From DEPARTMENT OF MEDICAL BIOCHEMISTRY AND BIOPHYSICS
Karolinska Institutet, Stockholm, Sweden

CROSS-TISSUE REGULATORY GENE NETWORKS IN CORONARY ATHEROSCLEROSIS

Husain Ahammad Talukdar

Karolinska Institutet

Stockholm 2016

Published by Karolinska Institutet.
Printed by Eprint AB 2016.

ISBN 978-91-7676-357-5

Cross-Tissue Regulatory Gene Networks in Coronary Atherosclerosis
THESIS FOR DOCTORAL DEGREE (Ph.D.)

By

Husain Ahammad Talukdar

Principal Supervisor:
Professor Johan LM Björkegren
Karolinska Institutet
Department of Medical Biochemistry and Biophysics
Division of Vascular Biology

Co-supervisor(s):
Associate Professor Josefin Skogsberg
Karolinska Institutet
Department of Medical Biochemistry and Biophysics
Division of Vascular Biology

Reader Tom Michoel
The University of Edinburgh
The Roslin Institute
Division of Genetics and Genomics

Opponent:
Associate Professor Frank Emmert-Streib
Tampere University of Technology
Department of Signal Processing

Examination Board:
Associate Professor Carsten Daub
Karolinska Institutet
Department of Biosciences and Nutrition

Professor Ann-Christine Syvänen
Uppsala Universitet
Department of Medical Sciences
Division of Molecular Medicine

Dr Sven Nelander
Uppsala Universitet
Department of Immunology, Genetics and Pathology
Division of Neuro-Oncology

Dedicated

to

My family

আব্বা-আম্মা, তোমাদের কষ্ট, ভালোবাসা, ত্যাগ, আর দোয়া ...

জিসা, তোমার প্রেরণা, আর হার না মানার মন্ত্র ...

জায়েফ, প্রশান্তির মহৌষধ ...

ABSTRACT

Coronary artery disease (CAD) is the underlying cause of myocardial infarction and stroke that together are responsible for nearly 30% of all global deaths. CAD is a common complex disease caused by the interactions of multiple genetic and environmental risk factors acting across several metabolic and vascular tissues. Owing to the complexity of these interactions, systems genetics is an increasingly recognized path to a better understanding of complex diseases. In this thesis, we applied systems genetics by integrating the analysis of genotype (DNA) and global gene expression (RNA) data from metabolic and vascular tissues with phenotype data from the clinically well-characterized subjects in the Stockholm Atherosclerosis Gene Expression (STAGE) study. We validated the initial findings using genome-wide association studies (GWAS) and several gene expression datasets from mice and cell models. As a result, we for the first time inferred regulatory gene networks (RGNs) with key drivers of CAD, several of its main risk factors and atherosclerosis regression.

In **paper I,** we designed a computational pipeline to reconstruct RGNs with key drivers in CAD using the STAGE study. Then, by integrating expression quantitative traits (eQTLs) of these RGNs with genotype data from several GWAS, 30 CAD-causal RGNs interconnected in blood, vascular and metabolic tissues were identified. Twelve of these RGNs were further validated in gene expression and phenotype data from the Hybrid Mouse Diversity Panel. As proof of concept, by targeting the key drivers *AIP*, *DRAP1*, *POLR2I*, and *PQBP1* in a cross-species-validated, arterial-wall RGN involving RNA-processing genes, we re-identified this RGN in THP-1 foam cells and independent gene expression data from CAD macrophages and carotid lesions.

In **paper II,** we developed a cross-tissue weighted gene co-expression network analysis (X-WGCNA) method (used in Paper I) that reliably captures gene activities both within and across tissues. X-WGCNA is implemented as a package in R and is available online.

In **paper III**, we inferred transcription factor (TF) RGNs from three plasma cholesterol lowering (PCL)-responsive gene sets causally related to regression of early, mature, and advanced mouse atherosclerosis. We then used THP-1 cells in an in vitro atherosclerosis regression model to successfully validate 3 key drivers in these RGNs driving regression in early (*PPARG*), mature (*MLL5*), and advanced (*SRSF10*/*XRN2*) atherosclerosis.

In **paper IV**, we inferred the STAGE eQTLs (used in papers I and II) and identified subsets with gene regulatory effects across multiple tissues that according to GWAS were highly enriched in association with CAD. To better understand the pathophysiological role of these multi-tissue eQTLs, we identified and analyzed a number of associated gene sets.

A key result of this thesis is a repository of RGNs with key drivers for CAD, CAD risk factors, and atherosclerosis regression. This repository together with the computational pipeline including X-WGCNA should be useful in future studies that aim to go beyond genetic loci identified by GWAS and provide opportunities for novel diagnostics and therapies.

LIST OF SCIENTIFIC PAPERS

I. Cross-Tissue Regulatory Gene Networks in Coronary Artery Disease.†‡
HUSAIN A. TALUKDAR, Hassan Foroughi Asl, Rajeev K. Jain, Raili Ermel, Arno Ruusalepp, Oscar Franzén, Brian A. Kidd, Ben Readhead, Chiara Giannarelli, Jason C. Kovacic, Torbjörn Ivert, Joel T. Dudley, Mete Civelek, Aldons J. Lusis, Eric E. Schadt, Josefin Skogsberg, Tom Michoel, and Johan L.M. Björkegren
Cell Systems, 2016, 2, 3:196-208

† ***Editorial Preview****:* Deciphering the Dark Matter of Complex Genetic Inheritance; Crawford, Nigel P.S.; *Cell Systems*, 2016, 2, 3:144-146
‡ ***ECCB 2016***: True scientific highlights of the year in bioinformatics and computational biology (Systems category)

II. X-WGCNA: an R package for Cross-Tissue Weighted Gene Coexpression Network Analysis.
HUSAIN A. TALUKDAR, Yifan Mo, Hassan Foroughi Asl, Josefin Skogsberg, Johan Bjorkegren and Tom Michoel.
Manuscript

III. Plasma Cholesterol–Induced Lesion Networks Activated before Regression of Early, Mature, and Advanced Atherosclerosis.
Johan L. M. Björkegren, Sara Hägg, **HUSAIN A. TALUKDAR***, Hassan Foroughi Asl*, Rajeev K. Jain, Cecilia Cedergren, Ming-Mei Shang, Aránzazu Rossignoli, Rabbe Takolander, Olle Melander, Anders Hamsten, Tom Michoel, and Josefin Skogsberg
PLOS GENETICS, 2014, 10(2), e1004201
*Equal contribution

IV. Expression quantitative trait Loci acting across multiple tissues are enriched in inherited risk for coronary artery disease.
Hassan Foroughi Asl, **HUSAIN A. TALUKDAR**, Alida S.D. Kindt, Rajeev K. Jain, Raili Ermel, Arno Ruusalepp, Khanh-Dung H. Nguyen, Radu Dobrin, Dermot F. Reilly, Heribert Schunkert, Nilesh J. Samani, Ingrid Braenne, Jeanette Erdmann, Olle Melander, Jianlong Qi, Torbjörn Ivert, Josefin Skogsberg, Eric E. Schadt, Tom Michoel, Johan L.M. Björkegren, CARDIoGRAM Consortium.
Circulation: Cardiovascular Genetics, 2015, 8, 305-315

Related scientific papers but not included in this thesis

I. Lim domain binding 2: a key driver of transendothelial migration of leukocytes and atherosclerosis.
Shang MM, **HUSAIN A. TALUKDAR**, Hofmann JJ, Niaudet C, Asl HF, Jain RK, Rossignoli A, Cedergren C, Silveira A, Gigante B, Leander K, de Faire U, Hamsten A, Ruusalepp A, Melander O, Ivert T, Michoel T, Schadt EE, Betsholtz C, Skogsberg J, Björkegren JL.
Arterioscler Thromb Vasc Biol. 2014 Sep;34(9):2068-77

II. Cardiometabolic Risk Loci Share Downstream *Cis*- and *Trans*-Genes Across Tissues and Diseases.
The Stockholm-Tartu Atherosclerosis Network Engineering Task (STARNET) Study
Oscar Franzén*, Raili Ermel*, Ariella Cohain*, Nicholas K. Akers, Antonio Di Narzo, **HUSAIN A. TALUKDAR**, Hassan Foroughi-Asl, Claudia Giambartolomei, John F. Fullard, Katyayani Sukhavasi, Sulev Köks, Li-Ming Gan, Chiara Gianarelli, Jason C. Kovacic, Christer Betsholtz, Bojan Losic, Tom Michoel, Ke Hao, Panos Roussos, Josefin Skogsberg, Arno Ruusalepp, Eric E. Schadt and Johan L.M. Björkegren
Accepted in Science
*Equal contribution

III. Poliovirus Receptor-Related 2– A Cholesterol Responsive Gene Affecting Atherosclerosis Development by Modulating Leukocyte Migration
Aránzazu Rossignoli, Ming-Mei Shang, Christine Mössinger, Hassan Foroughi Asl, **HUSAIN A. TALUKDAR**, Oscar Franzén, Erika Folestad, Johan L.M. Björkegren, Josefin Skogsberg
Manuscript -- in revision

IV. Alternatively Spliced Tissue Factor Isoform is a Critical Driver of Foam Cell Formation and Plaque Burden
Daniel Alicea Delgado, Aleksey Chudnovskiy, Claudia Calcagno, **HUSAIN A. TALUKDAR,** Dong Kwong Yang, Dongtak Jeong, Yifan Mo, Saboor Hekmaty, Gustav J. Strijkers, Gerald A. Soff , Jason C. Kovacic, Thomas Weber, Zahi A. Fayad, Juan Badimon, Valentin Fuster, Josefin Skogsberg, Miriam Merad, Johan L.M. Björkegren, Roger J. Hajjar, Chiara Giannarelli
Manuscript -- in revision

TABLE OF CONTENTS

1 Introduction 1
1.1 Background 2
1.1.1 Coronary artery disease (CAD) 2
1.1.2 CAD-risk factor 2
1.1.3 Atherosclerosis development 2
1.2 NEW: “Network-enabled wisdom” in biology, medicine and health 4
1.2.1 High-throughput data collection techniques 6
1.2.2 Biological network 6
1.2.3 Expression quantitative trait locus 9
1.2.4 Genome-wide association studies 9
1.2.5 Causal network 10
1.2.6 Clinical cohort 10
2 Aim of the thesis 12
3 Materials and methods 13
3.1 Genetics of gene expression cohort 13
3.1.1 STAGE 13
3.1.2 SÖS 16
3.1.3 HMDP 16
3.2 Atherosclerosis mouse model 17
3.2.1 Mouse model 17
3.2.2 Mouse dataset 17
3.3 Data pre-processing 17
3.3.1 Data normalization 17
3.3.2 Principal component analysis 17
3.4 GWAS datasets 17
3.5 Computation analysis of a set of genes of interest 18
3.5.1 Reconstruction of TF-RGNs 18
3.5.2 Key driver identification, part I 18
3.6 Computation analysis of genetics of gene expression data 18
3.6.1 Module identification 19
3.6.2 Cross-tissue modules robustness 20
3.6.3 GO function analysis 21
3.6.4 Disease-associated modules 21
3.6.5 Identification of causal modules 21
3.6.6 Reconstruction of regulatory gene networks 23
3.6.7 Key driver identification, part II 24
3.7 Validation of regulatory gene networks 24
3.7.1 Cross-species validation 24
3.7.2 Key driver validation 24
3.7.3 Module connectivity 25
3.7.4 Module differential connectivity 25

3.8 Reconstruction of super network 25
4 Result 26
4.1 Paper I 26
4.2 Paper II 30
4.3 Paper III 32
4.4 Paper IV 34
5 Discussion 36
6 Concluding remarks and future works 38
7 Acknowledgements 39
8 References 42

LIST OF ABBREVIATIONS

AAW	Atherosclerostic Arterial Wall
ACE	Angiotensin-Converting Enzyme
BIC	Bayesian Information Criterion
CABG	Coronary Artey Bypass Grafting
CAD	Coronary Artery Disease
CARDIoGRAM	Coronary ARtery DIsease Genome wide Replication and Meta-analysis
CCD	Common Complex Disease
CDF	Chip Description File
CLR	Context Likelihood of Relatedness
CRP	C-Reactive Protein
CT	Cross-Tissue
DAG	Directed Acyclic Graph
DNA	DeoxyriboNucleic Acid
DPI	Data Processing Inequality
DS	Diameter Stenosis
eQTL	Expression Quantitative Trait Locus
FDR	False Discovery Rate
GGE	Genetics of Gene Expression
GO	Gene Ontology
GTEx	Genotype-Tissue Expression
GWAS	Genome-Wide Association Studies
HDL	High-Density Lipoprotein
HMDP	Hybrid Mouse Diversity Panel
IMA	Internal Mammary Artery
IMT	Intima-Media Thickness
KD	Key Driver
LDL	Low-Density Lipoprotein
Mbp	Mega base pairs
MDC	Module Differential Connectivity
MI	Myocardial Infarction
MIGen	Myocardial Infarction Genetics Consortium
mRNA	Messenger RNA
NEW	Network-Enabled Wisdom
QCA	Quantitative Coronary Angiography
RGN	Regulatory Gene Network

RMA	Robust Multi-array average
RNA	RiboNucleic Acid
RNA-seq	RNA sequencing
SD	Standard Deviation
SF	Subcutaneous Fat
SM	Skeletal Muscle
SMC	Smooth Muscle Cell
SNP	Single Nucleotide Polymorphism
STAGE	Stockholm ATherosclerosis Gene Expression
STARNET	Stockholm-Tartu Atherosclerosis Reverse Network Engineering Task
TC	Total Cholesterol
TF	Transcription Factor
TOM	Topological Overlap Matrix
TS	Tissue-Specific
VF	Visceral Fat
VLDL	Very Low-Density Lipoprotein
WB	Whole Blood
WGCNA	Weighted Gene Coexpression Network Analysis
WHO	World Health Organization
WTCCC	Wellcome Trust Case Control Consortium
X-WGCNA	Cross-Tissue Weighted Gene Coexpression Network Analysis

1 INTRODUCTION

Atherosclerosis is a disease of the arterial wall where lipid-rich plaques form over a life time, leading to a gradual impairment of the blood flow. As the plaques expands into the arterial lumen, blood flow becomes increasingly restricted and eventually the stress to the endothelial layer leads to a rupture and clot formation by aggravating blood platelets. Together, the plaque and clot can cause a complete blockage of the local arterial blood flow; blockage of the coronary arteries of the heart leads to a myocardial infarction (MI). If a blood clot instead is released, it might travel with the circulation to the brain and cause a stroke. Atherosclerosis of the coronary arteries is called coronary artery disease (CAD) —the most common cause of mortality worldwide. In fact, according to a report by the World Health Organization (WHO) as well as other studies [1-5], cardiovascular diseases were responsible for 17.5 million deaths in 2012, which corresponds to 31% of all global deaths; 14 million of these deaths were due to CAD or stroke. These statistics are surprising, given lifestyle improvements (in particular, a sharp reduction in the number of active smokers) and the successful targeting of CAD risk factors, such as hypercholesterolemia by statins [6] and hypertension by beta-blockers [7], but at the same time underscore the urgency of developing new research strategies to battle atherosclerosis and CAD.

