00000558013	65	DM
chr19	CR055626	N/A	C	A	LDLR	HC	ENST00000558518	47	DM
chr19	CR075255	N/A	C	G	LDLR	HC	ENST00000558518	47	DM
chr1	CR084009	N/A	C	A	PCSK9	HC	ENST00000302118	42	DM
chr19	CR091488	N/A	C	T	LDLR	HC	ENST00000557933	48	DM
chr7	CR104943	rs17655652	A	G	NPC1L1	HC	ENST00000289547	76	DFP
chr19	CR108072	N/A	A	G	LDLR	HC	ENST00000558518	81	DM
chr19	CR116859	N/A	C	G	LDLR	HC	ENST00000558518	50	DM
chr19	CR127029	N/A	C	A	LDLR	HC	ENST00000558518	37	DM
chr19	CR127546	N/A	A	G	LDLR	HC	ENST00000558518	29	DM
chr19	CR920794	N/A	C	G	LDLR	HC	ENST00000557933	49	DM
chr19	CR941557	N/A	C	T	LDLR	HC	ENST00000558518	50	DM
chr19	CR973644	N/A	C	A	LDLR	HC	ENST00000558518	40	DM
chr19	CR992256	N/A	C	T	LDLR	HC	ENST00000558518	34	DM
chr2	CR092630	N/A	C	T	NR4A2	PD	ENST00000539077	51	DM
chr1	CR095375	N/A	C	G	PARK7	PD	ENST00000377491	36	DM
chrX	CR115755	N/A	C	G	GLA	PD	ENST00000316594	180	DM
chr16	CR116475	rs2270363	A	G	HMOX2	PD	ENST00000571291	7	DP
chr10	CR125662	N/A	C	G	SIRT1	PD	ENST00000212015	395	DM
chrX	CR128643	N/A	A	C	LAMP2	PD	ENST00000434600	857	DM

Table 3 - Continued

chr3	CR130993	rs76708041	G	A	ATG7	PD	ENST00000451513	547	DM
chr3	CR130994	N/A	T	C	ATG7	PD	ENST00000451513	185	DM
chr3	CR130995	N/A	G	A	ATG7	PD	ENST00000451513	83	DM
chr3	CR130996	rs77630528	G	A	ATG7	PD	ENST00000435760	6	DM
chr2	CR131136	N/A	C	G	NR4A2	PD	ENST00000409108	23	DM
chr6	CR132051	rs187978668	T	A	ATG5	PD	ENST00000369076	792	DM
chr14	CR000233	rs34086577	C	G	PSEN1	AD	ENST00000557356	228	DM
chr14	CR015066	rs1800839	C	T	PSEN1	AD	ENST00000557356	4	DP
chr3	CR056131	rs373602047	G	C	MME	AD	ENST00000491597	182	DP
chr21	CR062110	rs139885956	C	T	APP	AD	ENST00000448388	117	DM
chr21	CR062111	N/A	C	G	APP	AD	ENST00000448388	7	DM
chr21	CR062112	rs113926273	G	A	APP	AD	ENST00000448388	1388	DM
chr21	CR062113	N/A	G	A	APP	AD	ENST00000448388	9	DM
chr21	CR062114	N/A	C	A	APP	AD	ENST00000359726	64	DM
chr21	CR062115	rs187510057	G	A	APP	AD	ENST00000448388	172	DM
chr17	CR0911103	rs3744456	G	C	MAPT	AD	ENST00000446361	135	DFP
chr19	CR098250	N/A	T	C	PIN1	AD	ENST00000592184	1100	DM
chr1	CR099909	rs10752637	G	T	NCSTN	AD	ENST00000368069	774	DFP
chr11	CR100155	rs4938369	G	A	BACE1	AD	ENST00000528053	1144	DFP
chr1	CR1010177	rs117525971	G	T	IL6R	AD	ENST00000368485	94	DP
chr14	CR1010202	rs8003602	T	C	CYP46A1	AD	ENST00000554611	1681	DFP
chr14	CR1010203	rs3783320	A	G	CYP46A1	AD	ENST00000554611	1184	DFP
chr2	CR103125	rs3755166	C	T	LRP2	AD	ENST00000263816	685	DFP
chr1	CR116216	rs3754048	C	G	APH1A	AD	ENST00000493092	208	DFP
chr14	CR045986	rs8004654	T	C	PTGDR	AST	ENST00000306051	448	DFP
chr14	CR045987	rs803010	C	T	PTGDR	AST	ENST00000306051	340	DFP
chr6	CR057907	rs2071590	A	G	LTA	AST	ENST00000454783	64	DFP
chr5	CR077663	rs9313422	G	C	HAVCR1	AST	ENST00000518745	38	DFP
chr1	CR087758	rs10494132	T	C	CHIA	AST	ENST00000422815	1262	DFP
chr1	CR087759	rs3806448	G	A	CHIA	AST	ENST00000422815	1280	DFP
chr1	CR0911367	rs3806325	C	T	HLX	AST	ENST00000366903	92	DFP
chr1	CR0911368	rs2184658	C	G	HLX	AST	ENST00000366903	757	DFP
chr5	CR093694	rs2287774	A	G	SPINK5	AST	ENST00000398454	176	DFP

Table 3 - Continued									
chr1	CR104961	rs2038366	G	T	S1PR1	AST	ENST00000305352	1547	DFP
chr1	CR104962	rs59317557	C	G	S1PR1	AST	ENST00000305352	522	DFP
chr17	CR1111070	rs3091318	C	T	CCL7	AST	ENST00000378569	1313	DFP
chr5	CR112607	rs146866105	C	T	ITK	AST	ENST00000422843	45	DFP
chr7	CR118121	rs37973	A	G	GLCCI1	AST	ENST00000223145	550	DFP
chr5	CR119830	rs62382271	C	T	PPARGC1B	AST	ENST00000360453	383	DFP
chr14	CR123113	rs3751464	C	T	FRMD6	AST	ENST00000395718	685	DP
chr1	CR129592	rs10399931	G	A	CHI3L1	AST	ENST00000255409	202	DFP
chr5	CR136349	N/A	T	A	IL12B	AST	ENST00000231228	121	DFP
chr10	CR138579	rs6585018	A	G	PDCD4	AST	ENST00000280154	355	DFP
chr1	CR067511	rs2488457	G	C	PTPN22	T1D	ENST00000359785	986	DP
chr6	CR103219	N/A	C	G	CTGF	T1D	ENST00000367976	21	DFP
chr12	CR991538	N/A	G	C	HNF1A	T1D	ENST00000543427	7	DM
chr11	CR101139	N/A	C	A	INS	T1D	ENST00000381330	40	DM
chr11	CR101140	N/A	A	C	INS	T1D	ENST00000397262	16	DM
chr11	CR101141	N/A	A	G	INS	T1D	ENST00000481781	1134	DM
chr11	CR101142	N/A	C	G	INS	T1D	ENST00000381330	39	DM
chr4	CR011066	rs10011540	A	C	UCP1	T2D	ENST00000262999	36	DFP
chr12	CR031013	N/A	T	G	IAPP	T2D	ENST00000240652	198	DP
chr20	CR040573	rs2071023	C	G	PCK1	T2D	ENST00000467047	203	DFP
chr3	CR052437	rs5394	T	C	SLC2A2	T2D	ENST00000314251	353	DP
chr3	CR052438	rs5393	C	A	SLC2A2	T2D	ENST00000314251	380	DP
chr4	CR055621	rs2278862	A	G	BTC	T2D	ENST00000395743	136	DFP
chr11	CR067840	N/A	A	G	INS	T2D	ENST00000381330	53	DM
chr2	CR068525	rs6720415	T	C	GFPT1	T2D	ENST00000361060	220	DFP
chr11	CR140936	rs11603334	T	C	ARAP1	T2D	ENST00000426523	89	DFP
chr11	CR140942	rs1552224	G	T	ARAP1	T2D	ENST00000426523	24	DP
chr5	CR993027	rs1801704	T	C	ADRB2	T2D	ENST00000305988	218	DP
chr12	CR067513	rs201929640	C	G	serca2b	CRC	ENST00000539276	98	DM
chr12	CR067514	N/A	G	T	serca2b	CRC	ENST00000552636	151	DM
chr12	CR067515	N/A	G	T	serca2b	CRC	ENST00000552636	5	DM
chr1	CR050013	N/A	T	G	PTGS2	BLC	ENST00000367468	641	DP
chr3	CR072322	rs125701	A	G	OGG1	BLC	ENST00000302003	1151	DP

chr12	CR082025	rs937282	C	G	MDM2	BLC	ENST00000462284	160	DFP
chr16	CR102187	rs6498486	A	C	ERCC4	BLC	ENST00000575156	349	DFP
chr17	CR128492	rs9299	A	G	HOXB5	BLC	ENST00000498678	1795	DP

Table 4. List of known risk SNPs located in transcriptional enhancers. Pubmed ID is provided for the validation study. LD, linkage disequilibrium. BRC, breast cancer; NBL, neuroblastoma.

Disease	SNP ID	Target Gene(s)	PMID of Literature Reference	# eSNPs in the same LD block
Systemic lupus erythematosus	rs11568821	PDCD1	12402038	62
	rs148314165 rs200820567	TNFAIP3	24039598	168
Colorectal cancer	rs6983267	cMYC	19561604	41
Obesity	rs9930506	IRX3	24646999	15
Myocardial Infarction	rs12740374	SORT1	20686566	2
Breast cancer	rs4442975	IGFBP5	25248036	6
	rs4784227	TOX3	23001124	1
Chronic obstructive pulmonary disease	rs6537296	HHIP	22140090	10
	rs1542725			13
Prostate cancer	rs339331	RFX6	24390282	17
Schizophrenia	rs2159100	CACNA1C	25453756	33
Coronary artery disease	rs10811656	CDKN2B CDKN2BA IFNA21 MTAP	21307941	1
	rs10757278			3
Neuroblastoma	rs2168101	LMO1	23348506	25

Table 5. Number of NHGRI GWAS Catalog SNPs associated with autoimmune diseases and enhancer SNPs (eSNPs) in the same LD blocks with the GWAS Catalog lead SNPs.

Disease	# GWAS lead SNPs ($p<5x10^{-8}$)	# eSNPs in the same LD blocks with GWAS lead SNPs
Systemic lupus erythematosus	82	6,853
Psoriasis	58	3,520
Rheumatoid arthritis	143	10,346
Type 1 Diabetes	66	5,092
Crohn's Disease	273	15,015
Ulcerative Colitis	158	9,814
Multiple Sclerosis	82	5,966

Table 6. Summary of data sources used for constructing tissue/cell type specific enhancer-promoter networks. GSM*, accession IDs for NCBI GEO (Gene Expression Omnibus) database; E-MTAB*, accession IDs for EBI (European Bioinformatics Institute) Array Express database; GTEx*, Genotype-Tissue Expression Project tissue IDs used in Supplemental Table 8; HCAEC*, accession IDs for International Human Epigenome Consortium data portal; ERX*, accession IDs for European Nucleotide Archive; Reference 1, supplemental data portal of the study by Farh et al.

Tissue/Cell	Related Disease	Input	H3K4me1	H3K4me3	H3K27ac	RNA-Seq
CD4+ CD25+ CD127- (T_{reg})	Systemic lupus erythematous, Psoriasis, Rheumatoid arthritis, Type 1 diabetes, Crohn's disease, Ulcerative colitis, Multiple sclerosis	GSM772914	GSM772973	GSM772944	GSM997233	PMID: 25363779 [1]
CD4+ CD25- IL17+ (Th_{17})		GSM772988	GSM772985	GSM772986	GSM772987	PMID: 25363779 [1]
CD4+ CD25- IL17- (Th_{stim})		GSM772904 GSM1112784	GSM772902 GSM997268	GSM916071 GSM997232	GSM772905 GSM997266	PMID: 25363779 [1]
CD8+ (T_{mem})		GSM772874	GSM772873	GSM772967	GSM772880	E-MTAB-2319
CD4+ CD45RA+ Naïve T		GSM772916 GSM772876	GSM772860 GSM772869	GSM772836 GSM772948	GSM772835 GSM772934	GSM669617 GSM669583
CD4+ Memory T		GSM772881 GSM772930	GSM772884 GSM772924	GSM772790 GSM772925	GSM772963 GSM772997	GSM669584 GSM669618
B cell (GM12878)		GSM733742	GSM733772	GSM733708	GSM733771	GSM2072356 GSM2072357
CD14+ Monocyte		GSM1102807	GSM1102793	GSM1102797	GSM1102782	GSM1435495
Colonic mucosa	Colorectal cancer	GSM621669	GSM621670	GSM621671	GSM1112802	GTEx1
Rectal mucosa		GSM621647 GSM621677	GSM621639 GSM621659	GSM621643 GSM621658	GSM1112795 GSM1112801	
Sigmoid colon		GSM1059456	GSM956020	GSM956024	GSM915331	GTEx2
Adipose	Obesity	GSM621401	GSM621425	GSM621435	GSM916066	GTEx3
Skeletal muscle		GSM621680	GSM621686	GSM621685	GSM916064	GTEx4
Lung	Chronic obstructive pulmonary disease	GSM906417	GSM910572	GSM915336	GSM906395	GTEx5

Table 6 - Continued						
Substantia nigra	Schizophrenia	GSM772864	GSM772898	GSM772901	GSM1112778	GTEx6
Anterior caudate		GSM669978 GSM772826	GSM669970 GSM772830	GSM670031 GSM772829	GSM1112811 GSM772832	GTEx7
Hippocampus		GSM669971 GSM773019 GSM916037	GSM669962 GSM773021 GSM916039	GSM670022 GSM773022 GSM916040	GSM1112791 GSM773020 GSM916035	GTEx8
Mid frontal lobe		GSM669960 GSM773010	GSM670015 GSM773014	GSM670016 GSM773012	GSM1112810 GSM773015	GTEx9
Left ventricle	Myocardial Infarction	GSM908968	GSM906404	GSM906406	GSM908951	GTEx10
Endothelial cell	Coronary artery disease	HCAEC-03121 HCAEC-03193 HCAEC-03228	HCAEC-03121 HCAEC-03193 HCAEC-03228	HCAEC-03121 HCAEC-03193 HCAEC-03228	HCAEC-03121 HCAEC-03193 HCAEC-03228	HCAEC-03121 HCAEC-03193 HCAEC-03228
Mammary epithelial (HMEC)	Breast cancer	GSM733668	GSM733705	GSM733712	GSM733660	GSM765397
LNCAP (Prostate adenocarcinoma)	Prostate cancer	GSM686947	GSM686928	GSM686935	GSM686937	GSM1902621 GSM1902622 GSM1902623
SHSY5Y (Bone marrow neuroblast)	Neuroblastoma	GSM1532410	GSM1532409	NA	GSM1532408	ERX583734

Table 7. GTEx identifiers for RNA-Seq samples used in constructing enhancer-promoter networks.