One such strategy is systems biology, which integrates genetics (i.e., DNA associations with phenotypes) with genomics (i.e., RNA phenotypes) and is therefore increasingly referred to as *systems genetics* [8-13]. Systems genetics is based on the idea of understanding the flow of all biological information from inherited DNA variants that increase risk of disease through and their interactions with environmental factors to produce complex disease phenotypes. Systems genetics has largely been made possible by the development and refinement of high-throughput next-generation techniques. This development has also led to markedly reduced cost and allow genome-wide scale (omics) measurements of DNA, RNA (transcriptomics), proteins (proteomics), metabolites (metabolomics), and lipids (lipidomics). A chief reason for the development of systems genetics is genome-wide association studies (GWAS). In GWAS large cohorts consisting of tens of thousands of patients and controls are compared to find DNA variants associated with disease. For CAD, GWAS have identified more than 150 genetic risk loci [14, 15]. The challenge now is to identify the causal mechanisms by which these loci contribute to CAD development. It remains unclear if single variants within these loci affect the expression or function of a single gene in a given cell type or whether instead multiple DNA variants within loci affect several genes across cell types, tissues, or even organs to cause disease [8, 16]. In parallel to many single-gene approaches, we and others [8-10, 12] have proposed a network-enabled wisdom ("NEW") strategy, which relies on sophisticated algorithms incorporated into computational tools that can be applied to data from GWAS, genetic (DNA) and genomic (RNA) studies to identify disease-driving molecular networks.

In this thesis, I describe a computational pipeline we developed that integrates multi-tissue systems genetics and GWAS as well as cross-species analyses. When this pipeline was first

applied to the Stockholm Atherosclerosis Gene Expression (STAGE) study (to which I have had unique access during my PhD studies), we were able to robustly reconstruct and validate regulatory gene networks (RGNs) and key disease drivers in CAD. To implement this pipeline, we mainly relied on existing methods but also developed a new computational tool that extends the established method of weighted gene co-expression network analysis (WGCNA) by making it applicable to cross-tissue (CT) analysis (X-WGCNA).

1.1 BACKGROUND

As the name implies, CAD affects the coronary arteries of the heart to cause MI [17]. The development of CAD, however, is chiefly determined by a number of metabolic and inflammatory CAD risk factors.

1.1.1 Coronary artery disease (CAD)

Arteries are vessels within the body that carry oxygen-rich blood to all our organs, including the heart. Atherosclerosis affects most large to middle-sized arteries, including those in the heart, brain, legs, pelvis, and kidneys. CAD in the coronary circulation of the heart is a progressive life-long inflammatory disease [17-19]. If you do not have CAD you are highly unlikely to have an MI, although MI is a result of both CAD and thrombosis (i.e., the formation of blood clots). Thus, preventing CAD can also to a great extent prevent MI. It is therefore important to understand the risk factors and etiology of CAD. CAD is a multifactorial common complex disease. Although the central disease process in CAD is in the arterial walls of the heart, plaque formation (see further below) is also driven by metabolic factors [20], including hypercholesterolemia and diabetes as well as hypertension (mechanical stress to the arterial wall). Metabolic factors are mostly caused by molecular defects in the main metabolic organs, such as the liver, skeletal muscle, and fat deposits.

1.1.2 CAD-risk factor

CAD has both genetic risk factors, which cannot be modified, and environmental risk factors, which can be modified [17, 19]. Genetic risk factors include elevated levels of low-density lipoprotein (LDL)/very low-density lipoprotein (VLDL) [21] and lipoprotein (a), hypertension [22], age, sex, family history [23], obesity, [24] and diabetes [22]. Environmental risk factors include a high-fat diet, smoking, lack of exercise, and infectious agents.

1.1.3 Atherosclerosis development

Normal large artery consists of three layers, **Figure 1** [17, 19, 25]:

1) Intima (the innermost layer), a monolayer of endothelial cells oriented toward the circulating blood and a basal membrane oriented toward the media.
2) Media (the middle layer), consisting of smooth muscle cells (SMCs)
3) Adventitia (the outer layer), consisting mostly of structural cell types such fibroblasts.

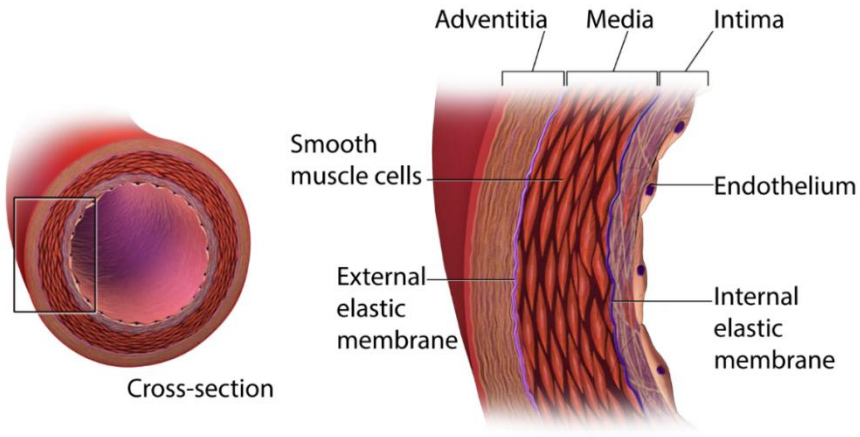

Figure 1: The structure of healthy aorta. *(left) Cross section of an aorta. (right) Different layers.* Modified from [26]

Atherosclerosis starts with the activation of the endothelium, the cell layer between the blood flow and the arterial wall (specifically, the intima), particularly at sites of turbulent blood flow, such as bifurcations. This leads to the accumulation of cholesterol-rich low-density lipoprotein (LDL) particles at these sites in the sub-endothelial matrix, where they become oxidized, which further activates the endothelium. The activated endothelium now attracts circulating inflammatory cells (i.e., white blood cells, primarily monocytes) that migrate into the arterial wall (intima). Inside the intima, migrating monocytes proliferate and also differentiate into macrophages. These macrophages react to and internalize oxidized LDL. Eventually, the macrophages become lipid-laden with an appearance in the microscope resembling *foam cells*, which together form so-called fatty streaks—the starting point for the development of atherosclerotic plaque (**Figure 2**).

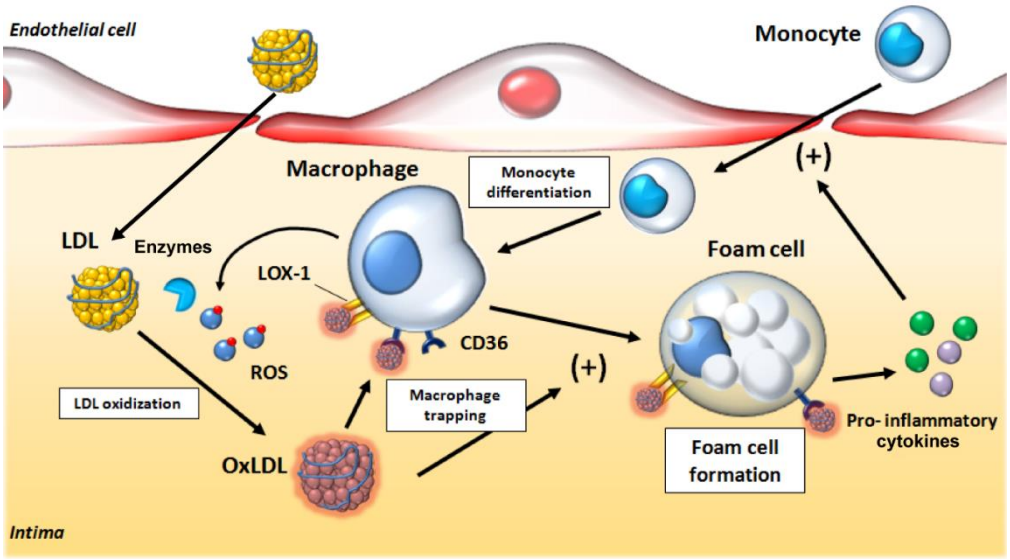

Figure 2: Foam cell formation. Modified from [27]

As foam cells increase in number, they begin to accumulate in a necrotic core consisting of live foam cells as well as foam cells undergoing apoptosis or necrosis. As a result, vascular smooth muscle cells start to migrate from the media to the intima layer of the artery, where they start

to produce collagen to form fibrous cap to contain the growing core of lipids. As the necrotic core with its fibrous cap grows into the lumen of the artery, it starts to impede blood flow. Over time, the plaques harden, and the fibrous cap can break, causing a so-called plaque rupture that triggers the coagulation systems to form a blood clot. An MI happens when a plaque ruptures in the coronary arteries, which together with the resulting blood clot may occlude blood flow and cause distal ischemia to heart muscle supplied by the affected artery. This leads typically to severe chest pain radiating to the left arm. The blood clot can also detach, forming a blood embolus that may end up in the brain and cause a stroke or elsewhere in the body and cause ischemia (typically muscle pain).

Atherosclerosis is believed to be a life-long progressive disease, which can start as early as the second decade of life. The initial stages of the disease are asymptomatic (**Figure 3**). In the later stages, when the clinical events occur, it is generally too late the regress CAD to healthy coronary arteries. Today, in patients suffering from CAD (e.g., chest-pain after heavy or mild labor) or MI, the acute event is often treated successfully. However, these patients eventually die from heart failure due to multiple smaller heart attacks, which result in insufficient healthy heart muscle to maintain adequate circulation.

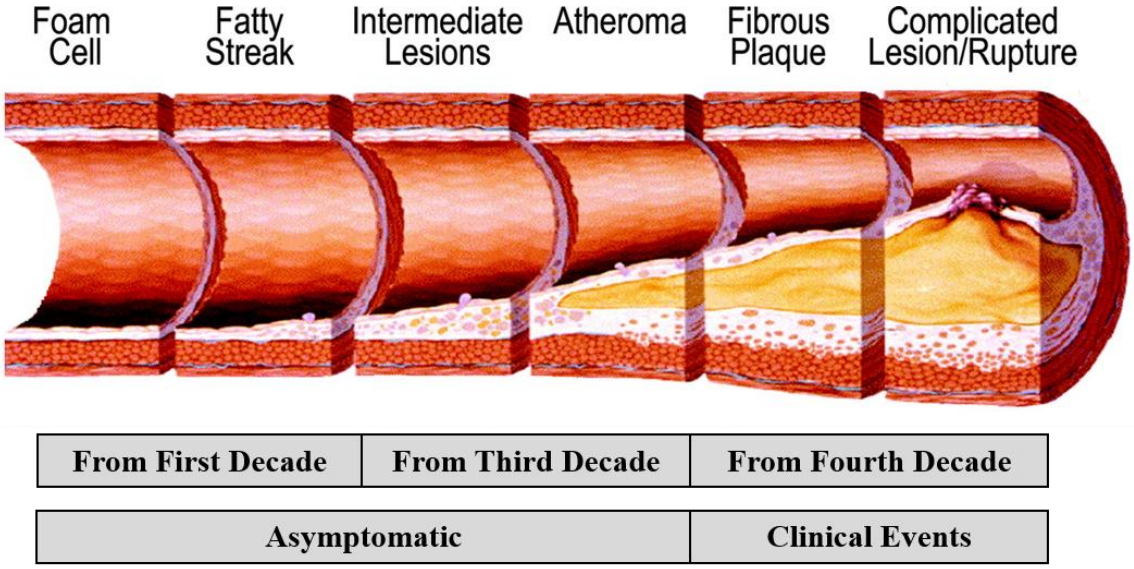

Figure 3: Different stages of atherosclerosis development. Modified from [28]

1.2 NEW: "NETWORK-ENABLED WISDOM" IN BIOLOGY, MEDICINE AND HEALTH

NEW biology is a data-driven approach based on accurate generation of high-throughput genetic and genomic data and high-performance computing [9-13, 29, 30]. In biology, medicine, and health, NEW rests on the fact that the human body consists of highly integrated organ systems. As a result, normal functioning as well as malfunctioning (i.e., disease states) must always involve the dynamic interaction of thousands of genes, proteins, and metabolites across multiple cell and tissue types shifting over time. This is in vast contrast to single- or so-called Mendelian diseases, in which frequently only one or a few genes in a single, linear pathway are responsible for the malfunction/disease. Thus, to capture the true complexity of

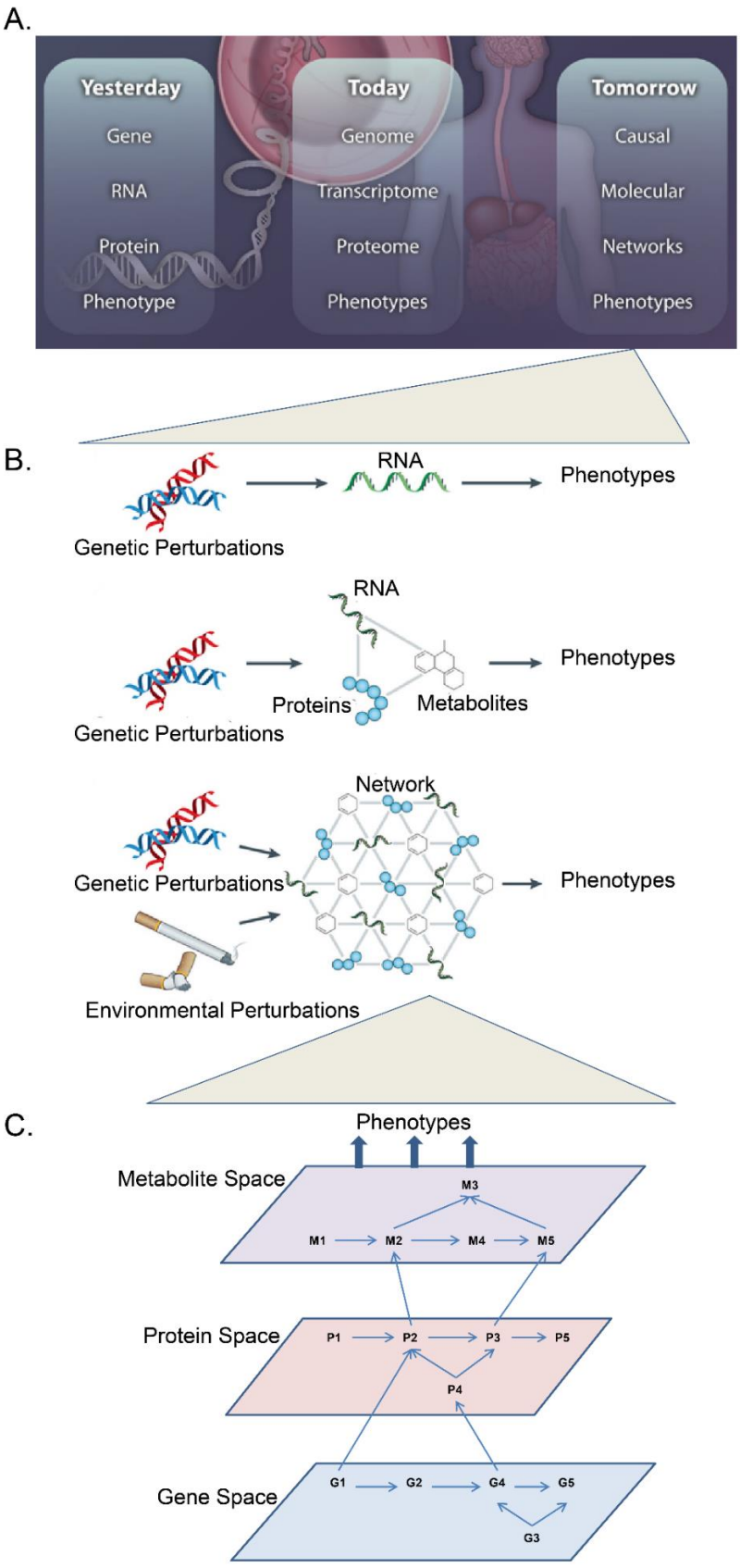

Figure 4: NEW biology strategies. *Panel **A**, generation of genetics-of-gene-expression (GGE) studies. Panel **B**, various model of GGE studies. Panel **C**, different layers of networks across multiple biological scales- genes, proteins, and metabolites.* Modified from [10, 12, 31]

normal molecular biology and how such states are perturbed in states of disease, the NEW biology puts forward RGNs as a much more appropriate representation of molecular biology than linear pathways. The development of high-throughput data collection and high-performance computing has allowed the NEW biology era to surface by provided the necessary means to infer RGNs from observational measurements in hundreds to thousands of individuals.

A challenge for the NEW biology is to design computational tools to reconstruct molecular networks that are causal (not reactive or independent) for variation in disease phenotypes across individuals (**Figure 4A**).

The first model assumes that an alteration in DNA alleles—such as a single nucleotide polymorphism (SNP), a copy-number variation, or insertions/deletions—is directly linked to a phenotype through a single intermediate phenotype such as a transcript. In the second model, multiple intermediate states (such as transcripts, proteins, and metabolites) can be activated by the genetic variation to cause a phenotype. These multiple intermediate effectors including environmental perturbations can, through omics measurements, be captured in regulatory networks (**Figure 4B**). According to the central dogma, different intermediate phenotypes can be ordered into a causal hierarchy in relation to the end phenotype/disease (**Figure 4C**).

1.2.1 High-throughput data collection techniques

With advanced technologies, it is now possible to quantify global transcript levels in relevant tissues by using hybridization-based microarray or next-generation sequencing methods of RNA sequencing (RNA-seq).

Microarray: Microarray analysis was the first high-throughput technology to measure the expression of thousands of genes simultaneously [32]. The main limitation of microarray analysis is that it examines only genes for which there are probes on the chip and thus cannot detect novel RNA transcripts [33].

Next-generation sequencing: Next-generation sequencing technology [34-37] is a major breakthrough to extract biological information from DNA and RNA samples. Unlike the microarray technology, RNA-seq is not limited to a certain set of probes, since RNA-sequences generated from a RNA sample are mapped to its unique region of the genome. Thus, novel transcripts can be detected [38].

1.2.2 Biological network

Biological networks [16, 31, 39-42] are the graphical representation of probabilistic or associative interactions (edges) between biological components (e.g., DNA, genes, proteins, metabolites (network "nodes")) that together characterize a biological system. In this thesis, we will study networks with edges between genes that are either undirected (representing co-

expression associations between the nodes) or directed (representing regulatory or probabilistic gene interactions (e.g., Bayesian networks)).

1.2.2.1 *Co-expression gene network*

A co-expression gene network is an undirected graph in which each network node has at least one edge to another node, as the node-pair is significantly co-expressed (**Figure 5A**). With expression profiles from several RNA samples, co-expression similarity for each gene pair can be computed. Co-expressed genes are biologically meaningful because nodes in a co-expression network collectively represent biological functions or pathways. Numerous computational methods and tools have been developed to construct co-expression gene networks [43-45]. Such methods measure the expression similarity between each gene pair in the data by using a method (Pearson or Spearman) to rank correlations that then are represented in a topology overlap matrix (TOM), where weak (i.e., non-significant) correlations are removed.

1.2.2.2 *Regulatory gene network*

A regulatory gene network (RGN) is a directed graph where edges between nodes/genes are directed from the regulatory/source node to the regulated/target node (**Figure 5B**). Thus, the variation in expression of the source node causes variation in expression in the target node, and not the other way around. Again, various computational methods are available to construct RGNs [11, 30, 46-50]. Details are further explained in "Materials and Methods" section (see below, pages 18, 23).

1.2.2.3 *Tissue-specific versus cross-tissue gene networks*

As the name implies, tissue-specific networks are derived from data isolated from a single tissue. However, complex multifactorial diseases like CAD are driven by molecular processes that are affected by multiple tissues. For example, cholesterol metabolism is regulated by the liver (synthesis and uptake) and adipose tissue and skeletal muscle (lipolysis). Therefore, it makes sense to also seek gene networks operating across different tissues. With studies like STAGE, we are fortunate to have gene expression data from multiple tissues of each CAD patient, allowing us to reconstruct network of genes also across tissues into so-called cross-tissue (CT) gene networks. In CT gene networks, each node corresponds to a gene-tissue pair. Thus, a given gene might occur more than once in a network if it originates from different tissues. A typical CT gene network looks like a bipartite graph (**Figure 5C**). Of note, CT correlations can only be calculated based on common samples between two tissues (**Figure 5D**). As a consequence, the sample size becomes smaller when assessing CT edges, further weakening correlations that are already weak, likely because intermediate CT network nodes are not detectable in the gene expression data connecting separate tissues such as signaling molecules in plasma and lymph and also because of neuronal signaling. We therefore reasoned that the cut-offs for significant CT gene-gene interactions (i.e., edges) need to be set at a more relaxed level than significant gene-gene interactions (i.e., edges) within a single tissue (see below, "Materials and Methods", pages 19-20).

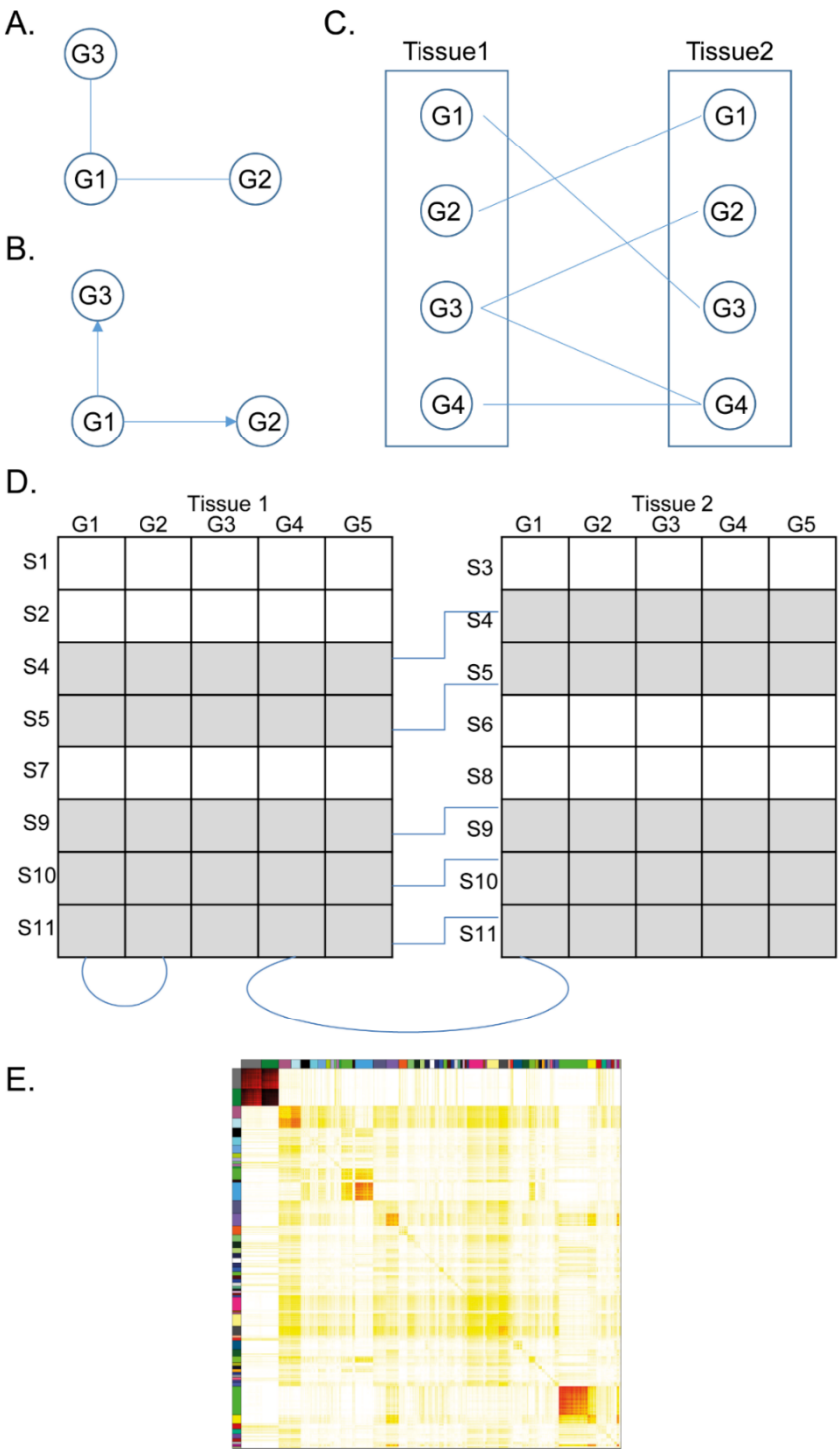

Figure 5: Schematic gene network and cluster. ***(A)*** *Co-expression gene network, where gene G1 is significantly co-expressed with genes G2 and G3.* ***(B)*** *RGN, where gene G1 is a regulator and regulates its target genes G2 and G3.* ***(C)*** *Typical CT connections between two tissues.* ***(D)*** *Correlation measurements; shaded samples are common in both tissues; therefore, for CT correlations we considered only those samples. Small arc shows TS correlations; big arc shows CT correlations* ***(E)*** *Gene clusters, a heatmap of gene-gene correlation. In top and left color-coded bars, each color represent a co-expression cluster (i.e., network). The diagonal indicates that the significantly co-expressed genes form co-expression clusters/networks.*

1.2.2.4 *Gene co-expression clustering*

The human genome consists of around 20,000 coding genes in each tissue and cell type to allow them fulfill their role in the human body. It has been observed that when genes are functionally associated (frequently termed a "gene module"), they are also to some extent co-expressed and therefore can be captured by co-expression clustering. Various algorithms and tools are available to perform co-expression clustering [43, 51]. We used weighted gene co-expression network analysis to generate co-expression clusters (**Figure 5E**).