Tissue Type and ID	Sample ID
(1) Colon transverse (Colonic mucosa, Rectal mucosa)	NFK9-2026-SM-3LK5K, O5YT-1426-SM-3MJHC, O5YW-1426-SM-3MJHF, OHPM-1426-SM-3TW8Y, OIZH-1426-SM-3NB1O, aOXRK-1726-SM-3NB16, OXRL-1426-SM-3NM9E, P4PP-1426-SM-3NM9L, P4QT-1426-SM-3NMCX, P78B-1726-SM-3P5ZV, PLZ5-1126-SM-3P613, PLZ6-0926-SM-3P5ZQ, PWCY-1026-SM-48TD4, PWN1-1426-SM-48TDF, PWOO-1326-SM-48TCJ, PX3G-1426-SM-48U1J, Q2AH-1226-SM-48TZL, Q2AI-0926-SM-48U1F, Q734-1126-SM-48TZY, QCQG-1626-SM-48U26, QDVJ-1326-SM-48U1X, QDVN-1326-SM-48TZ3, QLQW-0426-SM-447A7, QMRM-1226-SM-447C6, R53T-1326-SM-48FCQ, R55C-1126-SM-48FCJ, R55D-1826-SM-48FEF, R55G-1226-SM-48FDC, RM2N-0926-SM-48FD1, RU1J-1326-SM-46MUL, RWS6-1826-SM-47JXX, S341-1426-SM-4AD6U, S3XE-1126-SM-4AD4N, S4P3-1226-SM-4AD4Y, S4Q7-0826-SM-4AD5E, S4UY-0826-SM-4AD4Z, S7SF-1926-SM-4AT5B, SE5C-1526-SM-4BRWU, SNMC-1126-SM-4DM5M, SNOS-1226-SM-4DM5H, T5JW-1126-SM-4DM5V, T6MO-0826-SM-4DM51, TKQ1-0626-SM-4DXTS, TKQ2-1326-SM-4DXT9, TML8-1326-SM-4DXTO, U3ZH-1326-SM-4DXSF, U3ZM-1126-SM-4DXUB, U3ZN-2126-SM-4DXU1, U4B1-1126-SM-4DXV3, UJHI-1126-SM-4IHLN, UJMC-1326-SM-4IHLS, UPIC-1726-SM-4IHKG, V955-1626-SM-4JBHJ, VJYA-2026-SM-4KL1K, W5WG-2426-SM-4LMI6, WEY5-1426-SM-4LMJ3, WFG7-1526-SM-4LVMG, WFON-1426-SM-4LVMT, WH7G-1326-SM-4LVMS, X5EB-0926-SM-46MVT, XAJ8-0426-SM-47JYJ, XBED-1526-SM-4AT5W, XGQ4-1226-SM-4AT67, XMK1-1726-SM-4B64Z, XPVG-1826-SM-4B64X, XQ8I-2126-SM-4BOOM, XUJ4-2026-SM-4BOOW, XUW1-1926-SM-4BOP1, XUZC-1326-SM-4BRV2, XV7Q-2126-SM-4BRVX, XXEK-1926-SM-4BRVD, XYKS-2226-SM-4E3IU
(2) Sigmoid colon	V955-1726-SM-4JBHF, VJYA-2126-SM-4KL1O, W5WG-2026-SM-4LMIB, WFG7-1626-SM-4LVMF, WFG8-1726-SM-4LVM6, WFON-1326-SM-4LVMN, X4EO-2726-SM-4E3HS, X5EB-1126-SM-46MVV, XAJ8-0726-SM-47JY5, XBED-1726-SM-47JYO, XQ8I-2326-SM-4BOQC, XUJ4-2126-SM-4BOOX, XUW1-1826-SM-4BOQD, XUZC-1526-SM-4BRV4, XV7Q-2226-SM-4BRVY, XXEK-1826-SM-4BRVC

Table 7 - Continued

(3) Adipose	N7MS-0326-SM-4E3K2, NFK9-0326-SM-3MJGV, NPJ8-0226-SM-48TBN, O5YT-0226-SM-32PK5, O5YV-0226-SM-48TBY, OHPK-0226-SM-3MJH6, OHPM-0226-SM-3LK61, OHPN-0226-SM-48TBV, OIZF-0226-SM-2I5GR, OIZH-0226-SM-2YUMH, OIZI-0226-SM-2XCEE, OOBJ-0226-SM-2YUMM, OOBK-0226-SM-2YUMF, OXRK-0326-SM-2YUMQ, OXRK-0326-SM-3NB3R, OXRL-0226-SM-3NB18, OXRN-0226-SM-2I5EJ, OXRO-0226-SM-3LK6F, OXRP-0226-SM-3NB14, OXRP-0226-SM-48TDH, P44G-2326-SM-2XCCZ, P44H-0326-SM-2XCES, P4PP-0226-SM-2S1NN, P4PQ-0226-SM-2S1NK, P4QR-0426-SM-2S1NV, P4QS-0226-SM-3NB1U, P4QT-0226-SM-3LK68, P78B-0226-SM-3NB1Z, PLZ4-0226-SM-2S1NW, PLZ5-1826-SM-3NB22, PLZ6-1326-SM-3NB24, POMQ-2326-SM-2S1O8, POYW-0726-SM-2XCEO, PSDG-0326-SM-48TCP, PWCY-1926-SM-3NB25, PWN1-0226-SM-2S1OZ, PWOO-2226-SM-2S1P1, PX3G-0226-SM-2S1OU, PX3G-0226-SM-3NB2C, Q2AG-0226-SM-2S1P4, Q2AH-1726-SM-2S1OT, Q2AH-1726-SM-3NB2B, Q2AI-1426-SM-2S1P5, Q734-1826-SM-2I3EL, QCQG-1826-SM-2S1P2, QDT8-0226-SM-32PL4, QDVJ-1826-SM-2S1P3, QDVN-2126-SM-2I3FR, QDVN-2126-SM-33HBS, QEG4-0326-SM-2S1OS, QEG5-0326-SM-2S1PB, QEL4-0326-SM-3GAE5, QESD-1526-SM-2S1QT, QLQ7-1526-SM-2S1QA, QLQW-1226-SM-2S1Q9, QMRM-1726-SM-2S1QG, QV31-1326-SM-2S1QE, QV44-1825-SM-447CF, QVJO-0226-SM-2S1R2, R53T-1626-SM-3GAEW, R55C-1626-SM-48FEG, R55D-0326-SM-48FES, R55F-1426-SM-2TF53, R55G-2426-SM-2TC5I, REY6-0326-SM-2TF5A, RM2N-1726-SM-2TF55, RTLS-0226-SM-2TF5E, RU72-1026-SM-46MUG, RVPU-2326-SM-2TF6R, RWSA-0226-SM-2XCBA, S32W-2226-SM-2XCAY, S33H-1126-SM-2XCB6, S341-1626-SM-3K2B8, S3XE-1626-SM-3K2AJ, S4P3-1526-SM-3K2AV, S4Q7-1626-SM-3K2AE, S4UY-0226-SM-3K2AP, S4Z8-1726-SM-3K2AX, S7SE-0226-SM-2XCD4, S7SF-1826-SM-3K2AD, S95S-1326-SM-2XCDK, SIU8-0226-SM-2XCDS, SJXC-0226-SM-2XCDU, SN8G-0226-SM-4DM6B, SNMC-1326-SM-2XCFK, SNOS-1426-SM-32PLY, SSA3-0226-SM-32QPN, SUCS-1826-SM-32PM1, T2IS-0226-SM-32QPH, T5JC-0526-SM-32PM7, T5JW-1726-SM-3GADN, T6MN-0226-SM-32PMD, T6MO-1726-SM-33HB8, T8EM-1126-SM-3DB7D, TKQ1-1126-SM-4GIAZ, TMKS-0226-SM-3DB7X, TML8-2026-SM-32QOP, TMMY-0326-SM-33HBF, TMZS-0226-SM-3DB9N, TSE9-0226-SM-3DB84, U3ZN-2626-SM-3DB7T, U412-0526-SM-3DB9I, U4B1-1726-SM-3DB9F, U8T8-0226-SM-3DB97, U8XE-0426-SM-3DB91, UJHI-1626-SM-3DB9A, UTHO-0426-SM-3GAED, VJYA-1326-SM-3GIJC, VUSG-2426-SM-4KKZG, W5X1-2626-SM-4LMI8, WEY5-1926-SM-3GIL8, WFG8-2326-SM-3GILF, WFJO-1926-SM-3GILA, WFON-2226-SM-3TW8W, WH7G-2226-SM-3NMBN, WI4N-1126-SM-3LK7Q, WL46-0326-SM-3LK6Y, WVLH-0226-SM-3MJG6, WY7C-2426-SM-3NB2V, X261-0226-SM-3NMD2, X4LF-1726-SM-3NMBZ, X4XY-0326-SM-46MVZ, X5EB-2426-SM-4E3HX, X62O-0226-SM-4E3JB, X638-0226-SM-47JZ9, X88G-0226-SM-4GIE4, X8HC-0226-SM-4E3K1, XAJ8-0926-SM-47JXZ, XBEC-0326-SM-4AT4M, XBED-2326-SM-47JYR, XGQ4-2226-SM-4AT4Y, XK95-0426-SM-4AT4R, XLM4-0226-SM-4AT4N, XMK1-2226-SM-4B673, XOT4-0226-SM-4B66Z, XOTO-0226-SM-4B66H, XPVG-2726-SM-4B66W, XQ8I-0526-SM-4BOPS, XUJ4-2526-SM-4BOO4, XUW1-0526-SM-4BOP3, XUYS-0226-SM-47JX1, XUZC-1826-SM-4BRVO, XV7Q-2626-SM-4BRVA, XXEK-2426-SM-4BRUS, XYKS-2726-SM-4E3IC U8XE-1926-SM-3DB98, UPK5-1826-SM-3GAEB, UTHO-1826-SM-3GAFE, VUSG-1426-SM-3GIJN, W5X1-1426-SM-3GIKH, WEY5-1326-SM-3GILS, WFJO-1026-SM-3GIKL, WFON-1226-SM-3TW8F, WH7G-1126-SM-3NMBK, WHPG-0626-SM-3NMBD, WHWD-0826-SM-3LK6R, WK11-2426-SM-3NMAA, WL46-2026-SM-3LK7U, WOFM-0726-SM-3MJF8, WRHK-0826-SM-3MJFG, WXYG-0926-SM-3NB2O, WY7C-0926-SM-3NB34, WYJK-1426-SM-3NM8V, X3Y1-0926-SM-3P5YT, XAJ8-0226-SM-4GIB2, XBED-1326-SM-4AT4F, XGQ4-1026-SM-4AT4L, XLM4-1026-SM-4AT51, XMK1-0926-SM-4B66X, XOT4-0826-SM-4B66Y, XQ8I-1426-SM-4BOPW, XUJ4-1726-SM-4BONW, XUZC-1126-SM-4BOPZ, XV7Q-1726-SM-4BRUU, XXEK-1426-SM-4BRW1, XYKS-1826-SM-4E3JV

Table 7 - Continued

(4) Skeletal muscle	N7MS-0426-SM-2YUN6, NFK9-0626-SM-2HMIV, NPJ8-1626-SM-2HMIY, O5YT-1626-SM-32PK6, O5YV-2026-SM-2D7VS, OHPJ-1626-SM-2HMKO, OHPK-1626-SM-2YUN3, OHPL-1626-SM-2HMIR, OHPM-1626-SM-2HMK4, OHPN-2726-SM-2I5H4, OIZF-1626-SM-2YUMI, OIZG-1326-SM-2HMIQ, OIZH-1626-SM-2HMKI, OIZI-0626-SM-2XCEH, OOBJ-1626-SM-2I3F7, OOBK-1626-SM-2HMKG, OXRK-1826-SM-2HMJE, OXRL-1626-SM-2YUMU, OXRN-1326-SM-3LK5V, OXRO-1726-SM-3LK6C, OXRP-2326-SM-2S1NL, P44G-0526-SM-2XCD1, P44H-0426-SM-2XCEZ, P4PP-1626-SM-2HMJF, P4PQ-1626-SM-2HMKK, P4QR-0726-SM-2I5GO, P4QS-1626-SM-2S1NH, P4QT-1626-SM-2S1NP, P78B-1626-SM-2S1O1, PLZ4-0926-SM-2S1OI, PLZ5-1726-SM-2I5F6, PLZ6-1526-SM-2S1OC, POMQ-1926-SM-3NB1Y, POYW-0526-SM-2XCEY, PSDG-0426-SM-2S1OF, PW2O-1726-SM-2S1OO, PWCY-2026-SM-2S1NF, PWN1-1626-SM-2S1OL, PWOO-2326-SM-2S1PQ, PX3G-1626-SM-2S1PT, Q2AG-0426-SM-2S1PU, Q2AH-1826-SM-2S1Q2, Q2AI-1526-SM-3GIJ3, Q734-2026-SM-3GADA, QCQG-2126-SM-2S1P8, QDT8-0526-SM-32PL5, QDT8-0526-SM-3NMD8, QDT8-0526-SM-4LMJR, QDVJ-1926-SM-2S1PJ, QDVN-2426-SM-2S1Q4, QEG4-0626-SM-2S1OY, QEG5-0426-SM-2I5GJ, QEL4-0626-SM-3GIJM, QESD-1626-SM-2S1RB, QLQ7-1726-SM-2S1QQ, QLQW-1326-SM-2S1QS, QV31-1426-SM-2S1QD, QV44-2026-SM-2S1RD, QVJO-0126-SM-3GIK4, QVUS-0226-SM-3GIJY, QXCU-1726-SM-2TC6L, R3RS-0526-SM-3GADG, R53T-1826-SM-3GIJX, R55C-1726-SM-3GADJ, R55D-0626-SM-3GAD5, R55E-0526-SM-2TC6B, R55F-1526-SM-2TF4X, R55G-2326-SM-2TC61, REY6-0826-SM-2TF4S, RM2N-1626-SM-2TF5N, RN64-0326-SM-2TC5J, RNOR-0526-SM-2TF4O, RTLS-0526-SM-2TF64, RU1J-1726-SM-2TF5S, RU72-1326-SM-2TF6T, RUSQ-1726-SM-2TF68, RVPU-2426-SM-2XCAR, RVPV-0926-SM-47JXU, RWS6-2126-SM-2XCAV, RWSA-0726-SM-2XCBE, S32W-2326-SM-2XCAW, S33H-2226-SM-2XCB7, S341-1826-SM-3K2AB, S3XE-2026-SM-3K2B5, S4P3-1626-SM-3K2AZ, S4Q7-1526-SM-3K2AG, S4UY-0526-SM-3K2AN, S4Z8-1826-SM-3K2BH, S7PM-0526-SM-3NM92, S7SF-2026-SM-3K2AS, S95S-1426-SM-2XCDM, SE5C-1826-SM-2XCE4, SIU7-1826-SM-2XCE2, SIU8-0526-SM-2XCDP, SJXC-0526-SM-2XCFG, SN8G-0326-SM-32PLG, SNMC-1426-SM-2XCFM, SNOS-1526-SM-32PLW, SSA3-0326-SM-32QPS, SSA3-0326-SM-47JWY, SUCS-1626-SM-32PLS, T2IS-2626-SM-32QPP, T2YK-0226-SM-4DM5O, T5JC-0626-SM-3NMA6, T5JW-1826-SM-3GAE1, T6MN-0526-SM-32PMS, T8EM-1326-SM-3DB7G, TKQ1-1426-SM-4GICK, TKQ2-0826-SM-33HB6, TMKS-0326-SM-3DB8A, TML8-1826-SM-32QOR, TMMY-0426-SM-33HBB, TMZS-0326-SM-3DB9P, TSE9-0526-SM-3DB7Z, U3ZG-0326-SM-47JXN, U3ZH-1926-SM-4DXTR, U3ZM-1226-SM-3DB9G, U3ZN-2226-SM-3DB88, U412-0326-SM-3DB9L, U4B1-1626-SM-3DB8N, U8T8-1426-SM-3DB9H, U8XE-0726-SM-3DB8O, UJHI-1726-SM-3DB9B, UJMC-1826-SM-3GADT, UPJH-0526-SM-4IHK8, UTHO-0726-SM-3GAEN, V1D1-2426-SM-3GAER, V955-2426-SM-3GAEF, VJWN-0426-SM-3GIJI, VJYA-1926-SM-3GIJJ, VUSG-2626-SM-4KKZI, VUSH-0326-SM-3NB2I, W5WG-1926-SM-4KKZK, WEY5-2026-SM-3GILE, WFG7-2226-SM-3GIKP, WFG8-2426-SM-3GILL, WFON-2326-SM-3LK7M, WHPG-2226-SM-3NMBO, WHSB-1826-SM-3TW8M, WK11-2526-SM-3NM9Y, WL46-0626-SM-3LK7R, WOFL-0626-SM-3MJG3, WOFM-1326-SM-3MJFR, WRHK-1626-SM-3MJFH, WRHU-0826-SM-3MJFN, WWYW-0526-SM-3NB2W, WXYG-2526-SM-3NB3F, WY7C-2526-SM-3NB2N, WYJK-1726-SM-3NM9U, WYVS-2326-SM-3NMAI, WZTO-0826-SM-3NM8Q, X261-0326-SM-3NMD4, X4EO-0526-SM-3P5Z3, X4XX-0626-SM-3NMC1, X4XY-0626-SM-4E3IN, X5EB-2326-SM-46MW5, X638-0326-SM-47JY1, X88G-0326-SM-47JZ4, X8HC-0526-SM-4E3JA, XAJ8-1026-SM-47JY9, XBED-2626-SM-4E3J5, XGQ4-2326-SM-4AT53, XOT4-0526-SM-4B66O, XOTO-0526-SM-4B662, XPT6-2026-SM-4B64V, XPVG-2526-SM-4B66D, XQ3S-0426-SM-4BOOA, XQ8I-0626-SM-4BOPT, XUJ4-2626-SM-4BOQ3, XUW1-0826-SM-4BOP6, XUYS-0326-SM-47JX2, XUZC-2126-SM-4BRW8, XV7Q-2926-SM-4BRUL, XYKS-2426-SM-4AT43