1.2.3 Expression quantitative trait locus

Besides environmental influences, genetic variations affect gene expression [52-55]. Genomic regions (i.e., SNPs) that are associated with gene expressions are termed expression quantitative trait loci (eQTLs). When an eQTL affecting the expression level of transcript is located in the same region of the gene, it is called a *cis*-eQTL (i.e., SNP affecting the transcript is within 1 Mbp up- or downstream of the transcriptional starting site of the gene), (**Figure 6A**). In contrast, when an eQTL affects the expression level of transcript/gene outside 1 Mbp or on another chromosome it is termed a *trans*-eQTL (**Figure 6B**). Studies in mice, rats, and human cells and tissues have shown that ≥30% of variation in gene expression is influenced by eQTLs [55-57]. Genomic regions containing eQTLs affecting hundreds of genes are called "hot spot" regions [58].

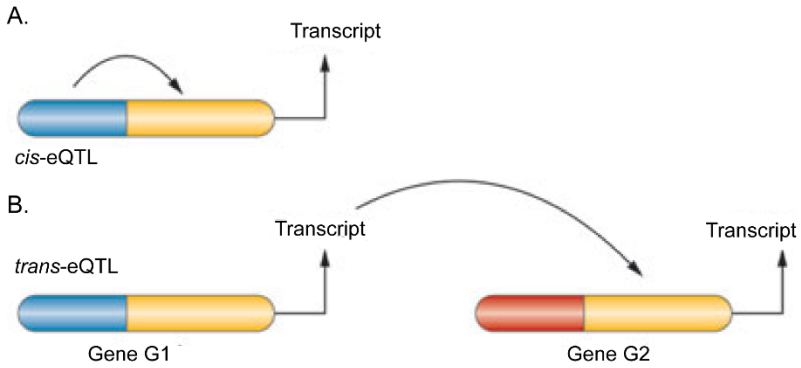

Figure 6: *cis*- and *trans*-eQTL Modified from [13]

1.2.4 Genome-wide association studies

Genome-wide association studies (GWAS) identify genomic loci where DNA variance is associated with a trait or disease. GWAS are most commonly performed in thousands of patients with a given disease who are compared to population controls assumed to have no disease. In this fashion, SNPs that occur more frequently in people with a particular disease (i.e., cases) than in people without the disease (i.e., controls) can be identified [59, 60]. GWAS have been performed for hundreds of human traits and diseases [61, 62]. Specifically for CAD, the largest current GWAS combine data from over 194,000 individuals and have robustly identified 150 loci associated with increased risk of CAD [14].

GWAS Catalog: The GWAS Catalog is a collection of all published SNP-trait associations with P-values < 1.0 x 10^{-5} identified by GAWS [61]. As of 1 August 2016, the GWAS Catalog documented 24069 unique SNP-trait associations from 2512 studies.

1.2.5 Causal network

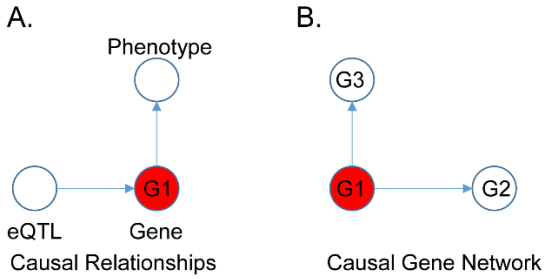

Figure 7: Causal relationships and causal gene network

According to the central dogma in molecular biology and in the heart of systems genetics, causal genes are defined as genes that either are known to be regulatory (i.e., transcription factors) or are regulatory in other ways, such as by acting through eQTLs. The causal nature of these genes propagate to downstream genes in signaling pathway that in turn affects other biochemical pathways [11, 49]. Following this principal, "CAD-causal" modules/networks are distinguished (as opposed to reactive or independent networks) (**Figure 7**) as follows:

- Networks containing one or more CAD candidate genes mapped to genome-wide significant loci for CAD by GWAS or
- Networks containing more eQTLs whose underlying SNPs are significantly associated with CAD (P-value < 0.05) according to the summary dataset of case-control CARDIoGRAM meta-analysis of CAD GWAS [14] than expected by chance (i.e., roughly 5 % of the eQTLs in a network are expected by chance to have a GWAS P-value < 0.05).

1.2.6 Clinical cohort

To unravel the complexity and causes of common complex diseases (CCDs) like CAD, more integrative biology is needed. The analysis of genetics-of-gene-expression (GGE) datasets has already proven useful for identifying disease-linked networks and, within these networks, novel candidate genes [9, 49, 63-65].

In NEW biology, the first major step is to collect clinical cohorts with intermediate phenotypes that besides DNA and clinical characteristics are RNA from disease-relevant tissues. Early GGE studies used blood samples, as these are easily obtainable. However, for complex diseases like CAD vascular and metabolic tissues are equally important. The Genotype–Tissue Expression (GTEx) [66, 67] is one of the largest GGE datasets, containing 1641 samples across 43 tissues collected from 175 individuals. A possible limitation of this cohort is that samples are collected postmortem. The Stockholm Atherosclerosis Gene Expression (STAGE) dataset [68, 69], used in this thesis, is a unique CAD GGE study that considers seven vascular and

metabolic tissues of well-characterized CAD patients. Two vascular and four metabolic tissues and blood were collected during coronary artery bypass grafting surgery from 121 well-characterized CAD patients. STAGE is now followed by The Stockholm-Tartu Atherosclerosis Reverse Network Engineering Task (STARNET) study [70], in which the same samples are gathered from 600 CAD patients.

2 AIM OF THE THESIS

The main aim of this thesis is to design and implement a computational pipeline using the unique GGE STAGE study to, for the first time, allow discovery of RGNs with key driver's across seven vascular and metabolic tissues acting to cause CAD. More specifically:

I. Develop and implement a computaional pipeline to identify and validate RGNs causal for CAD and some of its risk factors by primarily using the STAGE study.
II. Develop and validate a CT-weighted gene co-expression network analysis (X-WGCNA) method.
III. Reconstruct human TF-regulatory gene networks from genes involved in regression of atherosclerosis in mice.
IV. To infer eQTLs from the STAGE study and define key regulatory eQTLs acting across vascular and metabolic tissues and with these assocaited gene sets.

3 MATERIALS AND METHODS

3.1 GENETICS OF GENE EXPRESSION COHORT

3.1.1 STAGE

In the Stockholm Atherosclerosis Gene Expression (STAGE) study, seven vascular and metabolic tissues of well-characterized CAD patients were sampled during coronary artery bypass grafting (CABG) surgery [68, 69, 71]. Patients eligible for CABG were included but those with severe systematic non-CAD diseases were excluded. Tissue samples (**Table 1**) were obtained from atherosclerotic arterial wall (AAW), internal mammary artery (IMA, nonatherosclerotic arterial wall), liver, skeletal muscle (SM), subcutaneous fat (SF), visceral fat (VF), and blood.

Table 1: Number of genotype and global gene expression samples in STAGE

	AAW	**IMA**	**Liver**	**SM**	**SF**	**VF**	**Blood**
AAW	73 (68)	57	59	57	48	62	65
IMA		88(79)	68	70	58	77	77
Liver			87(77)	71	56	77	75
SM				89(78)	61	78	76
SF					72(63)	61	60
VF						98(88)	87
Blood							105(102)

The STAGE study comprises a total of 121 patients with global gene expression data from up to 7 tissues whereof 109 also were genotyped. The main diagonal of the table shows the numbers of samples with global gene expression data and in brackets the number of samples with both gene expression and genotype data. The off-diagonal numbers indicate the common number of gene expression samples between all pairs of tissues.

DNA from 109 patients with sufficient quantities and qualities (≥1 µg, 1.7 > 260/280 > 1.9 with Nanodrop, Agilent) were genotyped with the GenomeWideSNP_6 array (Affymetrix). Allele frequencies for 909,622 single-nucleotide polymorphisms (SNPs) were determined with the Birdseed algorithm in Affymetrix Power Tools (v 1.14.2); 530,222 autosomal SNPs with call rates of 100% and minor allele frequency >5% and in Hardy-Weinberg equilibrium (P-value > 10^{-6}) were used for downstream analysis (the "QC SNP set") [71].

Custom-made HuRSTA-2a520709 arrays (Affymetrix) were used for gene expression profiling of total RNA samples (≥5 µg, 1.95 > 260/280 > 2.05 with Nanodrop, Agilent) from 121 patients according to the manufacturer's instructions [69].

Phenotypic measurements, mostly standard plasma measures, were collected as described [68] and are shown in **Table 2**.

Table 2: Basic characteristics of the STAGE patients

Plasma cholesterol (mmol/L)		Plasma triglycerides (mmol/L)	
Total	4.09 ± 1.01	Total	1.41 ± 0.73
VLDL	0.32 ± 0.25	VLDL	1.04 ± 0.67
HDL	1.50 ± 0.29	HDL	0.16 ± 0.05
LDL	2.10 ± 0.79	LDL	0.26 ± 0.09
Patients (n)	124	CRP (mg/L)	8.8 ± 2.93
Age (years)	66 ± 8	HbA1c	5.24 ± 1.41
Gender (male)	110 (89%)	Diagnoses and therapies	
BMI (kg/m^2)	27 ± 3.7	Hypertension	74 (60%)
Waist/hip ratio (m)	0.94 ± 0.06	Beta-blockers	103 (83%)
Blood pressure (mm Hg)		Hyperlipidemia	86 (69%)
Systolic	141 ± 19	Lipid lowering	101 (81%)
Diastolic	80 ± 9.1	Diabetes mellitus	25 (20%)
Smokers	8 (7%)	Insulin	23 (19%)

Values are mean ± SD or n (% of all STAGE patients)

To determine the extent of atherosclerosis, all CABG patients underwent preoperative biplane coronary angiography (Judkins technique). Angiograms were evaluated with quantitative coronary angiography (QCA) techniques. The left and right coronary arteries and their branches were divided into segments [72]. Each segment was measured during end-diastole. Two measurements, stenosis score and segment area score, were calculated as explained below to define the degree of atherosclerosis:

i. **The stenosis score** measures stenosis in each patient and calculates the percentage diameter stenosis (%DS)—the reduction in lumen diameter caused by the plaque (see **Figure 8**). Stenosis was categorized as mild or severe, as shown below. A total occlusion is 100%DS. The score is calculated as 1 x #mild stenosis + 2 x #severe stenosis.

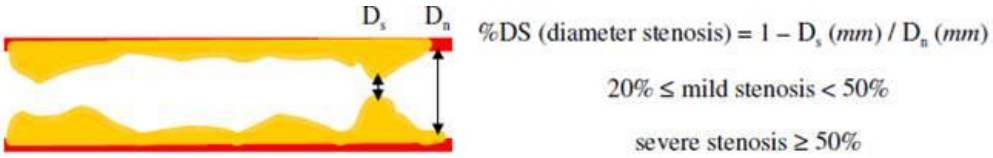

Figure 8: Stenosis score measurement

ii. **The segment area score** measures the amount of atherosclerosis in all the vessels in a patient. The measurement is the total percentage of plaque area in a segment (see **Figure 9**). To obtain a final score, the average percent stenosis over all the segments is calculated.

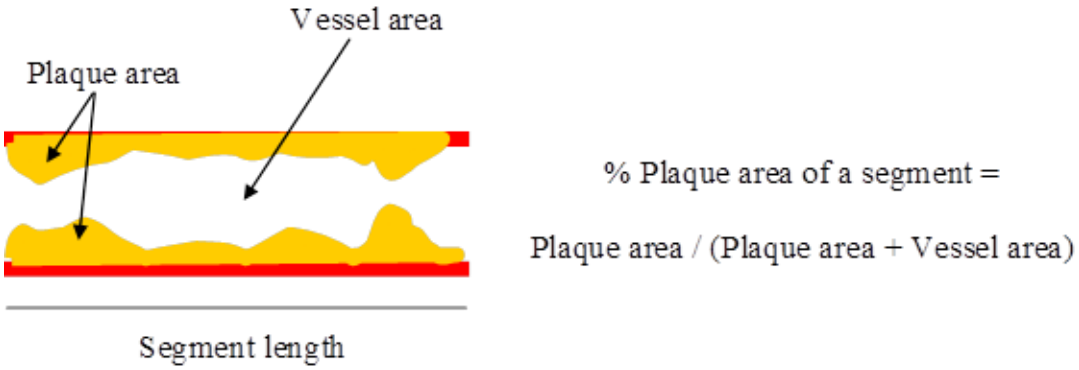

Figure 9: Segment area score measurement

Clinical phenotypes strongly and robustly associated with CAD were classified in four groups:

I. **Extent of atherosclerosis**
   1. Stenosis score
   2. Segment area score

II. **Cholesterol:** The following plasma levels considered as type of cholesterol associated with CAD,
   1. Total cholesterol (TC)
   2. Very low density lipoprotein (VLDL)
   3. Low-density lipoprotein (LDL)
   4. High-density lipoprotein (HDL)

III. **Glucose**: The following four measurements considered as type of glucose associated with CAD [73, 74]:
   1. Plasma glucose (fasting)
   2. HbA1c
   3. Pro-insulin
   4. Insulin

IV. **C-reactive protein (CRP)**: Inflammation is intimately and robustly linked to CAD and that is why we and others use CRP as a marker of inflammation in the context of CAD. [75, 76] These study shows CRP provides improved method of identifying persons at risk for cardiovascular diseases.