Table 7 - Continued

(5) Lung	N7MS-0926-SM-2HMIZ, N7MT-0126-SM-2D7VT, NFK9-1026-SM-2HMK1, NPJ8-0326-SM-2D7VV, O5YT-0526-SM-32PK8, O5YV-0526-SM-2D7VZ, O5YV-0526-SM-2I5GE, O5YW-0526-SM-2YUMX, OHPJ-1026-SM-2HMLE, OHPK-0526-SM-2HMJB, OHPL-0526-SM-2HMIX, OHPL-0526-SM-3NM8U, OHPM-0526-SM-2YUMJ, OIZF-0526-SM-2YUNO, OIZG-0526-SM-2HMLF, OIZH-0526-SM-2HMKV, OIZI-1026-SM-3NB1K, OOBJ-0526-SM-48TDK, OOBK-0526-SM-2HMJJ, OXRK-0926-SM-2HMKP, OXRL-0526-SM-2I3EZ, OXRN-0526-SM-2I5EN, OXRO-0326-SM-2I5EE, OXRO-0326-SM-33HBM, OXRP-0526-SM-2I3EW, P44H-1126-SM-48TBU, P4PP-0526-SM-2HMKE, P4PQ-0526-SM-2HMKR, P4QR-1026-SM-2I5GP, P4QS-0526-SM-2I3ET, P4QT-0526-SM-2I3EX, P78B-0926-SM-2I5FA, PLZ4-0726-SM-2TC6Q, PLZ5-0726-SM-2I5F9, PLZ6-0426-SM-2I5FG, POMQ-0526-SM-3GADD, POYW-1226-SM-2XCEP, PSDG-1126-SM-2S1ON, PVOW-1026-SM-2XCF9, PW2O-0526-SM-2I3DX, PWOO-0726-SM-2I3EB, PX3G-0526-SM-2I3EM, Q2AG-1026-SM-2HMJX, Q2AG-1026-SM-33HBW, Q2AH-0426-SM-2I3EP, Q2AI-0526-SM-2I3EJ, Q734-0626-SM-2I3EF, QCQG-0326-SM-2I3ES, QDT8-0926-SM-32PL2, QDVJ-0926-SM-2I5FU, QDVN-0726-SM-2I3FO, QDVN-0726-SM-4B64L, QEG4-0526-SM-48TZD, QEG5-1126-SM-2I5GH, QEG5-1126-SM-33HC2, QEL4-0826-SM-3GAF2, QESD-0626-SM-2I5G4, QMR6-1926-SM-32PL9, QMRM-0826-SM-2I5G7, QMRM-0826-SM-3NB33, QV44-0926-SM-2S1RH, QXCU-0626-SM-2TC69, R3RS-1026-SM-3GADF, R55C-0526-SM-3GIKA, R55D-0926-SM-3GAEU, R55G-0826-SM-2TC5U, REY6-0426-SM-2TF5G, RM2N-0426-SM-2TF4T, RN64-1226-SM-2TC6E, RNOR-0726-SM-2TF5I, RTLS-0926-SM-2TF5X, RU1J-0126-SM-2TF6Y, RU72-0526-SM-2TF5Z, RUSQ-0626-SM-2TF5V, RVPV-1726-SM-2TF5W, RWS6-0226-SM-2XCA9, RWSA-1126-SM-2XCAZ, S32W-0326-SM-2XCBI, S33H-0626-SM-2XCBJ, S341-0326-SM-2XCAU, S4Q7-0426-SM-3K2BJ, S4Z8-0426-SM-3K2AH, S7SE-0926-SM-2XCD6, S7SF-0426-SM-3K2B7, SE5C-0526-SM-2XCE1, SIU7-0526-SM-3NM8I, SIU8-0926-SM-2XCDO, SN8G-0926-SM-4DM5I, SNOS-0426-SM-32PMH, SUCS-0626-SM-32PM5, T2IS-0526-SM-32QP9, T5JC-0826-SM-32PMC, T6MN-0826-SM-32PM4, T6MO-0426-SM-32QOI, T8EM-0326-SM-3DB7F, TML8-0326-SM-4GICN, TMMY-0926-SM-33HBG, TSE9-0726-SM-3DB8C, U3ZH-0526-SM-3DB75, U3ZM-0426-SM-3DB73, U3ZN-0626-SM-3DB7U, U412-0826-SM-3DB9K, U8T8-2226-SM-3DB95, U8XE-1426-SM-3DB8Q, UJHI-0726-SM-3DB92, UJMC-0726-SM-3GADX, UPIC-0826-SM-3GADQ, UPK5-1126-SM-3GAEJ, V1D1-0826-SM-3P5ZA, VJYA-0326-SM-3GAEX, VUSG-0926-SM-3GIK6, W5X1-0526-SM-3GILH, WFG7-0526-SM-3GIKI, WFG8-0926-SM-3GIKJ, WFJO-0326-SM-3GIL3, WFON-0426-SM-3GIL4, WH7G-0726-SM-3NMBM, WHPG-1426-SM-3NMBB, WHSB-0326-SM-3LK6K, WK11-0526-SM-3NB3O, WOFM-0126-SM-3MJFE, WRHU-0226-SM-3MJFV, WWYW-0926-SM-3NB2Z, WY7C-0426-SM-3NB3C, WYBS-1126-SM-3NMAM, WYJK-0826-SM-3NM8Y, WYVS-0526-SM-3NM9W, WZTO-0426-SM-3NM99, X261-1026-SM-3NMDL, X3Y1-0626-SM-3P5YS, X4EO-0926-SM-3P5Z2, X4EP-0526-SM-3P5YW, X4LF-0526-SM-3NMB6, X4XY-1026-SM-46MVX, X585-1026-SM-46MW6, X5EB-0426-SM-46MVY, XBED-0826-SM-47JYC, XBEW-0226-SM-4AT6A, XGQ4-0826-SM-4AT4T, XOT4-1426-SM-4B65T, XPVG-1026-SM-4B64Y, XQ3S-0926-SM-4BOPI, XQ8I-1126-SM-4BOO2, XUJ4-1426-SM-4BONT, XV7Q-0426-SM-4BRVN, XXEK-0626-SM-4BRWE, XYKS-0526-SM-4BRW2
(6) Substantia nigra	N7MS-0011-R2a-SM-2HML6, N7MT-0011-R2a-SM-2I3GI, NL3H-0011-R2a-SM-2I3GG, NL4W-0011-R2a-SM-2I5GV, NPJ7-0011-R2a-SM-2I3GF, NPJ8-0011-R2a-SM-2TC6M, OHPN-0011-R2A-SM-2I5FB, OXRO-0011-R2A-SM-3NB1W, P44G-0011-R2A-SM-2XCD2, PWO3-0011-R2A-SM-2S1OX, Q2AG-0011-R2A-SM-2HMIT, QDT8-0011-R2A-SM-32PKQ, QMR6-0011-R2A-SM-32PKV, QVJO-0011-R2A-SM-2S1QK, RU72-0011-R2A-SM-2TF6O, RVPU-0011-R2A-SM-2XCAG, S7SE-0011-R2A-SM-2XCDC, T2IS-0011-R2A-SM-32QPF, T5JC-0011-R2A-SM-32PLZ, T6MN-0011-R2A-SM-32QOW, UTHO-0011-R2A-SM-3GIKC, WHSE-0011-R2A-SM-3P5ZL, WL46-0011-R2A-SM-3LK6O, WVLH-0011-R2A-SM-3MJFJ, X4EP-0011-R2B-SM-3P625, X4XX-0011-R2A-SM-3P623, X585-0011-R2B-SM-46MVF, XLM4-0011-R2B-SM-4AT5Z, XMD1-0011-R2B-SM-4AT5N

Table 7 - Continued	
(7) Anterior caudate	N7MS-0011-R5a-SM-2HMK8, N7MT-0011-R5a-SM-2I3G6, NL3H-0011-R5a-SM-2I3GB, NL4W-0011-R5a-SM-2I3GD, NPJ7-0011-R5a-SM-2I5GY, NPJ7-0011-R5a-SM-33HBK, NPJ8-0011-R5a-SM-2HMJY, OHPN-0011-R5A-SM-2I5FF, OXRN-0011-R5A-SM-2I5EF, OXRO-0011-R5A-SM-2I5EG, P44G-0011-R5A-SM-2I3FA, P44H-0011-R5A-SM-2XCEX, PVOW-0011-R5A-SM-32PL7, PWO3-0011-R5A-SM-2I5EZ, Q2AG-0011-R5A-SM-2HMJH, QDT8-0011-R5A-SM-32PKN, QMR6-0011-R5A-SM-32PKT, QVJO-0011-R5A-SM-2S1QM, R55E-0011-R5A-SM-2TC5N, R55F-0011-R5A-SM-2TF5L, RNOR-0011-R5A-SM-2TF4J, RU72-0011-R5A-SM-2TF6U, RVPU-0011-R5A-SM-2XCAD, RVPV-0011-R5A-SM-2TF69, S7PM-0011-R5A-SM-3NM8G, S7SE-0011-R5A-SM-2XCDA, T2IS-0011-R5A-SM-32QP4, T5JC-0011-R5A-SM-32PLK, T6MN-0011-R5A-SM-32QPD, TSE9-0011-R5A-SM-3DB7J, UTHO-0011-R5A-SM-3GIJD, WHSE-0011-R5A-SM-3P5ZO, WL46-0011-R5A-SM-3LK6V, WVLH-0011-R5A-SM-3MJFW, WWYW-0011-R5A-SM-3NB3E, WZTO-0011-R5B-SM-3NMC5, X261-0011-R5A-SM-3NMB4, X4EP-0011-R5A-SM-3P628, X4XX-0011-R5A-SM-46MWN, X585-0011-R5A-SM-46MVI, XMD1-0011-R5A-SM-4AT47, XOTO-0011-R5A-SM-4B657
(8) Hippocampus	N7MS-0011-R1a-SM-2HMJG, N7MT-0011-R1a-SM-2TC6G, NL3H-0011-R1a-SM-48TDJ, NPJ7-0011-R1a-SM-3GACT, NPJ8-0011-R1a-SM-2HMLC, NPJ8-0011-R1a-SM-33HCB, OHPN-0011-R1A-SM-2I5GB, P44G-0011-R1A-SM-2I3FE, P44H-0011-R1A-SM-3NM8J, PVOW-0011-R1A-SM-32PL6, PWO3-0011-R1A-SM-2I5EW, Q2AG-0011-R1A-SM-2HMJI, QDT8-0011-R1A-SM-32PKS, QMR6-0011-R1A-SM-32PKW, QVJO-0011-R1A-SM-2S1QI, QVUS-0011-R1A-SM-3GAD2, R55E-0011-R1A-SM-2TC6N, RNOR-0011-R1A-SM-2TF5D, RVPU-0011-R1A-SM-2XCAI, S7SE-0011-R1A-SM-2XCDE, T5JC-0011-R1A-SM-32PM6, T6MN-0011-R1A-SM-32QOY, TSE9-0011-R1A-SM-3DB7E, UTHO-0011-R1A-SM-3GIJO, WHSE-0011-R1A-SM-3P5ZK, WL46-0011-R1A-SM-3LK6M, WWYW-0011-R1A-SM-3TW8G, WZTO-0011-R1B-SM-3NMAR, X4EP-0011-R1A-SM-3P624, X4XX-0011-R1B-SM-3P622, X585-0011-R1B-SM-46MVE, XMD1-0011-R1A-SM-4AT4C, XOTO-0011-R1B-SM-4B65C
(9) Mid frontal lobe	N7MS-0011-R10A-SM-2HMJK, N7MT-0011-R10A-SM-2I3E1, NL3H-0011-R10A-SM-2I3E9, NL4W-0011-R10A-SM-2I3DY, NPJ7-0011-R10A-SM-2I3E5, NPJ8-0011-R10A-SM-2YUMO, OHPN-0011-R10A-SM-2I5FL, OHPN-0011-R10A-SM-33HBU, OXRN-0011-R10A-SM-2I5GC, OXRO-0011-R10A-SM-2I5EH, P44G-0011-R10A-SM-2I3FF, P44H-0011-R10A-SM-2XCEK, Q2AG-0011-R10A-SM-2HMLA, QDT8-0011-R10A-SM-32PKG, QMR6-0011-R10A-SM-32PKO, QVJO-0011-R10A-SM-2S1QJ, QVUS-0011-R10A-SM-3GIK3, RU72-0011-R10A-SM-2TF6D, RVPU-0011-R10A-SM-2XCAH, S7SE-0011-R10A-SM-2XCDF, T5JC-0011-R10A-SM-32PM2, T6MN-0011-R10A-SM-32QP7, TSE9-0011-R10A-SM-3DB7O, UTHO-0011-R10A-SM-3GIJQ, WL46-0011-R10A-SM-3MJFQ, WVLH-0011-R10A-SM-3MJFM, WWYW-0011-R10A-SM-3NB35, WZTO-0011-R10B-SM-4E3KB, X261-0011-R10B-SM-4E3JT, X4XX-0011-R10B-SM-46MWO, X4XY-0011-R10B-SM-46MWS, X585-0011-R10A-SM-46MUY, XLM4-0011-R10A-SM-4AT5P, XMD1-0011-R10A-SM-4AT4A