3.1.2 SÖS

SÖS is an extension of the STAGE study in which 39 CAD patients sampled during carotid stenosis surgery at the Södersjukhuset (SÖS) hospital in Stockholm [68, 77]. Patient exclusion criteria were same as STAGE (i.e., free from any other severe systematic diseases). Tissue samples were collected from carotid lesions (n = 25) as well as blood. From blood monocytes were isolated using Ficoll separation [78] and differentiated into macrophages (n = 36) *in vitro*. Gene expression profiling was done with the same platforms as in the STAGE study. **Table 3** shows basic characteristics of the STAGE SÖS patients.

Table 3: Basic characteristics of the STAGE SÖS patients

Plasma cholesterol (mmol/L)		Plasma triglycerides (mmol/L)	
Total	4.56 ± 1.09	Total	1.30 ± 0.49
VLDL	0.27 ± 0.19	VLDL	0.85 ± 0.41
HDL	1.70 ± 0.42	HDL	0.20 ± 0.07
LDL	2.45 ± 0.85	LDL	0.30 ± 0.09
Patients (n)	38	Diagnoses and therapies	
Age (years)	69 ± 10	Hypertension	16 (42%)
Gender (male)	25 (66%)	Beta-blockers	20 (53%)
BMI (kg/m^2)	25 ± 3.2	Hyperlipidemia	7 (18%)
Waist/hip ratio (m)	0.91 ± 0.07	Statins	25 (66%)
IMT, mm (mean ± SD)	1.20 ± 0.16	ACE inhibitors	12 (32%)
		Calcium channel blockers	7 (18%)
		Loop diuretics	4 (11%)

Values are mean ± SD or number (%) of patients. ACE indicates angiotensin-converting enzyme; IMT, intima-media thickness;

3.1.3 HMDP

Hybrid mouse diversity panel (HMDP) [53, 79] is a global gene expression datasets from DNA and RNA samples isolated from 105 strains of mice fed regular, high fat diet and cross-bred to the atherosclerosis-prone ApoE-Leiden background. Here are the definitions of different subgroups of mice from the HMDP:

- Healthy male mice fed a chow diet for 16 weeks and then sacrificed.
- High-fat diet mice were fed a chow diet for 8 weeks and then put on a high-fat, high-sucrose diet for 8 weeks before sacrifice.
- Atherosclerosis-prone mice where the 105 strains were bred onto the C57/BL6 background and were made transgenic for ApoE-Leiden and ApoB. These mice were fed a chow diet for 8 weeks and then fed a high-fat diet for 16 weeks before sacrifice.

Relevant to STAGE tissues/expression profiles were mouse tissues sampled from the aorta, liver, adipose and heart. From the atherosclerosis-prone ApoE-Leiden mice, atherosclerotic aortic arch samples were collected for gene expression profiling and for assessment of the extent of atherosclerosis. The phenotypes of the HMDP mice were similar to those obtained in STAGE.

3.2 ATHEROSCLEROSIS MOUSE MODEL

3.2.1 Mouse model

The *Ldlr*$^{-/-}$*Apob*$^{100/100}$*Mttp*$^{flox/flox}$*Mx1-Cre* mouse model [80, 81] used to study atherosclerosis regression. These mice have a plasma lipoprotein profile (*Ldlr*$^{-/-}$*Apob*$^{100/100}$) resembling that of familial hypercholesterolemia causing advanced and rapid atherosclerosis formation. These mice also have a genetic switch (*Mttp*$^{flox/flox}$*Mx1-Cre*) to lower plasma cholesterol at any time point during atherosclerosis progression.

3.2.2 Mouse dataset

Total RNA was isolated with an RNeasy Mini-kit with a DNAse I treatment step (Qiagen). RNA quality was assessed with a Bioanalyzer and RNA quantity with a Nanodrop. Mouse Gene 1.0 ST arrays (Affymetrix) were used for global mRNA expression profiling [81].

3.3 DATA PRE-PROCESSING

3.3.1 Data normalization

Robust multi-array average (RMA) was used for background correction, normalization, and summarization of raw microarray data through Affymetrix Power Tools, v 1.14.2. A custom-made Chip Description File (CDF) was used to match 381,707 probes on the array to 19,610 probe sets for unique genes (to avoid cross-hybridization between alternative transcripts) according to the hg19 human genome assembly [69].

3.3.2 Principal component analysis

To ensure the quality of data, we analyzed 612 mRNA profiles (considered as 19610-dimensional vectors) by multidimensional scaling using the Euclidean distance and Sammon's nonlinear mapping criterion [71].

3.4 GWAS DATASETS

To define causality for CAD, we used the CARDIoGRAM [14], WTCCC [82], and MIGen [83] GWAS datasets. For causality in CAD risk factors, we used the fasting glucose [84], blood lipids [85], HbA1c [86], and pro-insulin [87] GWAS datasets. In addition, we also used the GWAS Catalog [61, 88] to identify CAD GWAS candidate genes.

3.5 COMPUTATION ANALYSIS OF A SET OF GENES OF INTEREST

3.5.1 Reconstruction of TF-RGNs

To reconstruct TF-RGNs, we used the context likelihood of relatedness (CLR) method [46, 89] with co-expression similarities measured by Pearson correlation. After reconstructing TF-RGN, we used the data processing inequality (DPI) [90] technique to remove weak connections. The TF-RGN analysis flowchart is shown in **Figure 10**.

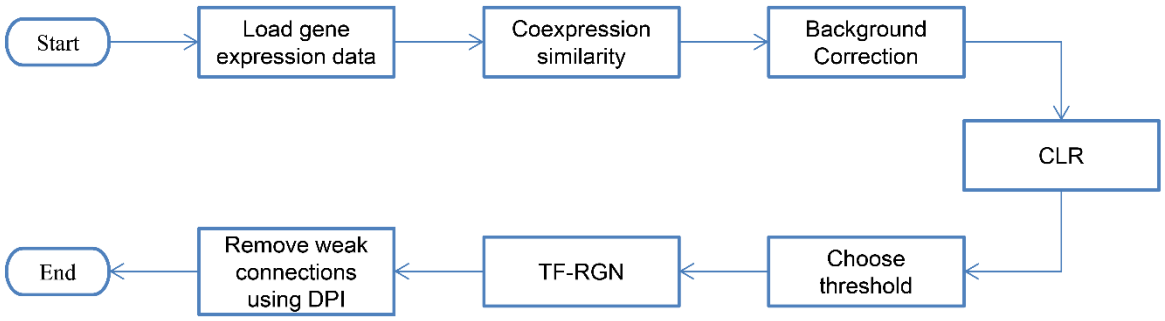

Figure 10: TF-RGN flowchart

The TF-RGN reconstruction algorithm was implemented in C++ and is available at https://github.com/hustal/TF-RGN.

3.5.2 Key driver identification, part I

In the first part of my thesis, I used a simplified definition of "key drivers" as network nodes belonging to the top 15% highly connected nodes. Subsequently, I used a more stringent and sophisticated way of defining key disease drivers in RGNs (see section below, "**Key driver identification, part II**", page 24).

3.6 COMPUTATION ANALYSIS OF GENETICS OF GENE EXPRESSION DATA

For the analysis GGE data, we established an overall computational pipeline (**Figure 11**). The individual steps are explained below.

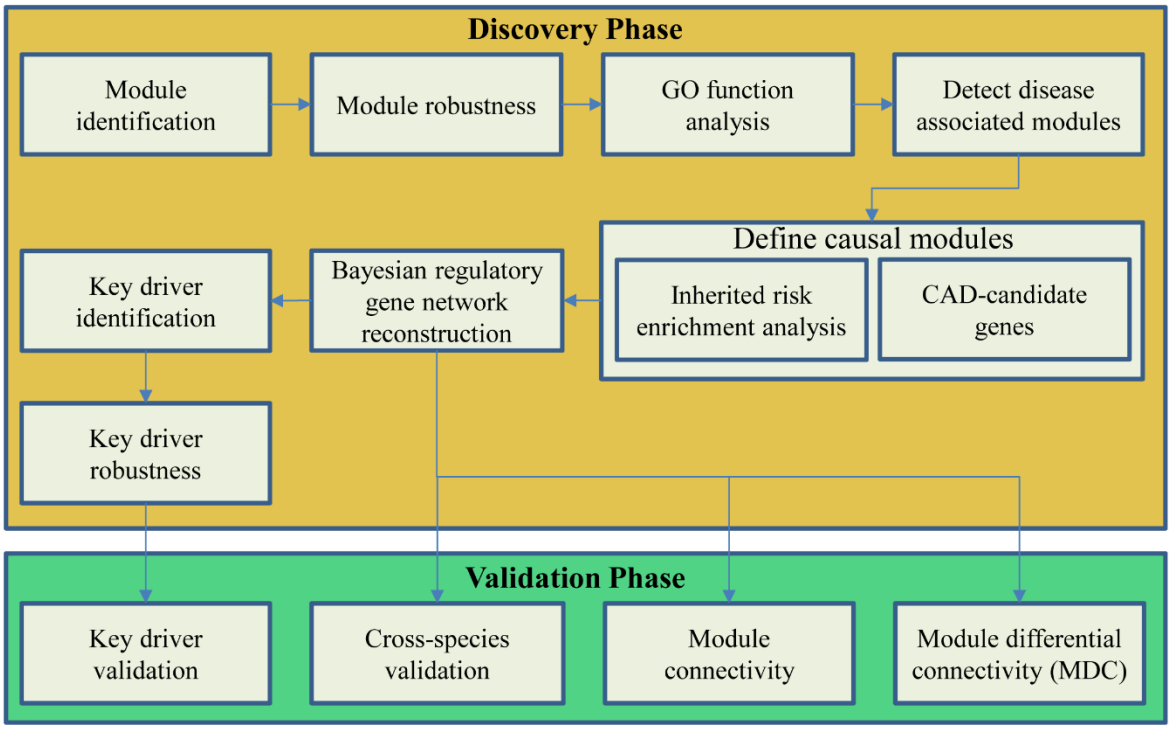

Figure 11: Computational pipeline for GGE analysis. *(top) Yellow color shows steps in discovery phase and (bottom) green color shows steps in validation phase.*

3.6.1 Module identification

To identify co-expressed genes in functionally associated modules (active in and across tissues) from global gene expression data, we used the CT-weighted gene co-expression network analysis (X-WGCNA) method, which we modified from the original weighted gene co-expression network analysis (WGCNA) method [43, 44] in Paper II.

In brief, X-WGCNA takes as input a set of normalized gene expression matrices, one for each tissue, where rows indicate samples (i.e., individuals) and columns indicate gene symbols. For each tissue, the most variant genes are selected on the basis of their standard deviation across the samples in each tissue. We either used a standard deviation or a "number of genes" cut-off. Then, an adjacency matrix A was calculated across all selected expression traits (i.e., gene-tissue pairs) as

$$A_{ij} = \begin{cases} |C_{ij}|^{\beta_1} & \text{if } i \text{ and } j \text{ belong to the same tissue} \\ |C_{ij}|^{\beta_2} & \text{if } i \text{ and } j \text{ belong to different tissues} \end{cases}$$

In WGCNA [43], the adjacency matrix is calculated using a single parameter $\beta = \beta_1 = \beta_2$. In X-WGCNA [Paper-II], the parameters β_1 and β_2 are determined independently to obtain scale-free tissue-specific as well as cross-tissue subnetworks, as measured by fitting index R^2 of the linear model which regresses logP(k) on log(k), where k ranges over the degree (i.e., weighted number of connections) values in the various subnetworks and P(k) is the frequency distribution of k. The next steps of X-WGCNA are identical to WGCNA: the topological overlap matrix (TOM) for A is calculated as described by [43], genes are grouped by average

linkage hierarchical modelling of the TOM, and a dynamic tree cut algorithm was used to cut the modelling dendrogram into gene clusters [91]. Finally, a user-defined threshold (default 95%) is used to define tissue-specific (percentage of genes from the same tissue exceeding the threshold) and cross-tissue (otherwise) clusters. X-WGCNA is implemented as an R package and is available from https://github.com/hustal/X-WGCNA. A summary of the workflow is shown below in **Figure 12**.

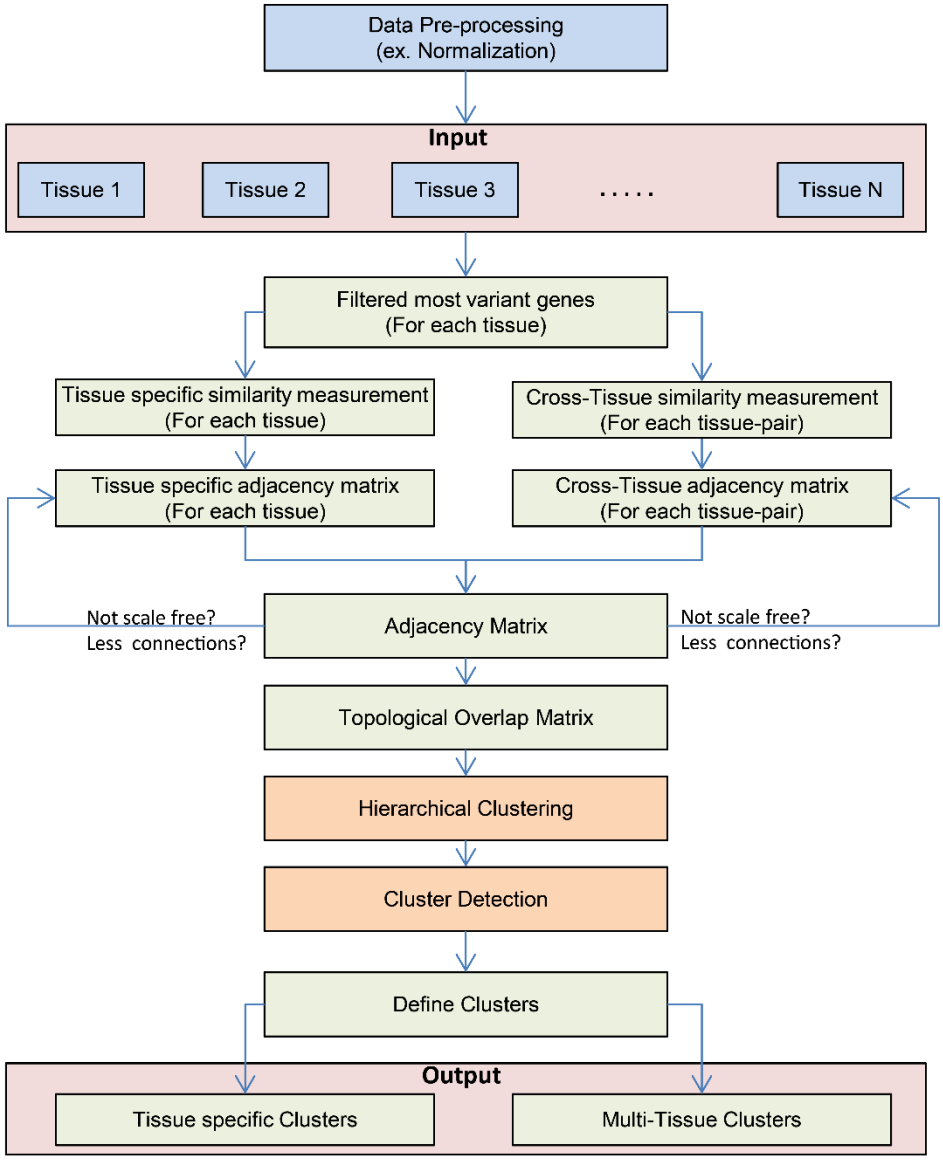

Figure 12: X-WGCNA workflow. *Blue colored steps are out of scope of X-WGCNA, for which any tool or software could be use. Lighter red color shows input and output state. Green color shows the steps which will execute by X-WGCNA. For orange colored steps one could use X-WGCNA or any other tool.*

3.6.2 Cross-tissue modules robustness

To test the robustness of the identified CT modules, we repeated the co-expression analysis seven times, each time removing the expression data from one tissue. For each tissue, we

compared the gene content of the original modules (excluding genes only expressed in the removed tissue) to that of the new modules.

3.6.3 GO function analysis

To assess the enrichment of modules in molecular functions, biological processes, and cellular components, we used Gene Ontology (GO) according to Bingo [92]. Fisher's exact test, and the Benjamini-Hochberg FDR method [93] were used to assess the statistical significances of GO-enriched modules (GO data version: 2014-01-29) [94, 95].

3.6.4 Disease-associated modules

To identify possible disease associations with the identified co-expression modules, we used two methods to detect both linear and stepwise phenotype associations with each module's gene expression levels. In this fashion, module associations with 4 main CAD phenotypes (described in section 3.1.1, page 15) were analyzed.

i. Non-linear associations: For a given co-expression module, patients who donated samples of all tissues represented in the module were grouped by K-means clustering. Next, for the two most distinct patient groups the extent to which one of the 4 main CAD phenotypes differed was assessed with a rank-sum test (P-value < 0.05)

ii. Linear associations: Again for the patients who donated samples of all tissues represented in the module, we determined non-zero Pearson correlations between the levels of any of the 4 main CAD phenotypes and the first principal component of the data matrix generated from the module gene expression levels (i.e., the "module eigengene" [96]).

We combined the P-values for both these statistical tests into a single P-value by using a weighted data integration method to maximize the overall statistical power of the combined associations [97]. To correct for multiple testing, we used Story's method to estimate a positive false discovery rate (FDR) for each P-value in the resulting table of module-phenotype associations [98]. An FDR threshold of 20% (corresponding to a nominal P-value threshold of 0.03) was used to define true module-phenotype associations. The individually most significant clinical measurement for each phenotype association was recorded, and the association was considered established only if the combined P-value < 0.03 (FDR = 20%).

3.6.5 Identification of causal modules

As alluded to in section 1.2.5, page 10, we defined "CAD-causal" modules as those either containing:

i. One or more of the 53 CAD candidate genes mapped to the lead SNPs within CAD loci having a genome-wide significance of P-value < 1.0×10^{-8} according to GWAS of CAD [14] or

ii. A larger-than-expected-by-chance number of eQTLs with CAD association (P-value < 0.05) according to the SNP summary dataset of the case-control CARDIoGRAM meta-analysis GWAS. The schematics of the risk-enrichment analysis is shown in **Figure 13**.

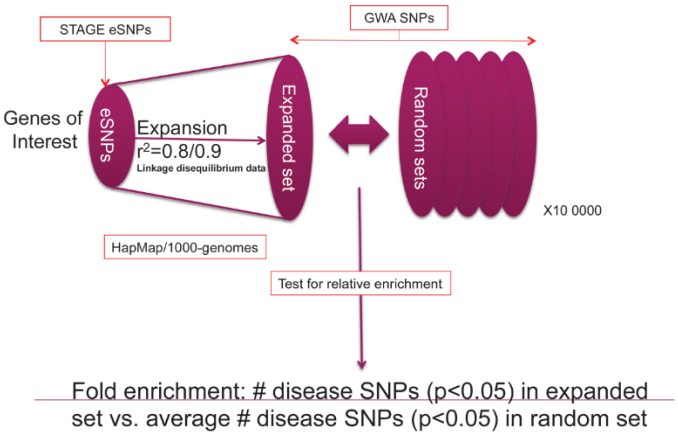

Figure 13: Schematic of inherited risk enrichment analysis

In brief, the set of eQTLs used to match module genes was that previously calculated from the STAGE cohort [71]. After matching individual eQTLs to module genes, each set of module eQTLs was first expanded using the 1000 Genomes to include SNPs in strong LD ($r^2 > 0.9$). The enrichment of CAD association (P-value < 0.05) according to the CARDIoGRAM meta-analysis dataset of the resulting expanded SNP set was then compared to the average enrichment of CAD association in 10,000 randomized, equal sized and chromosomal distributed SNP sets from the same data. The fold-enrichment was calculated by comparing the number of disease associated SNPs ($N^{real}_{\text{P-value} < 0.05}$) in the expanded SNP set with the average number of disease-associated SNPs in the 10,000 random set ($\bar{N}^{rand}_{\text{P-value} < 0.05}$). As the null distributions approximately followed a normal distribution, we defined Z-statistics as follow:

$$Z = \frac{N^{real}_{\text{P-value} < 0.05} - \bar{N}^{rand}_{\text{P-value} < 0.05}}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left|N^{i}_{\text{P-value} < 0.05} - \bar{N}^{rand}_{\text{P-value} < 0.05}\right|^2}} \quad , \qquad n = 10{,}000$$

The P-values were calculated from $\Phi(Z)$, standard normal cumulative probabilities [71, 99].

For modules with less than 10 eQTLs, we instead mapped SNPs of the GWA dataset to a region of ± 500kb of the transcription start or end site of a gene. Next, for G genes in a module, P_g for each gene using S mapped SNPs ($P_1 < P_2 < \cdots < P_s$) the CAD association P-value for each gene was calculated using Simes' combination test [100] as following:

$$P_g = \min_k \{\frac{S.P_k}{k}\},$$

Where, k is the rank of sorted P_S. The enrichment of CAD association (P-value < 0.001) was then calculated as described for eQTLs above.

3.6.6 Reconstruction of regulatory gene networks

Bayesian regulatory-gene networks (RGNs) were inferred taking into account eSNP and other prior information to derive the most probable causal interactions from undirected co-expression associations. Briefly, given a directed acyclic graph (DAG) *G* between *N* expression traits, the joint distribution of their expression levels x_i ($i = 1,...,N$) is assumed to take the form

$$p\,(x_1, \ldots\ldots, x_N \mid G) = \prod_{i=1}^{N} p\left(x_i \mid \{x_j : j \in Pa_i\}\right),$$

where Pa_i denotes the set of parent nodes of node *i* in the graph *G*. We further assume that the RGN is a linear Gaussian network [101] such that the conditional distributions are given by

$$p\left(x_i \mid \{x_j : j \in Pa_i\}\right) = \mathcal{N}\left(\alpha_i + \sum_{j \in Pa_i} \beta_{ij} x_j; \sigma_i^2\right),$$

where $\mathcal{N}(\mu;\sigma^2)$ denotes a normal probability distribution with mean μ and standard deviation σ. The parameters α_i, $\beta_{i,j}$ and σ_i are to be determined along with the graph structure *G*.

Given a dataset $D = (x_{im})_{im}$ of expression levels for *N* traits in *M* samples, assumed to be drawn independently from the distribution, the likelihood of observing the data given a DAG *G* is given by

$$p\,(D \mid G) = \prod_{m=1}^{M} \prod_{i=1}^{N} p\left(x_{im} \mid \{\, x_{jm} : j \in Pa_i\}\right).$$

Using Bayes' theorem we can therefore write the likelihood of observing *G* given the data *D* as

$$P\,(G \mid D) = \frac{p(D \mid G)\, P(G)}{Z},$$

where $P(G)$ is the prior probability of observing *G* and *Z* is a normalization constant. Expression traits with cis-acting eQTLs or known transcription factors (TFs) are more likely to act as causal regulators of other expression traits and this information can be encoded in the prior probability $P(G)$ to reconstruct causal networks [102]. Here we imposed a hard prior that only expression traits with eQTLs, GWAS candidate genes or known TFs are allowed to be parents of any other traits (i.e. $P(G) = 0$ if *G* contains an edge having a non-eQTL, non-TF or non-GWAS gene as its source node, and constant otherwise). A locally optimal DAG was then found starting from a random graph by randomly adding, removing and reversing edges until the likelihood no longer improves. Maximum-likelihood values for the model parameters α, β and σ are learned along with the graph structure by linearly regressing a node on its current parents.

Because it is computationally and statistically infeasible to reconstruct RGNs with 20,912 nodes (total number of expression traits in the co-expression network) with a hundred samples or less, we imposed an additional constraint in the structure prior $P(G)$, where traits were imposed to take their parent nodes from among traits with which they share a co-expression module. This prior effectively breaks down the 'large' problem of reconstructing a DAG on

the entire set of traits to a set of independent 'small' problems of reconstructing a DAG for each co-expression module, and significantly reduces the number of parameters to be estimated from the data.

Here we employed the Bayesian Information Criterion (BIC) to score models and a multiple restart greedy hill-climbing algorithm, using edge additions, deletions, and reversals to search locally optimal DAGs for each co-expression module [103].

Algorithm parameters were set to solve $25*N^2$ regression problems during each run of the search algorithm, where N is the number of traits in a module. A final consensus causal network was constructed by considering all edges that appeared in 30% or more of the locally optimal DAGs found during each run of the search algorithm. The consensus network was made acyclic by iteratively removing the weakest supported edge from every cycle in the network.

3.6.7 Key driver identification, part II

Key drivers—the key regulators, or upstream genes, in a gene network—were also identified according to a published method [104]. In its simplest form, a key driver of an RGN is its most highly connected regulator gene (see **Key driver identification, part I, page 18)**, the one with the highest number of outgoing edges. However, a gene that regulates one or more highly connected regulators is in fact more influential than the downstream regulators, even if it is not itself highly connected. Therefore, in the second key driver analysis, we also considered how many network genes can be reached from a given regulator in up to $H+1$ steps, where H is the parameter of the method. The value $H=1$ (i.e., considering direct neighbors and two-step neighbors) was used as the default value in the key driver software that has been implemented in R.

3.7 VALIDATION OF REGULATORY GENE NETWORKS

3.7.1 Cross-species validation

As briefly described above (page 16), to assess the cross-species conservation of the identified STAGE modules associated with any of the 4 main CAD phenotypes, we used the HMDP gene expression and phenotype data [53, 79]. First, we identified mouse orthologs of the human genes in the CAD-associated STAGE modules. Then, identical to the STAGE CAD phenotype-module associations (see above, page 21), we sought mouse gene expression and phenotype linear and nonlinear associations between gene expression matching STAGE and HMDP tissues (atherosclerotic aorta, adipose, SM-heart, liver) and matching phenotypes in STAGE and HMDP (i.e., extent of atherosclerosis assessed from angiograms (STAGE) and the aortic root (HMDP) as well as measures of plasma cholesterol and glucose). Since mice do not express CRP, CRP-associated CAD-causal modules were not validated.