Table 7 - Continued	
(10) Left Ventricle	N7MS-0826-SM-2HML4, N7MT-1326-SM-2I3FV, NFK9-0926-SM-2HMJU, NPJ8-0426-SM-2HMK6, O5YT-0326-SM-32PKA, O5YV-0326-SM-2I5H2, O5YW-0326-SM-2I5EI, OHPJ-0926-SM-2HMJ1, OHPK-0326-SM-2HMJO, OHPL-0326-SM-2HMKS, OHPL-0326-SM-33HC8, OHPM-0326-SM-2HMKT, OHPM-0326-SM-33HCA, OIZF-0326-SM-2YUNE, OIZG-1126-SM-2HMIU, OIZH-0326-SM-2HMKC, OOBJ-0326-SM-2I3F8, OOBJ-0326-SM-33HBO, OOBK-3025-SM-48TBP, OXRK-0826-SM-2HMK7, OXRL-0326-SM-2I3F2, OXRO-2026-SM-2YUMZ, P44G-0826-SM-2I5ES, P44H-0726-SM-48TBT, P4PP-0326-SM-2HML9, P4PP-0326-SM-33HC4, P4PQ-0326-SM-2HMJ8, P4QS-0326-SM-2I3EU, P78B-0426-SM-2I5F5, PLZ5-0626-SM-2I5F8, POMQ-0326-SM-2I5FO, PSDG-0926-SM-2I5FP, PVOW-0426-SM-2XCF8, PWCY-0526-SM-2I3ER, PWOO-0526-SM-2S1Q3, PX3G-0326-SM-2I3EO, Q2AH-0526-SM-2I3ED, QDVJ-0426-SM-2I5FW, QDVN-0326-SM-2I3FS, QEG4-0426-SM-2I5GK, QEG4-0426-SM-33HC3, QEG5-0926-SM-2TC64, QEL4-0926-SM-3GAD1, QESD-0526-SM-2I5G5, QLQ7-0526-SM-2I5G3, QMRM-0526-SM-2I5GA, QV44-0526-SM-2S1RE, QVJO-1926-SM-2S1QZ, QXCU-0826-SM-2TC6F, R45C-0926-SM-3GAD4, R53T-0926-SM-3GADH, R55C-0326-SM-3GAF1, R55E-1026-SM-2TC5S, R55G-0526-SM-2TC5O, REY6-1026-SM-2TF4Y, RN64-0826-SM-2TC62, RNOR-0826-SM-2TF5C, RTLS-0826-SM-2TF5Q, RU72-0326-SM-2TF5T, RUSQ-0526-SM-2TF72, RWS6-0326-SM-2XCAP, RWSA-0626-SM-2XCBD, S32W-0626-SM-2XCBG, S33H-0526-SM-2XCBK, S3XE-0426-SM-3K2AC, S7SF-0526-SM-3K2BC, SE5C-0626-SM-2XCDV, SE5C-0626-SM-3P5ZJ, SE5C-0626-SM-4IHLJ, SIU7-0426-SM-2XCDX, SIU8-0826-SM-2XCDQ, SNMC-0126-SM-2XCFO, SUCS-0326-SM-32PLL, T2IS-0426-SM-32QPE, T6MN-0926-SM-32PLX, U3ZH-0326-SM-3DB7A, U3ZN-1426-SM-3DB87, U4B1-0326-SM-3DB8K, U8XE-1126-SM-3DB8W, UJHI-0426-SM-3DB8Y, UJMC-0526-SM-3GAE3, UPK5-0326-SM-3GAF3, V1D1-0526-SM-4JBGW, V955-0726-SM-3GAFG, WEY5-0426-SM-3GIKT, WFG7-0726-SM-3GIKO, WFG8-0626-SM-3GILJ, WFON-0326-SM-3GIKX, WH7G-0426-SM-3NMBJ, WHPG-0826-SM-3NMBF, WHSE-0926-SM-3NMBS, WHWD-0426-SM-3LK83, WI4N-0626-SM-3TW8Z, WL46-0926-SM-3LK7T, WQUQ-1426-SM-3MJFD, WRHU-1226-SM-4E3IJ, WWYW-1326-SM-3NB2S, WY7C-0526-SM-3NB3D, WYJK-1026-SM-3NM8W, WZTO-1326-SM-3NM8X, X3Y1-0426-SM-3P5Z4, X5EB-0826-SM-46MVS, X8HC-1626-SM-46MWE, XBEC-1326-SM-4AT69, XBED-0526-SM-47JY3, XGQ4-0326-SM-4GIEE, XPT6-0126-SM-4B65S, XPVG-0826-SM-4B654, XQ3S-0626-SM-4BOOB, XQ8I-0126-SM-4BOPL, XUJ4-0526-SM-4BOON, XV7Q-0826-SM-4BRV7, XXEK-0926-SM-4BRWH

Table 8. List of gene expression data used for accessing differential gene expression between case and control for diseases studied in this report.

Disease	Accession Number	Source Tissue/Cell Types
Alzheimer's Disease	GSE48350	Hippocampus
Asthma	GSE43696	Bronchial epithelial cells
Autism Spectrum Disorder	GSE7329	Lymphoblastoid cells
Bladder cancer	GSE3167	Bladder
Breast cancer	EGEOD-54002	Mammary gland cells
Chronic obstructive pulmonary disease	GSE47460	Whole lung homogenate
Colorectal cancer	GSE21510	Colon
Coronary artery disease	GSE20686	Whole blood
Crohn's disease	GSE20881	Sigmoid colon Terminal colon Ascending colon Descending colon
	GSE36807	Intestine
	GSE9686	Colon
Cystic Fibrosis	GSE15568	Epithelial cells
Hypercholesterolaemia	GSE6054	Monocytes
Multiple sclerosis	GSE21942	Peripheral blood mononuclear cells
	GSE43592	T cells
	GSE21942	Peripheral blood mononuclear cells
Myocardial infarction	GSE66360	CD146+ circulating endothelial
Neuroblastoma	E-MEXP-669	Fetal sympathetic neuroblasts
Obesity	GSE55200	Subcutaneous adipose tissue
Parkinson Disease	GSE7621	Substantia nigra
Prostate cancer	GSE55945	Prostate
Psoriasis	GSE13355	Skin
	GSE52471	
	GSE32407	
	GSE14905	
	GSE10500	
Rheumatoid arthritis	GSE12021	Synovial membrane
	GSE10500	Synovial macrophage
Schizophrenia	GSE25673	Hippocampus
Systemic lupus erythematous	GSE29536	Whole blood
	GSE13887	CD3+ T cells
	GSE10325	CD4+ T, CD19+ B, myeloid
	GSE46907	Monocytes

	GSE30153	Peripheral B cells
Thalassemia Beta	GSE62430	Erythroid progenitor cells
Type 1 diabetes	GSE9006	Peripheral blood mononuclear cells
	GSE55098	Peripheral blood mononuclear cells
Type 2 diabetes	GSE29226	subcutaneous adipose
Ulcerative colitis	GSE10191	Colon
	GSE36807	Intestine
	GSE9686	Colon

Table 8 - Continued

Table 9. List of eQTL tissue/cell types that are relevant to a given autoimmune disease and are used in this study.

Diseases	Tissue/Cell types
CRH	Colon_Transverse, Colon_Sigmoid, Small_Intestine_Terminal_Ileum, Stomach, Esophagus_Mucosa, Esophagus_Gastroesophageal_Junction, Whole_Blood, Cells_EBV-transformed_lymphocytes
MS	Brain_Anterior_cingulate_cortex_BA24, Brain_Cortex, Brain_Frontal_Cortex_BA9, Brain_Nucleus_accumbens_basal_ganglia, Brain_Hippocampus, Brain_Cerebellum, Brain_Cerebellar_Hemisphere, Brain_Putamen_basal_ganglia, Brain_Caudate_basal_ganglia, Brain_Hypothalamus, Cells_EBV-transformed_lymphocytes
PSO	Skin_Sun_Exposed_Lower_leg, Skin_Not_Sun_Exposed_Suprapubic, Whole_Blood, Cells_EBV-transformed_lymphocytes
RA	Whole_Blood, Cells_EBV-transformed_lymphocytes
SLE	Whole_Blood, Cells_EBV-transformed_lymphocytes
T1D	Pancreas, Whole_Blood, Cells_EBV-transformed_lymphocytes
ULC	Colon_Transverse, Colon_Sigmoid, Whole_Blood, Cells_EBV-transformed_lymphocytes

Table 10. List of selected and all features based on recursive feature elimination.

Features	Rank	Type
Betweenness_centrality	1	Network
Differential_expression	2	Network
Pagerank_centrality	3	Network
Weighted_degree	4	Network
Module_score	5	Network
tss_dist	6	GWAVA
avg_het	7	GWAVA
target_gene_known	8	FunSeq
TRAN	9	GWAVA
Closeness	10	Network
avg_daf	11	GWAVA
ss_dist	12	GWAVA
GC	13	GWAVA
INTRON	14	GWAVA
is_annotated_in_encode	15	FunSeq
dnase_fps	16	GWAVA
repeat	17	GWAVA
REP	18	GWAVA
taget_gene_is_hub	19	FunSeq
TSS	20	GWAVA
ENH	21	GWAVA
H3K9ac	22	GWAVA
UTR5	23	GWAVA
bound_motifs	24	GWAVA
ETS1	25	GWAVA
H4K20me1	26	GWAVA
EXON	27	GWAVA
FAIRE	28	GWAVA
cpg_island	29	GWAVA
is_motif_breaking	30	FunSeq
is_sensitive	31	FunSeq
EP300	32	GWAVA
H3K9me1	33	GWAVA
pwm	34	GWAVA
MXI1	35	GWAVA

CHAPTER 5: DISCUSSION AND FUTURE PERSPECTIVE

5.1 Summary

Network biology has proven to be a powerful tool for representing and analyzing complex biological networks. However, gene interactions are dynamical processes resulting in pathway rewiring. Thus, static networks are insufficient to capture corresponding dynamic events. On the other hand, many gene interactions are cell/tissue type specific. Construction of cell/tissue type specific network is still challenging due to limited data. The objective of my thesis research is to develop novel network-based computational methods to address the limitations of previous methods and solve new biological problems. First, our current knowledge about the dynamics of molecular networks during disease progression is rather limited. Particularly, only two conditions were usually considered in the previous work. Therefore, I developed the *i*MDM algorithm to study network dynamics. *i*MDM can identify both unique and shared modules from multiple gene networks, each of which denoting a different condition. Using *i*MDM algorithm, I identified different types of gene modules to understand heart failure progression and disease dynamics. Second, network construction is a prerequisite of network analysis. When the number of samples is limited, state-of-the-art computational methods for network construction are not robust. To address this issue, I developed a computational method to construct condition specific transcriptional regulatory network. I also developed a computational method to rank transcription factors in the transcriptional regulatory network. Applying this framework to RNA-seq data for hematopoietic stem cell development, I successfully constructed corresponding transcriptional regulatory networks and identified key transcriptional factors that play

important roles in endothelial-to-hematopoietic transition. Finally, I developed ARVIN, a network-based algorithm, to identify causal noncoding genetic variants for diseases. I validated ARVIN using gold-standard promoter and enhancer SNPs for a range of human diseases. After applying ARVIN to seven autoimmune diseases, we obtained a systematic understanding of the gene circuitry that is affected by all enhancer mutations in a given disease.

5.1.1 iMDM, an algorithm for the analysis of multiple gene networks

In this thesis, I first developed *i*MDM which is a network-based algorithm extracting both unique and shared gene modules across multiple gene networks. Previously, several lines of investigations have leveraged dynamic changes in molecular networks to predict disease outcomes. Focusing on hub genes in human protein interactome, several groups have shown that they can be categorized into different types based on topological measures such as degree and modularity (de Lichtenberg et al., 2005; Han et al., 2004). It was further demonstrated that such topological features of hub genes can be used to improve the prognosis of breast cancer patients (Taylor et al., 2009). Chuang *et al.* used a different strategy by examining the differentially expressed subnetworks (instead of hub genes) between two cohorts of breast cancer patients (Chuang et al., 2007).

After applying the *i*MDM to heart failure data, I found that condition-specific modules mediating different biological processes are associated with known cardiovascular phenotypes. By contrast, non-condition-specific modules have very different topological features. In addition, these modules that are more dynamic show

higher correlation with cardiac disease phenotype. In this way, we can understand disease progression and pathway dynamics.

5.1.2 Construction and analysis of condition-specific TRN

Reconstructing transcriptional regulatory networks from high-throughput data is a long-standing challenge. So far diverse computational methods or frameworks have been developed to solve this challenge (Marbach et al., 2012). However, those existing methods all rely on a relatively large number of samples to construct the network. To address this issue, I developed a computational framework to construct condition specific TRN. I first build a gene expression compendium by collecting data in related cell types from public databases or previous studies. Next, two networks are constructed using expression compendium with and without samples of interest. By comparing those two networks, we can get a condition-specific network by extracting edges unique to the network constructed using all samples. This method has been validated using a benchmarking dataset consisting of both gene expression profiling data and corresponding ChIP-Seq data.

After applying this method to HSC development data, I built 8 condition specific TRNs. I also developed a method to identify key transcriptional factors that play important roles in TRNs using their distance to differentially expressed genes. In this way, I found two lists of key TFs which potentially affect the development of HE cells and determine the difference between AGM and YS. Furthermore, many of those identified TFs are supported by previous studies or have related biological functions. For instance, we identified Spi1, Runx1 and Gfi1 which were previously found as required

factors to reprogram endothelial cells into hematopoietic cells (Sandler et al., 2014). In addition, TEAD factors I found were also reported to regulate hematopoietic specification such as Tead3 and Tead1 (Goode et al., 2016). For HE comparisons, we found quite a few clustered Hox family genes that are known to involve in hematopoiesis and hematopoietic stem cell function (Argiropoulos and Humphries, 2007). Therefore, those uncovered transcriptional factors significantly enhance our understanding in HSC development.

5.1.3 Identification of causal genetic variants for diseases

GWAS studies have revealed many genetic variants that are associated with different traits and diseases. However, interpretation of these variants especially variants in non-coding regions is still challenging. Molecular networks have been used extensively to improve the inference accuracy of causal coding variants. This potential has not been investigated to the same extent for noncoding variants. Therefore, we developed a novel computational framework ARVIN to prioritize non-coding genetic variants that are likely to be causal to certain diseases. We first developed a strategy to construct disease-relevant gene regulatory networks integrating epigenomic, transcriptomic and interactomic data. We then developed and characterized multiple network-based features which are more discriminative than existing genomic and epigenomics features. We applied our method to 233 regulatory promoter variants annotated in the Human Gene Mutation Database (HGMD) for 20 diseases and 15 enhancer variants for 10 diseases from published literature. ARVIN outperforms state-of-the-art methods using genomic and epigenomic features alone.

We next applied ARVIN to gene regulatory networks for 7 autoimmune diseases.

We identified causal SNP candidates and pathways which are likely to play roles in those autoimmune diseases. Our identified eSNPs are significantly supported by various evidence including eQTL, disease-associated genes, and drug targets. Further, enhancers harboring causal SNPs potentially have a combinatorial regulatory role in disease pathogenesis. Interestingly, genes targeted by mutiple risk eSNP-containing enhancers had more significant expression change than genes targeted by single enhancer. In addition, those multi-target genes are more central in the gene regulatory networks compared to single-targeted genes. This may suggest those enhancers that target multi-target genes have very important roles in disease progression.

5.2 Future directions

5.2.1 iMDM that integrates emerging genetic and epigenetic data

Currently, we have only applied *i*MDM to study diseases using gene expression profiling data. A key step of *i*MDM is to select seeds in the network for module expansion. When no prior information is available, genes will be chosen as seeds based on their topological features. For well-studied diseases, we can choose those seed genes which are known as disease-associated. Nevertheless, the set of disease-associated genes is either not available or very incomplete in most cases. Therefore, integrating multiple types of omics data beyond gene expression can further expand our ability to identify dynamic molecular events that are associated with phenotypic dynamics. For instance, genetic mutation data such as those from exome and whole-genome sequencing can be used as prior information to guide module search under the assumption that mutated sequences are likely to be involved in the diseases. For instance, genetic variants

databases provide plenty of prior information including GWAS, COSMIC, HGMD and dbGaP (Bamford et al., 2004; Mailman et al., 2007).

In addition, epigenomic data can also be integrated with transcriptome data to understand how environmental factors perturb gene networks. To date, the best understood epigenetic mechanisms are CpG DNA methylation and histone modification. DNA methylation is an epigenetic mechanism used by cells to control gene expression in many cellular processes, including embryonic development, transcription, chromatin, structure, X chromosome inactivation, genomic imprinting and chromosome stability (Robertson, 2005). DNA methylation has been the subject of intense interest because of its recognized roles in both healthy and disease development. Previous studies found that genomes of cancer cells are hypomethylated compared to their normal counterparts (Esteller, 2002; Feinberg and Tycko, 2004). DNA hypermethylation has been shown to silence tumor suppressor genes in cancer cells (Esteller, 2002; Wajed et al., 2001). On the other hand, it has become more and more evident that histone modifications are key players in the regulation of chromatin states and dynamics as well as in gene expression (Cohen et al., 2011; Portela and Esteller, 2010). Hence, integrating epigenetic data can greatly help network analysis and enhancer our understanding in normal and disease development.

5.2.2 TRN construction for single-cell RNA-seq data

Despite advances in cancer treatment, many patients still fail therapy, resulting in disease progression, recurrence, and reduced overall survival rate. A tumor is not simply a “bag” of homogeneous malignant cells. Instead, a tumor is a complex system consisting of different types of cells that can affect the tumor function. In this complex system,

tumor cells can show distinct morphological and phenotypic profiles, which is known as tumor heterogeneity (Zellmer and Zhang, 2014). Heterogeneity is common in all cancer types and at various levels including genetic, epigenetic, positional, and at the population level. The different subpopulations within a tumor mass interact with each other and influence other tumor cells both locally and at a distance. The extent of tumor heterogeneity is an emerging topic that researchers are only starting to understand (Alizadeh et al., 2015; Marusyk and Polyak, 2010). There are two models used to explain the heterogeneity of tumor cells: cancer stem cell model and the clonal evolution model. Cancer stem cells are cancer cells that have the ability to give rise to all cell types found in a particular cancer site. In this model, a smaller population of stem cells primarily triggers cancers. Cancer stem cell may generate tumors through the stem cell processes of self-renewal and differentiation into multiple cell types (Kreso and Dick, 2014). In clonal evolution model, tumors arise from a single mutated cell, accumulating additional mutations as it progresses. These changes result into more subpopulations, and each of these subpopulations is able to divide and mutate further (Greaves and Maley, 2012).

Single-cell analysis allows for a better understanding of invasion, metastasis, and therapy resistance during cancer progression. Recent technical development has enabled the transcriptomes of hundreds of cells to be profiled in an unbiased manner. Single-cell analysis can be accomplished by a number of technologies, including imaging approaches and combination of mRNA amplification and microfluidics approaches. Imaging techniques such as RNA-FISH can only detect the expression of only a small number of genes in each experiment (Kolodziejczyk et al., 2015). However, protocols with microfluidics approaches are able to examine the entire transcriptome of larger numbers

of single cells (Streets et al., 2014). Single-cell sequencing has emerged as a revolutionary tool that allows us to investigate scientific questions that have eluded examination such as understanding tumor heterogeneity.

With single-cell sequencing, the transcriptome of all tumors cells are profiled simultaneously. To study the network dynamic among all those cells, a cell-specific TRN needs to be constructed for each individual cell. Gene expression profiles for all sequenced cells can be used to build an expression compendium. Thus, cell-specific TRNs can be constructed by network comparison using method described in Chapter 3. Important TFs can be identified across all TRNs to understand their roles in cancer progression. On the other hand, *i*MDM algorithm is also feasible to identify gene modules across all cell-specific networks. For instance, modules shared by cells within the same subpopulation can help us characterize their features.

5.2.3 Identification of causal somatic mutations using ARVIN

Somatic mutation is an alteration in DNA that occurs after conception. Somatic mutations can occur in any of the cells of the body except for germ cells. In contrast, SNPs are germ line mutations that are detectable in a human population. Although most somatic mutations accumulated in our cells are harmless, occasionally a mutation influences a gene or regulatory element and leads to a phenotypic change. A subset of these mutations can confer a selective advantage to the cell resulting to preferential growth or survival. Cancers arise as a result of these somatic mutations that confer growth advantage (Martincorena and Campbell, 2015).

Recent technological advances such as whole-genome sequencing (WGS) have allowed the comprehensive characterization of somatic mutations in a large number of

tumor samples. WGS allows discovery of novel cancer-associated variants, including single nucleotide variants (SNVs), copy number variations (CNVs), and structural variants (SVs). Through comparing tumor and normal DNA, WGS is able to provide a comprehensive view of alterations in specific tumor sample. Therefore, we can identify causal noncoding mutations in specific cancer using ARVIN algorithm. Regulatory mutations can be found by overlapping with enhancers predicted using corresponding histone modification marks. In addition, regulatory mutations will be linked to target genes using computational predictions or chromatin interaction data. In this way, we can construct cancer specific gene regulatory networks and apply ARVIN to identify causal mutations.

5.2.4 Therapy development and drug discovery

Integrating network biology and pharmacology holds the promise of expanding the current opportunity space for drug discovery (Hopkins, 2008). Network-based approaches to human disease have multiple potential biological and clinical applications. It can provide a systematic understanding of the effects of cellular inter-connectedness on disease progression via identification of disease genes and pathways. These identified biomarkers can provide better targets for drug development and help disease classification.

*i*MDM algorithm allows us to find gene modules taking into account of disease dynamics and intricate network effects. Therefore, genes in identified modules can be further prioritized for drug target selection. Besides, gene modules can be also used to search for drug candidates in databases such as Library of Integrated Network-based Cellular Signatures (Vempati et al., 2014). Stage-specific gene modules can provide

candidates for developing therapies for patients at particular disease stage. For causal SNPs identified by ARVIN for a particular disease, their targets show very interesting topological features. Many of genes are also overlapped with known drug targets genes (described in Chapter 4). Hence, target genes of risk eSNP are also ideal candidates for drug development.

5.3 Conclusions

In this thesis work, I developed three network-based methods, *i*MDM, condition-specific TRN construction and ARVIN. Application of *i*MDM to RNA-seq data of heart failure model has provided a better understanding of network dynamics in the disease development. Condition-specific TRN construction is a powerful computational method for modeling transcriptional regulatory relationships between TFs and their targets. I demonstrated that ARVIN achieved improved accuracy in finding causal noncoding genetic variants for complex human diseases. I expect further applications and extensions of these three methods will continue to contribute to our understanding of heath and disease development.

REFERENCES

Adams, P.L., and Turnbull, D.M. (1996). Disorders of the electron transport chain. Journal of inherited metabolic disease *19*, 463-469.

Aittokallio, T., and Schwikowski, B. (2006). Graph-based methods for analysing networks in cell biology. Briefings in bioinformatics *7*, 243-255.

Akavia, U.D., and Benayahu, D. (2008). Meta-analysis and profiling of cardiac expression modules. Physiological genomics *35*, 305-315.

Albert, F.W., and Kruglyak, L. (2015). The role of regulatory variation in complex traits and disease. Nature reviews Genetics *16*, 197-212.

Alizadeh, A.A., Aranda, V., Bardelli, A., Blanpain, C., Bock, C., Borowski, C., Caldas, C., Califano, A., Doherty, M., Elsner, M.*, et al.* (2015). Toward understanding and exploiting tumor heterogeneity. Nature medicine *21*, 846-853.

Alonso-Lopez, D., Gutierrez, M.A., Lopes, K.P., Prieto, C., Santamaria, R., and De Las Rivas, J. (2016). APID interactomes: providing proteome-based interactomes with controlled quality for multiple species and derived networks. Nucleic acids research *44*, W529-535.

Andersson, R., Gebhard, C., Miguel-Escalada, I., Hoof, I., Bornholdt, J., Boyd, M., Chen, Y., Zhao, X., Schmidl, C., Suzuki, T.*, et al.* (2014). An atlas of active enhancers across human cell types and tissues. Nature *507*, 455-461.

Argiropoulos, B., and Humphries, R.K. (2007). Hox genes in hematopoiesis and leukemogenesis. Oncogene *26*, 6766-6776.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T.*, et al.* (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nature genetics *25*, 25-29.

Bader, G.D., Betel, D., and Hogue, C.W. (2003). BIND: the Biomolecular Interaction Network Database. Nucleic acids research *31*, 248-250.

Bagger, F.O., Rapin, N., Theilgaard-Monch, K., Kaczkowski, B., Jendholm, J., Winther, O., and Porse, B. (2012). HemaExplorer: a Web server for easy and fast visualization of gene expression in normal and malignant hematopoiesis. Blood *119*, 6394-6395.

Bailly-Bechet, M., Borgs, C., Braunstein, A., Chayes, J., Dagkessamanskaia, A., Francois, J.M., and Zecchina, R. (2011). Finding undetected protein associations in cell signaling by belief propagation. Proceedings of the National Academy of Sciences of the United States of America *108*, 882-887.

Bamford, S., Dawson, E., Forbes, S., Clements, J., Pettett, R., Dogan, A., Flanagan, A., Teague, J., Futreal, P.A., Stratton, M.R.*, et al.* (2004). The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website. British journal of cancer *91*, 355-358.

Bandyopadhyay, S., Mehta, M., Kuo, D., Sung, M.K., Chuang, R., Jaehnig, E.J., Bodenmiller, B., Licon, K., Copeland, W., Shales, M.*, et al.* (2010). Rewiring of genetic networks in response to DNA damage. Science *330*, 1385-1389.

Barabasi, A.L., Gulbahce, N., and Loscalzo, J. (2011). Network medicine: a network-based approach to human disease. Nature reviews Genetics *12*, 56-68.

Barabasi, A.L., and Oltvai, Z.N. (2004). Network biology: understanding the cell's functional organization. Nature reviews Genetics *5*, 101-113.

Barrat, A., Barthelemy, M., Pastor-Satorras, R., and Vespignani, A. (2004). The architecture of complex weighted networks. Proceedings of the National Academy of Sciences of the United States of America *101*, 3747-3752.
Bavelas, A. (1950). Communication patterns in task-oriented groups. J Acoust Soc Am *22*, 725–730.
Bebek, G., and Yang, J. (2007). PathFinder: mining signal transduction pathway segments from protein-protein interaction networks. BMC bioinformatics *8*, 335.
Benjamini Y, a.H.Y. (1995). Controlling the False Discovery Rate—a Practical and Powerful Approach to Multiple Testing. J Roy Stat Soc B Met *57*, 289–300.
Bertrand, J.Y., Chi, N.C., Santoso, B., Teng, S., Stainier, D.Y., and Traver, D. (2010). Haematopoietic stem cells derive directly from aortic endothelium during development. Nature *464*, 108-111.
Bisson N, J.D., Ivosev G, Tate SA, Bonner R, Taylor L, et al. (2011). Selected reaction monitoring mass spectrometry reveals the dynamics of signaling through the GRB2 adaptor. Nature biotechnology *29*, 653–658.
Blake, J.A., Bult, C.J., Eppig, J.T., Kadin, J.A., Richardson, J.E., and Mouse Genome Database, G. (2014). The Mouse Genome Database: integration of and access to knowledge about the laboratory mouse. Nucleic acids research *42*, D810-817.
Boisset, J.C., and Robin, C. (2012). On the origin of hematopoietic stem cells: progress and controversy. Stem cell research *8*, 1-13.
Boisset, J.C., van Cappellen, W., Andrieu-Soler, C., Galjart, N., Dzierzak, E., and Robin, C. (2010). In vivo imaging of haematopoietic cells emerging from the mouse aortic endothelium. Nature *464*, 116-120.
Bonneau, R., Reiss, D.J., Shannon, P., Facciotti, M., Hood, L., Baliga, N.S., and Thorsson, V. (2006). The Inferelator: an algorithm for learning parsimonious regulatory networks from systems-biology data sets de novo. Genome biology *7*, R36.
Brenowitz, M., Senear, D.F., and Kingston, R.E. (2001). DNase I footprint analysis of protein-DNA binding. Current protocols in molecular biology *Chapter 12*, Unit 12 14.
Brown, K.R., and Jurisica, I. (2005). Online predicted human interaction database. Bioinformatics *21*, 2076-2082.
Bruckner, A., Polge, C., Lentze, N., Auerbach, D., and Schlattner, U. (2009). Yeast two-hybrid, a powerful tool for systems biology. International journal of molecular sciences *10*, 2763-2788.
Carithers, L.J., and Moore, H.M. (2015). The Genotype-Tissue Expression (GTEx) Project. Biopreservation and biobanking *13*, 307-308.
Chatterjee, S., Kapoor, A., Akiyama, J.A., Auer, D.R., Lee, D., Gabriel, S., Berrios, C., Pennacchio, L.A., and Chakravarti, A. (2016). Enhancer Variants Synergistically Drive Dysfunction of a Gene Regulatory Network In Hirschsprung Disease. Cell *167*, 355-368 e310.
Chen, Y., Jiang, T., and Jiang, R. (2011). Uncover disease genes by maximizing information flow in the phenome-interactome network. Bioinformatics *27*, i167-176.
Cho, D.Y., Kim, Y.A., and Przytycka, T.M. (2012). Chapter 5: Network biology approach to complex diseases. PLoS computational biology *8*, e1002820.