3.7.2 Key driver validation

Key drivers identified in RGNs of the atherosclerotic arterial wall were validated by siRNA silencing in THP-1 macrophages, which were then incubated with Ac-LDL to induce foam cell

formation, as described [69, 105]. Gene expression data from siRNA-treated and control THP-1 cells were generated with Agilent Human Custom Gene Expression Microarray 8x15K, containing the three module genes (in total 245 unique genes from the modules 42, 58, and 98) (spotted in triplicate), according to the manufacturer's instructions. The quantile normalization method was used to normalize the data, and differentially expressed genes were identified by maintaining P-value (< 0.05) and FDR (< 10%) with the Benjamini-Hochberg procedure [93] through the 'limma' package in R.

The probability that a key driver was specific for its network rather than not affecting it more than expected by chance, or affecting all three networks indiscriminately, was calculated by using hypergeometric distribution P-values in R, with the expression of all module genes as background.

3.7.3 Module connectivity

To assess module connectivity, we compared the total connectivity (sum of adjacency values, i.e. weighted correlation coefficients using scaling parameter beta = 6) of module genes in independent data to the total connectivity of 10000 random gene sets of the same size. We confirmed that the random connectivities were normally distributed and used their Z-score and P-values from the normal distribution to test if the real connectivity value deviated significantly from the random ones.

3.7.4 Module differential connectivity

Module differential connectivity (MDC) between pairs of similar tissues (i.e., AAW and IMA) was calculated as the ratio of total module connectivity (sum of adjacency values, i.e. weighted correlation coefficients using scaling parameter beta = 6, between all pairs of module genes).

3.8 RECONSTRUCTION OF SUPER NETWORK

To construct a super-network of the CAD-causal modules to assess how causal modules may communicate, we calculated the Pearson correlation (considering common samples between module pairs) between all "module eigengenes" [96] and keeping all edges with absolute correlation threshold, r > 0.40 and P-value < 0.05.

4 RESULT

4.1 PAPER I

The main goal of the Paper I was to establish and implement a pipeline (**Figure 11, page 19**) to reconstruct and validate CAD RGNs and key drivers from GGE cohort. Key findings of this paper are shown in Box1.

We used 20,912 (7703 unique) of the most variant genes from 612 gene expression profiles in the seven STAGE tissues for CT-weighted gene co-expression network analysis. One hundred seventy-one modules (94 tissue-specific and 77 cross-tissue) were identified; 33% of the module genes were previously related to atherosclerosis or CAD, and 147 module genes were found to be CAD candidate genes identified by GWAS.

Box1

AIP, DRAP1, POLR2I and _PQBP1_ are THP-1 foam cell validated key drivers in RGN 42, which was identified as a cross-species, independently validated CAD-causal atherosclerosis module involved in RNA processing.

Sixty-one of 171 modules were associated with one of the four main CAD phenotypes: 14 atherosclerosis modules, 29 cholesterol modules, 14 glucose modules, and 15 CRP modules. Ten modules were associated with more than one phenotype. Atherosclerosis, cholesterol, and CRP modules were found to be mostly tissue-specific, whereas glucose modules were mostly CT modules.

Thirty of 61 phenotype-associated modules were also identified as CAD-causal: eight atherosclerosis modules (of 14), 10 cholesterol modules (of 29), five glucose modules (of 14), and four CRP modules (of 15). Three additional CAD-causal modules were associated with more than one of the four CAD phenotypes. Of the genes in the CAD-causal modules, 42% had previously been related to atherosclerosis or CAD. The CAD-causal modules also contained 59 unique CAD candidate genes identified by GWAS.

Twelve of 26 CAD-causal modules were validated through phenotypic associations in the HMDP. The CAD-causal atherosclerosis modules 42, 58, and 98 were also validated against the extent of atherosclerosis in mice (**Figure 14**). RGN 42 was re-identified in independent in-house data from atherosclerotic carotid artery lesions (module 42 genes had 9.5-fold higher connectivity compared to background), as well as in associated blood macrophages (3.7-fold higher connectivity). Moreover, RGN 42 was also validated in two external datasets: lipopolysaccharide-stimulated monocytes (2.2-fold higher connectivity) and macrophages (2.0-fold higher connectivity) isolated from patients with CAD (GEO: GSE9820).

Finally, four of seven key drivers in RGN 42 (*AIP*, *DRAP1*, *POLR2I*, and *PQBP1*) were validated in the THP-1 foam cell model in that siRNA targeting affected cholesterol-ester accumulation primarily by affecting RGN 42 genes.

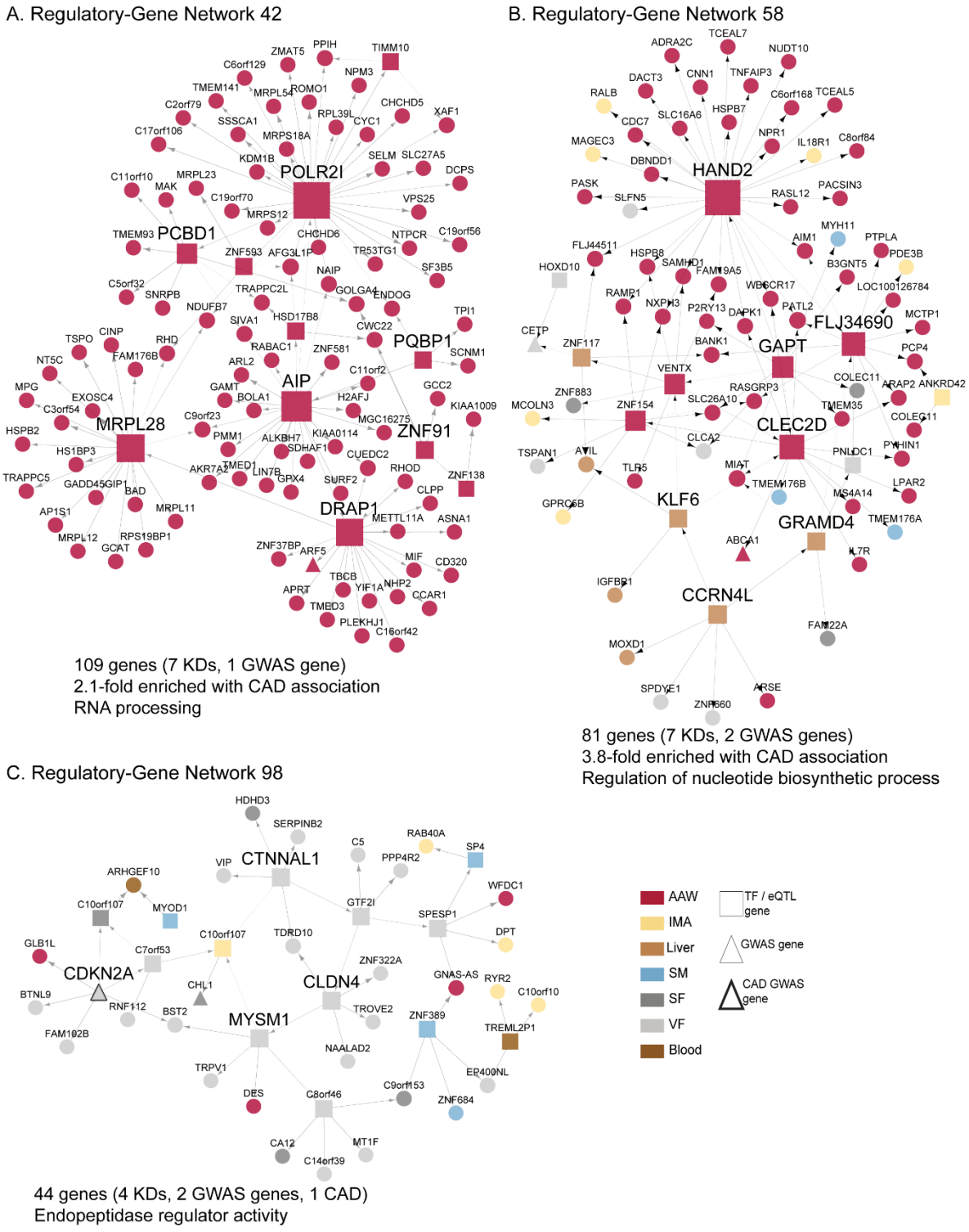

Figure 14: Cross-species validated atherosclerosis RGNs. *Bayesian RGNs with key drivers inferred from mice validated CAD-causal modules (i.e., module eQTLs are risk-enriched or contain CAD GWA candidate genes) linked to extent of coronary atherosclerosis (A–C). KD, key driver; GWA genes, total and CAD candidate genes identified in GWAS of CAD, plasma lipid/glucose levels, and type 2 diabetes. Fold enrichment for CAD association was assessed for network eQTLs/SNPs using the CARDIoGRAM dataset. The molecular process with the strongest functional enrichment assessed by Gene Ontology is indicated.*

We also tested the network connectivity of CAD-phenotype-associated modules by comparing gene connectivity between genes in modules identified in AAW (diseased artery) and IMA (healthy artery). Overall, we found that the connectivity for CAD-causal AAW/IMA modules increased in the diseased compared to healthy state. In contrast, the connectivity of CAD-reactive modules (i.e., modules that were not CAD-causal) either did not change or was lower in AAW than in IMA. We concluded that in CAD-driving modules, which frequently represented disease activity such as inflammation, gene-gene interactions increased in the disease state. The non-causal modules, however, reflected more normal vascular functions, and the connectivity of those modules was reduced in the disease state, suggesting that normal vascular physiology is impaired in the disease state.

Outside the main study pipeline, we also examined the connectivity between the 30 CAD-causal RGNs in a super-network based on RGN eigengene correlations. All 30 RGNs were connected to at least one of these 30 RGNs (**Figure 15**). CT modules 98 (dominated by VF genes, including its four key drivers) and 113 (dominated by SF genes, including its key drivers) appeared to act as hub modules and thus as key driver RGNs in this super-network. These two RGNs mediated all but one connection (from RGN 27 in SF) from the metabolic tissues to coronary atherosclerosis (i.e., AAW) RGNs.

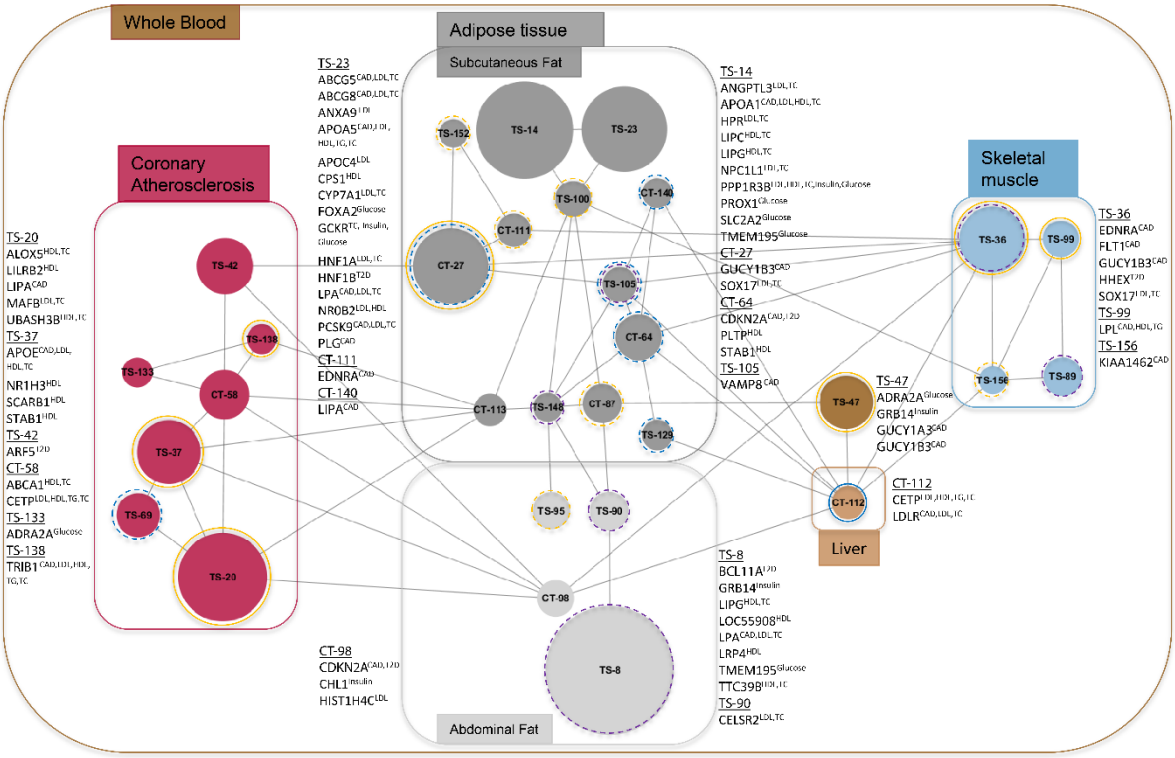

Figure 15: Super network of CAD-causal modules. *Eigengene associations (r > 0.4,* P-value *< 0.05) were used to link the 30 CAD-causal modules. RGNs/modules (circles) are oriented according to dominating tissue-belonging of genes and are color-coded accordingly. Numbers in circles are the module IDs. Circle size corresponds to the number of RGN/module genes. Colored circle circumferences indicate phenotype associations; dotted lines, RGNs/modules that are reactive to indicated phenotype; solid line, RGNs/modules that are causal for indicated*

phenotype. Orange, plasma cholesterol measures; blue, plasma glucose metabolism measures; and purple, plasma CRP levels. Next to each box, the names of GWA genes in the RGNs/modules and related trait (in superscript) are indicated.