Chorley, B.N., Wang, X., Campbell, M.R., Pittman, G.S., Noureddine, M.A., and Bell, D.A. (2008). Discovery and verification of functional single nucleotide polymorphisms in regulatory genomic regions: current and developing technologies. Mutat Res *659*, 147-157.
Chotinantakul, K., and Leeanansaksiri, W. (2012). Hematopoietic stem cell development, niches, and signaling pathways. Bone marrow research *2012*, 270425.
Chuang, H.Y., Lee, E., Liu, Y.T., Lee, D., and Ideker, T. (2007). Network-based classification of breast cancer metastasis. Molecular systems biology *3*, 140.
Cohen, I., Poreba, E., Kamieniarz, K., and Schneider, R. (2011). Histone modifiers in cancer: friends or foes? Genes & cancer *2*, 631-647.
Consortium, E.P. (2004). The ENCODE (ENCyclopedia Of DNA Elements) Project. Science *306*, 636-640.
Consortium, U.K., Walter, K., Min, J.L., Huang, J., Crooks, L., Memari, Y., McCarthy, S., Perry, J.R., Xu, C., Futema, M.*, et al.* (2015). The UK10K project identifies rare variants in health and disease. Nature *526*, 82-90.
Corradin, O., Saiakhova, A., Akhtar-Zaidi, B., Myeroff, L., Willis, J., Cowper-Sal lari, R., Lupien, M., Markowitz, S., and Scacheri, P.C. (2014). Combinatorial effects of multiple enhancer variants in linkage disequilibrium dictate levels of gene expression to confer susceptibility to common traits. Genome research *24*, 1-13.
Costantini, G., and Perugini, M. (2014). Generalization of clustering coefficients to signed correlation networks. PloS one *9*, e88669.
Cowper-Sal lari, R., Zhang, X., Wright, J.B., Bailey, S.D., Cole, M.D., Eeckhoute, J., Moore, J.H., and Lupien, M. (2012). Breast cancer risk-associated SNPs modulate the affinity of chromatin for FOXA1 and alter gene expression. Nat Genet *44*, 1191-1198.
Croft, D., Mundo, A.F., Haw, R., Milacic, M., Weiser, J., Wu, G., Caudy, M., Garapati, P., Gillespie, M., Kamdar, M.R.*, et al.* (2014). The Reactome pathway knowledgebase. Nucleic acids research *42*, D472-477.
Croft, D., O'Kelly, G., Wu, G., Haw, R., Gillespie, M., Matthews, L., Caudy, M., Garapati, P., Gopinath, G., Jassal, B.*, et al.* (2011). Reactome: a database of reactions, pathways and biological processes. Nucleic acids research *39*, D691-697.
Das, J., Mohammed, J., and Yu, H. (2012). Genome-scale analysis of interaction dynamics reveals organization of biological networks. Bioinformatics *28*, 1873-1878.
Davis, C.J., Gurbel, P.A., Gattis, W.A., Fuzaylov, S.Y., Nair, G.V., O'Connor, C.M., and Serebruany, V.L. (2000). Hemostatic abnormalities in patients with congestive heart failure: diagnostic significance and clinical challenge. International journal of cardiology *75*, 15-21.
De Las Rivas, J., and Fontanillo, C. (2010). Protein-protein interactions essentials: key concepts to building and analyzing interactome networks. PLoS computational biology *6*, e1000807.
de Lichtenberg, U., Jensen, L.J., Brunak, S., and Bork, P. (2005). Dynamic complex formation during the yeast cell cycle. Science *307*, 724-727.
Dekker, J., Rippe, K., Dekker, M., and Kleckner, N. (2002). Capturing chromosome conformation. Science *295*, 1306-1311.

Desler, C., Lykke, A., and Rasmussen, L.J. (2010). The effect of mitochondrial dysfunction on cytosolic nucleotide metabolism. Journal of nucleic acids *2010*.
Dewey, F.E., Perez, M.V., Wheeler, M.T., Watt, C., Spin, J., Langfelder, P., Horvath, S., Hannenhalli, S., Cappola, T.P., and Ashley, E.A. (2011). Gene coexpression network topology of cardiac development, hypertrophy, and failure. Circulation Cardiovascular genetics *4*, 26-35.
Diez J, G.A., Lopez B, and Querejeta R (2005). Mechanisms of disease: pathologic structural remodeling is more than adaptive hypertrophy in hypertensive heart disease. Nature clinical practice Cardiovascular medicine *2*, 209–216.
Dostie, J., Richmond, T.A., Arnaout, R.A., Selzer, R.R., Lee, W.L., Honan, T.A., Rubio, E.D., Krumm, A., Lamb, J., Nusbaum, C.*, et al.* (2006). Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. Genome research *16*, 1299-1309.
Doulatov, S., Notta, F., Laurenti, E., and Dick, J.E. (2012). Hematopoiesis: a human perspective. Cell stem cell *10*, 120-136.
Drozdov, I., Didangelos, A., Yin, X., Zampetaki, A., Abonnenc, M., Murdoch, C., Zhang, M., Ouzounis, C.A., Mayr, M., Tsoka, S.*, et al.* (2013). Gene network and proteomic analyses of cardiac responses to pathological and physiological stress. Circulation Cardiovascular genetics *6*, 588-597.
Eilken, H.M., Nishikawa, S., and Schroeder, T. (2009). Continuous single-cell imaging of blood generation from haemogenic endothelium. Nature *457*, 896-900.
Ellis, J.D., Barrios-Rodiles, M., Colak, R., Irimia, M., Kim, T., Calarco, J.A., Wang, X., Pan, Q., O'Hanlon, D., Kim, P.M.*, et al.* (2012). Tissue-specific alternative splicing remodels protein-protein interaction networks. Molecular cell *46*, 884-892.
Epstein, D.J. (2009). Cis-regulatory mutations in human disease. Briefings in functional genomics & proteomics *8*, 310-316.
Ernst, J., Kheradpour, P., Mikkelsen, T.S., Shoresh, N., Ward, L.D., Epstein, C.B., Zhang, X., Wang, L., Issner, R., Coyne, M.*, et al.* (2011). Mapping and analysis of chromatin state dynamics in nine human cell types. Nature *473*, 43-49.
Esteller, M. (2002). CpG island hypermethylation and tumor suppressor genes: a booming present, a brighter future. Oncogene *21*, 5427-5440.
Ezhkova, E., and Tansey, W.P. (2006). Chromatin immunoprecipitation to study protein-DNA interactions in budding yeast. Methods in molecular biology *313*, 225-244.
Faith, J.J., Hayete, B., Thaden, J.T., Mogno, I., Wierzbowski, J., Cottarel, G., Kasif, S., Collins, J.J., and Gardner, T.S. (2007). Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles. PLoS Biol *5*, e8.
Farh, K.K., Marson, A., Zhu, J., Kleinewietfeld, M., Housley, W.J., Beik, S., Shoresh, N., Whitton, H., Ryan, R.J., Shishkin, A.A.*, et al.* (2015). Genetic and epigenetic fine mapping of causal autoimmune disease variants. Nature *518*, 337-343.
Feinberg, A.P., and Tycko, B. (2004). The history of cancer epigenetics. Nature reviews Cancer *4*, 143-153.
Feuk, L., Carson, A.R., and Scherer, S.W. (2006). Structural variation in the human genome. Nature reviews Genetics *7*, 85-97.

Firpi, H.A., Ucar, D., and Tan, K. (2010). Discover regulatory DNA elements using chromatin signatures and artificial neural network. Bioinformatics *26*, 1579-1586.
Fisher, R.A. (1925). Statistical Methods for Research Workers. Oliver and Boyd.
Freedman, M.L., Monteiro, A.N., Gayther, S.A., Coetzee, G.A., Risch, A., Plass, C., Casey, G., De Biasi, M., Carlson, C., Duggan, D.*, et al.* (2011). Principles for the post-GWAS functional characterization of cancer risk loci. Nat Genet *43*, 513-518.
Fullwood, M.J., Han, Y., Wei, C.L., Ruan, X., and Ruan, Y. (2010). Chromatin interaction analysis using paired-end tag sequencing. Current protocols in molecular biology *Chapter 21*, Unit 21 15 21-25.
Garber, M., Grabherr, M.G., Guttman, M., and Trapnell, C. (2011). Computational methods for transcriptome annotation and quantification using RNA-seq. Nature methods *8*, 469-477.
Gargalovic, P.S., Imura, M., Zhang, B., Gharavi, N.M., Clark, M.J., Pagnon, J., Yang, W.P., He, A., Truong, A., Patel, S.*, et al.* (2006). Identification of inflammatory gene modules based on variations of human endothelial cell responses to oxidized lipids. Proceedings of the National Academy of Sciences of the United States of America *103*, 12741-12746.
Gilad, Y., Rifkin, S.A., and Pritchard, J.K. (2008). Revealing the architecture of gene regulation: the promise of eQTL studies. Trends in genetics : TIG *24*, 408-415.
Gill, R., Datta, S., and Datta, S. (2010). A statistical framework for differential network analysis from microarray data. BMC bioinformatics *11*, 95.
Gitter, A., Klein-Seetharaman, J., Gupta, A., and Bar-Joseph, Z. (2011). Discovering pathways by orienting edges in protein interaction networks. Nucleic acids research *39*, e22.
Goh, K.I., Cusick, M.E., Valle, D., Childs, B., Vidal, M., and Barabasi, A.L. (2007). The human disease network. Proceedings of the National Academy of Sciences of the United States of America *104*, 8685-8690.
Goode, D.K., Obier, N., Vijayabaskar, M.S., Lie, A.L.M., Lilly, A.J., Hannah, R., Lichtinger, M., Batta, K., Florkowska, M., Patel, R.*, et al.* (2016). Dynamic Gene Regulatory Networks Drive Hematopoietic Specification and Differentiation. Developmental cell *36*, 572-587.
Greaves, L.C., and Taylor, R.W. (2006). Mitochondrial DNA mutations in human disease. IUBMB life *58*, 143-151.
Greaves, M., and Maley, C.C. (2012). Clonal evolution in cancer. Nature *481*, 306-313.
Griffith, O.L., Montgomery, S.B., Bernier, B., Chu, B., Kasaian, K., Aerts, S., Mahony, S., Sleumer, M.C., Bilenky, M., Haeussler, M.*, et al.* (2008). ORegAnno: an open-access community-driven resource for regulatory annotation. Nucleic Acids Res *36*, D107-113.
Grubb, S.C., Bult, C.J., and Bogue, M.A. (2014). Mouse phenome database. Nucleic acids research *42*, D825-834.
Guan, Y., Myers, C.L., Lu, R., Lemischka, I.R., Bult, C.J., and Troyanskaya, O.G. (2008). A genomewide functional network for the laboratory mouse. PLoS computational biology *4*, e1000165.
Guenole A, S.R., Vreeken K, Wang ZZ, Wang S, Krogan NJ, et al. (2013). Dissection of DNA damage responses using multiconditional genetic interaction maps. . Molecular cell *49*, 346–358.

Guyatt, G.H. (1993). Measurement of health-related quality of life in heart failure. Journal of the American College of Cardiology *22*, 185A-191A.
Hamaneh, M.B., and Yu, Y.K. (2014). Relating diseases by integrating gene associations and information flow through protein interaction network. PloS one *9*, e110936.
Han, J.D., Bertin, N., Hao, T., Goldberg, D.S., Berriz, G.F., Zhang, L.V., Dupuy, D., Walhout, A.J., Cusick, M.E., Roth, F.P.*, et al.* (2004). Evidence for dynamically organized modularity in the yeast protein-protein interaction network. Nature *430*, 88-93.
Hao, D., Ren, C., and Li, C. (2012). Revisiting the variation of clustering coefficient of biological networks suggests new modular structure. BMC systems biology *6*, 34.
Harbison, C.T., Gordon, D.B., Lee, T.I., Rinaldi, N.J., Macisaac, K.D., Danford, T.W., Hannett, N.M., Tagne, J.B., Reynolds, D.B., Yoo, J.*, et al.* (2004). Transcriptional regulatory code of a eukaryotic genome. Nature *431*, 99-104.
Hartwell, L.H., Hopfield, J.J., Leibler, S., and Murray, A.W. (1999). From molecular to modular cell biology. Nature *402*, C47-52.
Haury, A.C., Mordelet, F., Vera-Licona, P., and Vert, J.P. (2012). TIGRESS: Trustful Inference of Gene REgulation using Stability Selection. Bmc Syst Biol *6*.
He, B., Chen, C., Teng, L., and Tan, K. (2014). Global view of enhancer-promoter interactome in human cells. Proceedings of the National Academy of Sciences of the United States of America *111*, E2191-2199.
Heintzman, N.D., Hon, G.C., Hawkins, R.D., Kheradpour, P., Stark, A., Harp, L.F., Ye, Z., Lee, L.K., Stuart, R.K., Ching, C.W.*, et al.* (2009). Histone modifications at human enhancers reflect global cell-type-specific gene expression. Nature *459*, 108-112.
Hellman, L.M., and Fried, M.G. (2007). Electrophoretic mobility shift assay (EMSA) for detecting protein-nucleic acid interactions. Nature protocols *2*, 1849-1861.
Hermjakob, H., Montecchi-Palazzi, L., Lewington, C., Mudali, S., Kerrien, S., Orchard, S., Vingron, M., Roechert, B., Roepstorff, P., Valencia, A.*, et al.* (2004). IntAct: an open source molecular interaction database. Nucleic acids research *32*, D452-455.
Hindorff, L.A., Sethupathy, P., Junkins, H.A., Ramos, E.M., Mehta, J.P., Collins, F.S., and Manolio, T.A. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci U S A *106*, 9362-9367.
Hirschi, K.K. (2012). Hemogenic endothelium during development and beyond. Blood *119*, 4823-4827.
Hofmann, I., Stover, E.H., Cullen, D.E., Mao, J., Morgan, K.J., Lee, B.H., Kharas, M.G., Miller, P.G., Cornejo, M.G., Okabe, R.*, et al.* (2009). Hedgehog signaling is dispensable for adult murine hematopoietic stem cell function and hematopoiesis. Cell stem cell *4*, 559-567.
Hofree, M., Shen, J.P., Carter, H., Gross, A., and Ideker, T. (2013). Network-based stratification of tumor mutations. Nat Methods *10*, 1108-1115.
Holloway, B., Luck, S., Beatty, M., Rafalski, J.A., and Li, B. (2011). Genome-wide expression quantitative trait loci (eQTL) analysis in maize. BMC genomics *12*, 336.
Hong, J.W., Hendrix, D.A., and Levine, M.S. (2008). Shadow enhancers as a source of evolutionary novelty. Science *321*, 1314.
Hopkins, A.L. (2008). Network pharmacology: the next paradigm in drug discovery. Nature chemical biology *4*, 682-690.

Hou, J., and Kang, Y.J. (2012). Regression of pathological cardiac hypertrophy: signaling pathways and therapeutic targets. Pharmacology & therapeutics *135*, 337-354.
Huan, T., Zhang, B., Wang, Z., Joehanes, R., Zhu, J., Johnson, A.D., Ying, S., Munson, P.J., Raghavachari, N., Wang, R.*, et al.* (2013). A systems biology framework identifies molecular underpinnings of coronary heart disease. Arteriosclerosis, thrombosis, and vascular biology *33*, 1427-1434.
Huang, S.S., and Fraenkel, E. (2009). Integrating proteomic, transcriptional, and interactome data reveals hidden components of signaling and regulatory networks. Science signaling *2*, ra40.
Hunter, J.J., and Chien, K.R. (1999). Signaling pathways for cardiac hypertrophy and failure. The New England journal of medicine *341*, 1276-1283.
Huynh-Thu, V.A., Irrthum, A., Wehenkel, L., and Geurts, P. (2010). Inferring regulatory networks from expression data using tree-based methods. PLoS One *5*.
I. Ljubić, R.W., U. Pferschy, G. W. Klau, P. Mutzel, M. Fischetti (2006). An algorithmic framework for the exact solution of the Prize-Collecting Steiner Tree Problem. Mathematical Programming *105*.
Javierre, B.M., Burren, O.S., Wilder, S.P., Kreuzhuber, R., Hill, S.M., Sewitz, S., Cairns, J., Wingett, S.W., Varnai, C., Thiecke, M.J.*, et al.* (2016). Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoters. Cell *167*, 1369-1384 e1319.
Jia, P., Zheng, S., Long, J., Zheng, W., and Zhao, Z. (2011). dmGWAS: dense module searching for genome-wide association studies in protein-protein interaction networks. Bioinformatics *27*, 95-102.
Kahan, T., and Bergfeldt, L. (2005). Left ventricular hypertrophy in hypertension: its arrhythmogenic potential. Heart *91*, 250-256.
Kaimakis, P., Crisan, M., and Dzierzak, E. (2013). The biochemistry of hematopoietic stem cell development. Biochimica et biophysica acta *1830*, 2395-2403.
Kandoth, C., McLellan, M.D., Vandin, F., Ye, K., Niu, B., Lu, C., Xie, M., Zhang, Q., McMichael, J.F., Wyczalkowski, M.A.*, et al.* (2013). Mutational landscape and significance across 12 major cancer types. Nature *502*, 333-339.
Kanehisa, M., Goto, S., Sato, Y., Furumichi, M., and Tanabe, M. (2012). KEGG for integration and interpretation of large-scale molecular data sets. Nucleic acids research *40*, D109-114.
Karamanlidis, G., Lee, C.F., Garcia-Menendez, L., Kolwicz, S.C., Jr., Suthammarak, W., Gong, G., Sedensky, M.M., Morgan, P.G., Wang, W., and Tian, R. (2013). Mitochondrial complex I deficiency increases protein acetylation and accelerates heart failure. Cell metabolism *18*, 239-250.
Karwacz, K., Miraldi, E.R., Pokrovskii, M., Madi, A., Yosef, N., Wortman, I., Chen, X., Watters, A., Carriero, N., Awasthi, A.*, et al.* (2017). Critical role of IRF1 and BATF in forming chromatin landscape during type 1 regulatory cell differentiation. Nat Immunol.
Kathiresan, S., and Srivastava, D. (2012). Genetics of human cardiovascular disease. Cell *148*, 1242-1257.
Kendziorski, C.M., Chen, M., Yuan, M., Lan, H., and Attie, A.D. (2006). Statistical methods for expression quantitative trait loci (eQTL) mapping. Biometrics *62*, 19-27.

Khong, D.M., Dudakov, J.A., Hammett, M.V., Jurblum, M.I., Khong, S.M., Goldberg, G.L., Ueno, T., Spyroglou, L., Young, L.F., van den Brink, M.R.*, et al.* (2015). Enhanced hematopoietic stem cell function mediates immune regeneration following sex steroid blockade. Stem cell reports *4*, 445-458.
Khurana, E., Fu, Y., Colonna, V., Mu, X.J., Kang, H.M., Lappalainen, T., Sboner, A., Lochovsky, L., Chen, J., Harmanci, A.*, et al.* (2013). Integrative annotation of variants from 1092 humans: application to cancer genomics. Science *342*, 1235587.
Kilpinen, H., Waszak, S.M., Gschwind, A.R., Raghav, S.K., Witwicki, R.M., Orioli, A., Migliavacca, E., Wiederkehr, M., Gutierrez-Arcelus, M., Panousis, N.I.*, et al.* (2013). Coordinated effects of sequence variation on DNA binding, chromatin structure, and transcription. Science *342*, 744-747.
Kircher, M., Witten, D.M., Jain, P., O'Roak, B.J., Cooper, G.M., and Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. Nature genetics *46*, 310-315.
Kissa, K., and Herbomel, P. (2010). Blood stem cells emerge from aortic endothelium by a novel type of cell transition. Nature *464*, 112-115.
Kolodziejczyk, A.A., Kim, J.K., Svensson, V., Marioni, J.C., and Teichmann, S.A. (2015). The technology and biology of single-cell RNA sequencing. Molecular cell *58*, 610-620.
Komuro, I. (2001). Molecular mechanism of cardiac hypertrophy and development. Japanese circulation journal *65*, 353-358.
Kreso, A., and Dick, J.E. (2014). Evolution of the cancer stem cell model. Cell stem cell *14*, 275-291.
Krymskii, L.D. (1958). [Causes of decompensation in cardiac enlargement; review of literature]. Eksperimental'naia khirurgiia *3*, 57-64.
Kuang, S.Q., Geng, L., Prakash, S.K., Cao, J.M., Guo, S., Villamizar, C., Kwartler, C.S., Peters, A.M., Brasier, A.R., and Milewicz, D.M. (2013). Aortic remodeling after transverse aortic constriction in mice is attenuated with AT1 receptor blockade. Arteriosclerosis, thrombosis, and vascular biology *33*, 2172-2179.
Kuhn, M. (2008). Building predictive models in R using the caret package. Journal of Statistical Software *28*.
Lancrin, C., Sroczynska, P., Stephenson, C., Allen, T., Kouskoff, V., and Lacaud, G. (2009). The haemangioblast generates haematopoietic cells through a haemogenic endothelium stage. Nature *457*, 892-895.
Langfelder, P., and Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. BMC bioinformatics *9*, 559.
Lee, E., Cho, S., Kim, K., and Park, T. (2009). An integrated approach to infer causal associations among gene expression, genotype variation, and disease. Genomics *94*, 269-277.
Lee, I., Blom, U.M., Wang, P.I., Shim, J.E., and Marcotte, E.M. (2011). Prioritizing candidate disease genes by network-based boosting of genome-wide association data. Genome Res *21*, 1109-1121.
Lee, I., Date, S.V., Adai, A.T., and Marcotte, E.M. (2004). A probabilistic functional network of yeast genes. Science *306*, 1555-1558.

Lee, M.J., Ye, A.S., Gardino, A.K., Heijink, A.M., Sorger, P.K., MacBeath, G., and Yaffe, M.B. (2012). Sequential application of anticancer drugs enhances cell death by rewiring apoptotic signaling networks. Cell *149*, 780-794.
Leek, J.T., Johnson, W.E., Parker, H.S., Jaffe, A.E., and Storey, J.D. (2012). The sva package for removing batch effects and other unwanted variation in high-throughput experiments. Bioinformatics *28*, 882-883.
Leng, N., Dawson, J.A., Thomson, J.A., Ruotti, V., Rissman, A.I., Smits, B.M., Haag, J.D., Gould, M.N., Stewart, R.M., and Kendziorski, C. (2013). EBSeq: an empirical Bayes hierarchical model for inference in RNA-seq experiments. Bioinformatics *29*, 1035-1043.
Levo, M., and Segal, E. (2014). In pursuit of design principles of regulatory sequences. Nature reviews Genetics *15*, 453-468.
Li, G., Ruan, X., Auerbach, R.K., Sandhu, K.S., Zheng, M., Wang, P., Poh, H.M., Goh, Y., Lim, J., Zhang, J.*, et al.* (2012). Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. Cell *148*, 84-98.
Li, Y., Esain, V., Teng, L., Xu, J., Kwan, W., Frost, I.M., Yzaguirre, A.D., Cai, X., Cortes, M., Maijenburg, M.W.*, et al.* (2014). Inflammatory signaling regulates embryonic hematopoietic stem and progenitor cell production. Genes Dev *28*, 2597-2612.
Liao, Y., Smyth, G.K., and Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics *30*, 923-930.
Lim, Y., and Matsui, W. (2010). Hedgehog signaling in hematopoiesis. Critical reviews in eukaryotic gene expression *20*, 129-139.
Linghu, B., Snitkin, E.S., Hu, Z., Xia, Y., and Delisi, C. (2009). Genome-wide prioritization of disease genes and identification of disease-disease associations from an integrated human functional linkage network. Genome Biol *10*, R91.
Lissy, N.A., Davis, P.K., Irwin, M., Kaelin, W.G., and Dowdy, S.F. (2000). A common E2F-1 and p73 pathway mediates cell death induced by TCR activation. Nature *407*, 642-645.
Luscombe, N.M., Babu, M.M., Yu, H., Snyder, M., Teichmann, S.A., and Gerstein, M. (2004). Genomic analysis of regulatory network dynamics reveals large topological changes. Nature *431*, 308-312.
Ma X, G.L., and Tan K (2014). Modeling Disease Progression Using Dynamics of Pathway Connectivity. Bioinformatics.
Magli, M.C., Largman, C., and Lawrence, H.J. (1997). Effects of HOX homeobox genes in blood cell differentiation. Journal of cellular physiology *173*, 168-177.
Maillet, M., van Berlo, J.H., and Molkentin, J.D. (2013). Molecular basis of physiological heart growth: fundamental concepts and new players. Nature reviews Molecular cell biology *14*, 38-48.
Mailman, M.D., Feolo, M., Jin, Y., Kimura, M., Tryka, K., Bagoutdinov, R., Hao, L., Kiang, A., Paschall, J., Phan, L.*, et al.* (2007). The NCBI dbGaP database of genotypes and phenotypes. Nature genetics *39*, 1181-1186.
Mant, J., Al-Mohammad, A., Swain, S., Laramee, P., and Guideline Development, G. (2011). Management of chronic heart failure in adults: synopsis of the National Institute For Health and clinical excellence guideline. Annals of internal medicine *155*, 252-259.

Marbach, D., Costello, J.C., Kuffner, R., Vega, N.M., Prill, R.J., Camacho, D.M., Allison, K.R., Consortium, D., Kellis, M., Collins, J.J., *et al.* (2012). Wisdom of crowds for robust gene network inference. Nature methods *9*, 796-804.
Martin, P., McGovern, A., Orozco, G., Duffus, K., Yarwood, A., Schoenfelder, S., Cooper, N.J., Barton, A., Wallace, C., Fraser, P., *et al.* (2015). Capture Hi-C reveals novel candidate genes and complex long-range interactions with related autoimmune risk loci. Nat Commun *6*, 10069.
Martincorena, I., and Campbell, P.J. (2015). Somatic mutation in cancer and normal cells. Science *349*, 1483-1489.
Marusyk, A., and Polyak, K. (2010). Tumor heterogeneity: causes and consequences. Biochimica et biophysica acta *1805*, 105-117.
Maurano, M.T., Humbert, R., Rynes, E., Thurman, R.E., Haugen, E., Wang, H., Reynolds, A.P., Sandstrom, R., Qu, H., Brody, J., *et al.* (2012). Systematic localization of common disease-associated variation in regulatory DNA. Science *337*, 1190-1195.
McDowall, M.D., Scott, M.S., and Barton, G.J. (2009). PIPs: human protein-protein interaction prediction database. Nucleic acids research *37*, D651-656.
McKinney-Freeman, S., Cahan, P., Li, H., Lacadie, S.A., Huang, H.T., Curran, M., Loewer, S., Naveiras, O., Kathrein, K.L., Konantz, M., *et al.* (2012). The transcriptional landscape of hematopoietic stem cell ontogeny. Cell Stem Cell *11*, 701-714.
Medvinsky, A., Taoudi, S., Mendes, S., and Dzierzak, E. (2008). Analysis and manipulation of hematopoietic progenitor and stem cells from murine embryonic tissues. Curr Protoc Stem Cell Biol *Chapter 2*, Unit 2A 6.
Meyers, D.E., Basha, H.I., and Koenig, M.K. (2013). Mitochondrial cardiomyopathy: pathophysiology, diagnosis, and management. Texas Heart Institute journal *40*, 385-394.
Michael J. Chen, T.Y., Brandon M. Zeigler, Elaine Dzierzak, Nancy A. Speck (2009). Runx1 is required for the endothelial to haematopoietic cell transition but not thereafter. Nature *457*, 887-891.
Moreau, Y., and Tranchevent, L.C. (2012). Computational tools for prioritizing candidate genes: boosting disease gene discovery. Nat Rev Genet *13*, 523-536.
Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods *5*, 621-628.
Nacher, J.C., Ueda, N., Yamada, T., Kanehisa, M., and Akutsu, T. (2004). Clustering under the line graph transformation: application to reaction network. BMC bioinformatics *5*, 207.
Nishimura (2001). BioCarta. Biotech Software & Internet Report *2*, 117–120.
Noonan, J.P., and McCallion, A.S. (2010). Genomics of long-range regulatory elements. Annual review of genomics and human genetics *11*, 1-23.
North, T., Gu, T.L., Stacy, T., Wang, Q., Howard, L., Binder, M., Marin-Padilla, M., and Speck, N.A. (1999). Cbfa2 is required for the formation of intra-aortic hematopoietic clusters. Development *126*, 2563-2575.
Nottingham, W.T., Jarratt, A., Burgess, M., Speck, C.L., Cheng, J.F., Prabhakar, S., Rubin, E.M., Li, P.S., Sloane-Stanley, J., Kong, A.S.J., *et al.* (2007). Runx1-mediated hematopoietic stem-cell emergence is controlled by a Gata/Ets/SCL-regulated enhancer. Blood *110*, 4188-4197.

Oldridge, D.A., Wood, A.C., Weichert-Leahey, N., Crimmins, I., Sussman, R., Winter, C., McDaniel, L.D., Diamond, M., Hart, L.S., Zhu, S.*, et al.* (2015). Genetic predisposition to neuroblastoma mediated by a LMO1 super-enhancer polymorphism. Nature *528*, 418-421.
Ourfali, O., Shlomi, T., Ideker, T., Ruppin, E., and Sharan, R. (2007). SPINE: a framework for signaling-regulatory pathway inference from cause-effect experiments. Bioinformatics *23*, i359-366.
Ouwerkerk, P.B., and Meijer, A.H. (2001). Yeast one-hybrid screening for DNA-protein interactions. Current protocols in molecular biology *Chapter 12*, Unit 12 12.
Page, L. (1998). The PageRank Citation Ranking: Bringing Order to the Web. Technical Report.
Pajcini, K.V., Speck, N.A., and Pear, W.S. (2011). Notch signaling in mammalian hematopoietic stem cells. Leukemia *25*, 1525-1532.
Patwardhan, R.P., Hiatt, J.B., Witten, D.M., Kim, M.J., Smith, R.P., May, D., Lee, C., Andrie, J.M., Lee, S.I., Cooper, G.M.*, et al.* (2012). Massively parallel functional dissection of mammalian enhancers in vivo. Nat Biotechnol *30*, 265-270.
Pearson, K. (1905). The problem of the random walk. Nature *72*, 294.
Pepe, M., Longton, G., and Janes, H. (2009). Estimation and Comparison of Receiver Operating Characteristic Curves. The Stata journal *9*, 1.
Peri, S., Navarro, J.D., Kristiansen, T.Z., Amanchy, R., Surendranath, V., Muthusamy, B., Gandhi, T.K., Chandrika, K.N., Deshpande, N., Suresh, S.*, et al.* (2004). Human protein reference database as a discovery resource for proteomics. Nucleic acids research *32*, D497-501.
Portela, A., and Esteller, M. (2010). Epigenetic modifications and human disease. Nature biotechnology *28*, 1057-1068.
Pujana, M.A., Han, J.D., Starita, L.M., Stevens, K.N., Tewari, M., Ahn, J.S., Rennert, G., Moreno, V., Kirchhoff, T., Gold, B.*, et al.* (2007). Network modeling links breast cancer susceptibility and centrosome dysfunction. Nature genetics *39*, 1338-1349.
Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., de Bakker, P.I., Daly, M.J.*, et al.* (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet *81*, 559-575.
Quan, C., Wang, M., and Ren, F. (2014). An unsupervised text mining method for relation extraction from biomedical literature. PloS one *9*, e102039.
R. Ahlswede, N.C. (2000). Network information flow. IEEE Transactions on Information Theory *46*, 1204-1216.
Ramamoorthy, H., Abraham, P., and Isaac, B. (2014). Mitochondrial dysfunction and electron transport chain complex defect in a rat model of tenofovir disoproxil fumarate nephrotoxicity. Journal of biochemical and molecular toxicology *28*, 246-255.
Ritchie, G.R., Dunham, I., Zeggini, E., and Flicek, P. (2014). Functional annotation of noncoding sequence variants. Nature methods *11*, 294-296.
Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., and Smyth, G.K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res *43*, e47.

Robertson, K.D. (2005). DNA methylation and human disease. Nature reviews Genetics *6*, 597-610.
Robinson, M.D., McCarthy, D.J., and Smyth, G.K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics *26*, 139-140.
Rockman, M.V., and Kruglyak, L. (2006). Genetics of global gene expression. Nature reviews Genetics *7*, 862-872.
Romanoski, C.E., Che, N., Yin, F., Mai, N., Pouldar, D., Civelek, M., Pan, C., Lee, S., Vakili, L., Yang, W.P*., et al.* (2011). Network for activation of human endothelial cells by oxidized phospholipids: a critical role of heme oxygenase 1. Circulation research *109*, e27-41.
Ruwhof, C., and van der Laarse, A. (2000). Mechanical stress-induced cardiac hypertrophy: mechanisms and signal transduction pathways. Cardiovascular research *47*, 23-37.
Sadeghi, A., and Frohlich, H. (2013). Steiner tree methods for optimal sub-network identification: an empirical study. BMC bioinformatics *14*, 144.
Saetre, R., Yoshida, K., Miwa, M., Matsuzaki, T., Kano, Y., and Tsujii, J. (2010). Extracting protein interactions from text with the unified AkaneRE event extraction system. IEEE/ACM transactions on computational biology and bioinformatics *7*, 442-453.
Sandler, V.M., Lis, R., Liu, Y., Kedem, A., James, D., Elemento, O., Butler, J.M., Scandura, J.M., and Rafii, S. (2014). Reprogramming human endothelial cells to haematopoietic cells requires vascular induction. Nature *511*, 312-318.
Sanyal, A., Lajoie, B.R., Jain, G., and Dekker, J. (2012). The long-range interaction landscape of gene promoters. Nature *489*, 109-113.
Segal, E., Shapira, M., Regev, A., Pe'er, D., Botstein, D., Koller, D., and Friedman, N. (2003). Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. Nature genetics *34*, 166-176.
Segovia Cubero, J., Alonso-Pulpon Rivera, L., Peraira Moral, R., and Silva Melchor, L. (2004). [Heart failure: etiology and approach to diagnosis]. Revista espanola de cardiologia *57*, 250-259.
Shakya, A., Goren, A., Shalek, A., German, C.N., Snook, J., Kuchroo, V.K., Yosef, N., Chan, R.C., Regev, A., Williams, M.A*., et al.* (2015). Oct1 and OCA-B are selectively required for CD4 memory T cell function. J Exp Med *212*, 2115-2131.
Shih, Y.K., and Parthasarathy, S. (2012). A single source k-shortest paths algorithm to infer regulatory pathways in a gene network. Bioinformatics *28*, i49-58.
Shimizu, I., and Minamino, T. (2016). Physiological and pathological cardiac hypertrophy. Journal of molecular and cellular cardiology *97*, 245-262.
Simonis, M., Klous, P., Splinter, E., Moshkin, Y., Willemsen, R., de Wit, E., van Steensel, B., and de Laat, W. (2006). Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). Nature genetics *38*, 1348-1354.
Stark, C., Breitkreutz, B.J., Reguly, T., Boucher, L., Breitkreutz, A., and Tyers, M. (2006). BioGRID: a general repository for interaction datasets. Nucleic acids research *34*, D535-539.

Stefano Monti, P.T., Jill Mesirov, Todd Golub (2003). Consensus Clustering: A Resampling-Based Method
for Class Discovery and Visualization of Gene
Expression Microarray Data. Machine Learning *52*, 91-118.
Stenson, P.D., Ball, E.V., Howells, K., Phillips, A.D., Mort, M., and Cooper, D.N. (2009). The Human Gene Mutation Database: providing a comprehensive central mutation database for molecular diagnostics and personalized genomics. Hum Genomics *4*, 69-72.
Streets, A.M., Zhang, X., Cao, C., Pang, Y., Wu, X., Xiong, L., Yang, L., Fu, Y., Zhao, L., Tang, F.*, et al.* (2014). Microfluidic single-cell whole-transcriptome sequencing. Proceedings of the National Academy of Sciences of the United States of America *111*, 7048-7053.
Stuart, J.M., Segal, E., Koller, D., and Kim, S.K. (2003). A gene-coexpression network for global discovery of conserved genetic modules. Science *302*, 249-255.
Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S.*, et al.* (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences of the United States of America *102*, 15545-15550.
Swiers, G., de Bruijn, M., and Speck, N.A. (2010). Hematopoietic stem cell emergence in the conceptus and the role of Runx1. The International journal of developmental biology *54*, 1151-1163.
Tan, N., Chung, M.K., Smith, J.D., Hsu, J., Serre, D., Newton, D.W., Castel, L., Soltesz, E., Pettersson, G., Gillinov, A.M.*, et al.* (2013). Weighted gene coexpression network analysis of human left atrial tissue identifies gene modules associated with atrial fibrillation. Circulation Cardiovascular genetics *6*, 362-371.
Tardiff, J.C. (2006). Cardiac hypertrophy: stressing out the heart. The Journal of clinical investigation *116*, 1467-1470.
Tarnavski O, M.J., Schinke M, Nie Q, Kong S, and Izumo S (2004). Mouse cardiac surgery: comprehensive techniques for the generation of mouse models of human diseases and their application for genomic studies. Physiol Genomics *16*, 349–360.
Taylor, I.W., Linding, R., Warde-Farley, D., Liu, Y., Pesquita, C., Faria, D., Bull, S., Pawson, T., Morris, Q., and Wrana, J.L. (2009). Dynamic modularity in protein interaction networks predicts breast cancer outcome. Nature biotechnology *27*, 199-204.
Touzet, H., and Varre, J.S. (2007). Efficient and accurate P-value computation for Position Weight Matrices. Algorithms Mol Biol *2*, 15.
Trapnell, C., Pachter, L., and Salzberg, S.L. (2009). TopHat: discovering splice junctions with RNA-Seq. Bioinformatics *25*, 1105-1111.
Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J., and Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nature biotechnology *28*, 511-515.
Trynka, G., Hunt, K.A., Bockett, N.A., Romanos, J., Mistry, V., Szperl, A., Bakker, S.F., Bardella, M.T., Bhaw-Rosun, L., Castillejo, G.*, et al.* (2011). Dense genotyping identifies

and localizes multiple common and rare variant association signals in celiac disease. Nat Genet *43*, 1193-1201.
Tuupanen, S., Turunen, M., Lehtonen, R., Hallikas, O., Vanharanta, S., Kivioja, T., Bjorklund, M., Wei, G., Yan, J., Niittymaki, I.*, et al.* (2009). The common colorectal cancer predisposition SNP rs6983267 at chromosome 8q24 confers potential to enhanced Wnt signaling. Nat Genet *41*, 885-890.
Ussher, J.R., Jaswal, J.S., and Lopaschuk, G.D. (2012). Pyridine nucleotide regulation of cardiac intermediary metabolism. Circulation research *111*, 628-641.
Vasilescu, J., Guo, X., and Kast, J. (2004). Identification of protein-protein interactions using in vivo cross-linking and mass spectrometry. Proteomics *4*, 3845-3854.
Vempati, U.D., Chung, C., Mader, C., Koleti, A., Datar, N., Vidovic, D., Wrobel, D., Erickson, S., Muhlich, J.L., Berriz, G.*, et al.* (2014). Metadata Standard and Data Exchange Specifications to Describe, Model, and Integrate Complex and Diverse High-Throughput Screening Data from the Library of Integrated Network-based Cellular Signatures (LINCS). Journal of biomolecular screening *19*, 803-816.
Verfaillie, A., Svetlichnyy, D., Imrichova, H., Davie, K., Fiers, M., Kalender Atak, Z., Hulselmans, G., Christiaens, V., and Aerts, S. (2016). Multiplex enhancer-reporter assays uncover unsophisticated TP53 enhancer logic. Genome research *26*, 882-895.
Visel, A., Rubin, E.M., and Pennacchio, L.A. (2009). Genomic views of distant-acting enhancers. Nature *461*, 199-205.
Vo, L.T., and Daley, G.Q. (2015). De novo generation of HSCs from somatic and pluripotent stem cell sources. Blood *125*, 2641-2648.
von Mering, C., Jensen, L.J., Snel, B., Hooper, S.D., Krupp, M., Foglierini, M., Jouffre, N., Huynen, M.A., and Bork, P. (2005). STRING: known and predicted protein-protein associations, integrated and transferred across organisms. Nucleic acids research *33*, D433-437.
Wajed, S.A., Laird, P.W., and DeMeester, T.R. (2001). DNA methylation: an alternative pathway to cancer. Annals of surgery *234*, 10-20.
Waring, C.D., Vicinanza, C., Papalamprou, A., Smith, A.J., Purushothaman, S., Goldspink, D.F., Nadal-Ginard, B., Torella, D., and Ellison, G.M. (2014). The adult heart responds to increased workload with physiologic hypertrophy, cardiac stem cell activation, and new myocyte formation. European heart journal *35*, 2722-2731.
Watson, I.R., Takahashi, K., Futreal, P.A., and Chin, L. (2013). Emerging patterns of somatic mutations in cancer. Nature reviews Genetics *14*, 703-718.
Watson-Haigh, N.S., Kadarmideen, H.N., and Reverter, A. (2010). PCIT: an R package for weighted gene co-expression networks based on partial correlation and information theory approaches. Bioinformatics *26*, 411-413.
Weber, K.T., and Brilla, C.G. (1991). Pathological hypertrophy and cardiac interstitium. Fibrosis and renin-angiotensin-aldosterone system. Circulation *83*, 1849-1865.
Weirauch, M.T., Yang, A., Albu, M., Cote, A.G., Montenegro-Montero, A., Drewe, P., Najafabadi, H.S., Lambert, S.A., Mann, I., Cook, K.*, et al.* (2014). Determination and inference of eukaryotic transcription factor sequence specificity. Cell *158*, 1431-1443.

Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., Klemm, A., Flicek, P., Manolio, T., Hindorff, L.*, et al.* (2014). The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Res *42*, D1001-1006.
Westra, H.J., Peters, M.J., Esko, T., Yaghootkar, H., Schurmann, C., Kettunen, J., Christiansen, M.W., Fairfax, B.P., Schramm, K., Powell, J.E.*, et al.* (2013). Systematic identification of trans eQTLs as putative drivers of known disease associations. Nat Genet *45*, 1238-1243.
White, M.A., Myers, C.A., Corbo, J.C., and Cohen, B.A. (2013). Massively parallel in vivo enhancer assay reveals that highly local features determine the cis-regulatory function of ChIP-seq peaks. Proceedings of the National Academy of Sciences of the United States of America *110*, 11952-11957.
Wilson, C.L., and Miller, C.J. (2005). Simpleaffy: a BioConductor package for Affymetrix Quality Control and data analysis. Bioinformatics *21*, 3683-3685.
Workman, C.T., Mak, H.C., McCuine, S., Tagne, J.B., Agarwal, M., Ozier, O., Begley, T.J., Samson, L.D., and Ideker, T. (2006). A systems approach to mapping DNA damage response pathways. Science *312*, 1054-1059.
Wu, S., Shao, F., Ji, J., Sun, R., Dong, R., Zhou, Y., Xu, S., Sui, Y., and Hu, J. (2015). Network propagation with dual flow for gene prioritization. PloS one *10*, e0116505.
Yang, H., Qin, C., Li, Y.H., Tao, L., Zhou, J., Yu, C.Y., Xu, F., Chen, Z., Zhu, F., and Chen, Y.Z. (2016). Therapeutic target database update 2016: enriched resource for bench to clinical drug target and targeted pathway information. Nucleic Acids Res *44*, D1069-1074.
Yang, J., Ferreira, T., Morris, A.P., Medland, S.E., Genetic Investigation of, A.T.C., Replication, D.I.G., Meta-analysis, C., Madden, P.A., Heath, A.C., Martin, N.G.*, et al.* (2012). Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. Nat Genet *44*, 369-375, S361-363.
Yang, Z., Fujii, H., Mohan, S.V., Goronzy, J.J., and Weyand, C.M. (2013). Phosphofructokinase deficiency impairs ATP generation, autophagy, and redox balance in rheumatoid arthritis T cells. J Exp Med *210*, 2119-2134.
Yang, Z., Matteson, E.L., Goronzy, J.J., and Weyand, C.M. (2015). T-cell metabolism in autoimmune disease. Arthritis Res Ther *17*, 29.
Yokomizo, T., Yamada-Inagawa, T., Yzaguirre, A.D., Chen, M.J., Speck, N.A., and Dzierzak, E. (2012). Whole-mount three-dimensional imaging of internally localized immunostained cells within mouse embryos. Nat Protoc *7*, 421-431.
Zanzoni, A., Montecchi-Palazzi, L., Quondam, M., Ausiello, G., Helmer-Citterich, M., and Cesareni, G. (2002). MINT: a Molecular INTeraction database. FEBS letters *513*, 135-140.
Zellmer, V.R., and Zhang, S. (2014). Evolving concepts of tumor heterogeneity. Cell & bioscience *4*, 69.
Zhang, B., Gaiteri, C., Bodea, L.G., Wang, Z., McElwee, J., Podtelezhnikov, A.A., Zhang, C., Xie, T., Tran, L., Dobrin, R.*, et al.* (2013a). Integrated systems approach identifies genetic nodes and networks in late-onset Alzheimer's disease. Cell *153*, 707-720.

Zhang, B., Tian, Y., Jin, L., Li, H., Shih Ie, M., Madhavan, S., Clarke, R., Hoffman, E.P., Xuan, J., Hilakivi-Clarke, L.*, et al.* (2011). DDN: a caBIG(R) analytical tool for differential network analysis. Bioinformatics *27*, 1036-1038.
Zhang, J., Jiang, M., Yuan, F., Feng, K.Y., Cai, Y.D., Xu, X., and Chen, L. (2013b). Identification of age-related macular degeneration related genes by applying shortest path algorithm in protein-protein interaction network. BioMed research international *2013*, 523415.
Zhou, Q., Kesteven, S., Wu, J., Aidery, P., Gawaz, M., Gramlich, M., Feneley, M.P., and Harvey, R.P. (2015). Pressure Overload by Transverse Aortic Constriction Induces Maladaptive Hypertrophy in a Titin-Truncated Mouse Model. BioMed research international *2015*, 163564.
Zhu, X., Gerstein, M., and Snyder, M. (2007). Getting connected: analysis and principles of biological networks. Genes & development *21*, 1010-1024.