We noted that CT RGNs/modules appeared to have more connections in this super-network of 30 CAD-causal modules than in TS RGNs/modules. This notion was reinforced by computational assessment of the super-network's connectivity. From this we found greater and stronger connectivity for CT modules than for TS modules (mean number of connections: 5.33 ± 1.73 vs. 3.43 ± 2.04, P-value < 0.02; mean connectivity strength: 2.96 ± 0.98 vs. 1.98 ± 1.20, P-value < 0.03). Notably, module 98 contains genes involved in endopeptidase activity previously implicated in atherosclerosis [106], and one of its four key drivers, *CDKN2A*, is a CAD candidate gene for the well-established Chr9p21 CAD risk locus [61].

4.2 PAPER II

In this study, we invented and validated a CT-weighted gene co-expression network analysis (X-WGCNA) method as an extension of weighted gene co-expression network analysis (WGCNA) [43]. We also implemented X-WGCNA in R.

X-WGCNA considers gene expression matrix from *n* number of tissues as an input to produce both TS and CT modules. We tested the X-WGCNA with seven tissues and 100,000 genes. **Figure 16** shows a comparison of WGCNA and X-WGCNA modules constructed from the STAGE and GTEx [67] dataset. Regardless of the threshold cut-off, this comparison indicated that X-WGCNA captures both CT and TS modules using both STAGE and GTEx datasets, whereas WGCNA mostly capture TS modules.

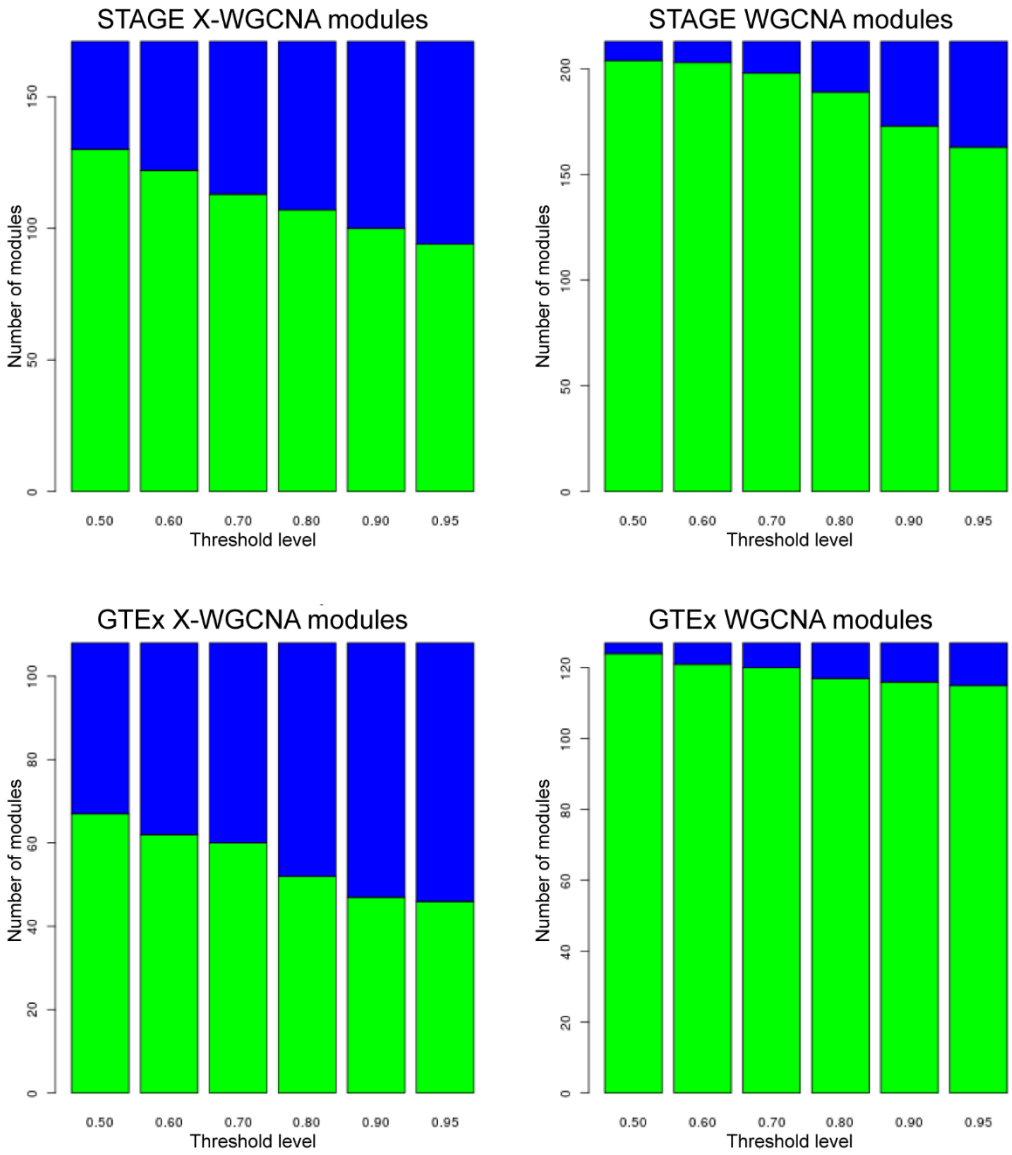

Figure 16: Comparison between X-WGCNA and WGCNA. *X-axis for threshold level to detect module type and y-axis for number of modules. Blue, CT modules; green, TS modules.*

With the X-WGCNA, the GO enrichment of CT modules was higher (68%, 999 categories vs. 58%, 881 categories), especially if we consider only categories that are exclusively found in CT modules (296 categories or 20% of total vs. 107 categories or 7% of total). These findings reinforce the idea that CT-modules are identifiable and real.

To further compare cross-tissue modules inferred with X-WGCNA and WGCNA, we took advantage of having the genotype data available for the STAGE study participants. Another proof of the biological relevance of the CT-modules was that the X-WGCNA CT modules contained 2.5-fold more genes with eQTLs than the TS modules (344 vs. 134), which was not the case for WGCNA-inferred modules (230 vs. 286). We also found that X-WGCNA CT modules were more often enriched with inherited risk for CAD than TS modules (15 modules vs. 1 module), which was not the case when comparing WGCNA CT versus TS modules (6 vs. 9).

Finally, X-WGCNA was found to be flexible in terms of choosing parameters like scale-free fitness, total tissue, most variant genes, minimum module size, and module type.

4.3 PAPER III

In this study, we examined gene expression in the atherosclerotic arterial wall in mice during atherosclerosis regression as a result of plasma cholesterol lowering (PCL) at three different stages of atherosclerosis progression: early (30 weeks), mature (40 weeks), and advanced (50 weeks) (**Figure 17**).

First, the extent of atherosclerosis progression and regression was studied in atherosclerosis-prone mice. We found that atherosclerosis regression in response to PCL occurred at all three stages (early, mature, and advanced), but only at the early stage did PCL lead to complete regression after 20 weeks. PCL at the mature and advanced stages led to substantial, but not complete, atherosclerosis regression (**Figure 17A**).

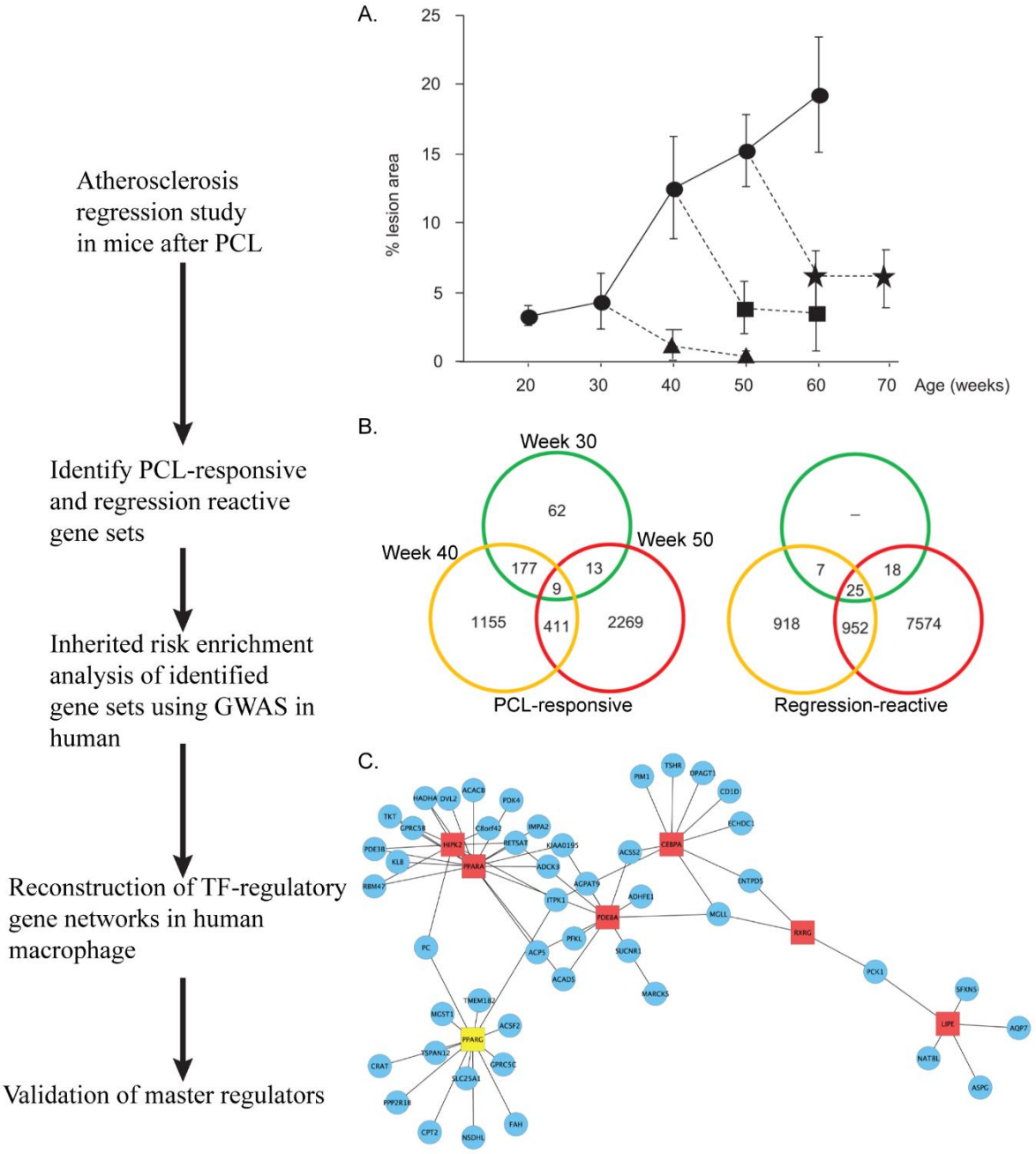

Figure 17: Major steps and results of cholesterol responsive gene network study. ***(A)*** *Atherosclerosis progression and regression curves (Straight line and dotted line for progression and regression curve respectively). Values are surface lesion area. Lesion development in controls without PCL (●) and in mice after PCL started at week 30 (▲), 40 (■), or 50 (✸).* ***(B)*** *Venn diagram of identified PCL-responsive and regression-reactive gene sets for the three stages. Green circle, early stage; yellow circle, mature stage; red circle, advanced stage.* ***(C)*** *TF-RGN from early PCL-responsive gene sets. Red rectangles, master regulators; yellow rectangle, validated master regulator.* Modified from [81]

Next, we identified PCL-responsive and regression-reactive genes for the three different stages (**Figure 17B**). The three PCL-responsive gene sets were less shared among the stages than the regression reactive gene sets. Analysis of inherited CAD-risk enrichment showed that only the PCL-responsive gene sets were risk-enriched (early, 2.0-fold, P-value = 3.1×10^{-14}; mature, 1.4-fold, P-value = 6.8×10^{-4}; advanced, 1.5-fold, P-value = 1.3×10^{-6}), indicating that the PCL-responsive genes were causal for the regression of atherosclerosis.

Given the causal role of the PCL-responsive gene sets, we reconstructed PCL-responsive TF-RGNs and within these identified master regulators:

Out of 215 human orthologs matching the early mouse PCL-responsive gene set, 53 were found in a TF RGN with *PPARA* and *PPARG* as top master regulators (**Figure 17C**).

Similarly, 185 PCL-responsive genes identified in the mature lesions formed a TF-RGN with the master regulators *HMGB2*, *ADORA2A*, *TERF1*, and *MLL5*.

In the advanced PCL-responsive TF-RGN, the top master regulators were *SRSF10*, *XRN2*, and *HMGB1*, which regulated a total of 379 network genes.

Using the THP-1 foam cell model, we validated four master regulators as affecting cholesterol-esters accumulation *in vitro*: *PPARG* in early (**Figure 17C**), *MLL5* in the mature, and *SRSF10*, and *XRN2* in advanced TF-RGNs.

4.4 PAPER IV

In this study, STAGE eQTLs were inferred from seven CAD-relevant tissues using matrix based eQTL discovery method kruX [107] and examined for inherited CAD-risk enrichment. In **Figure 18** (left side), the principal steps of this study are shown. Principal component analysis of gene expression data showed that gene expression clearly clustered according to individual tissues (**Figure 18A**).

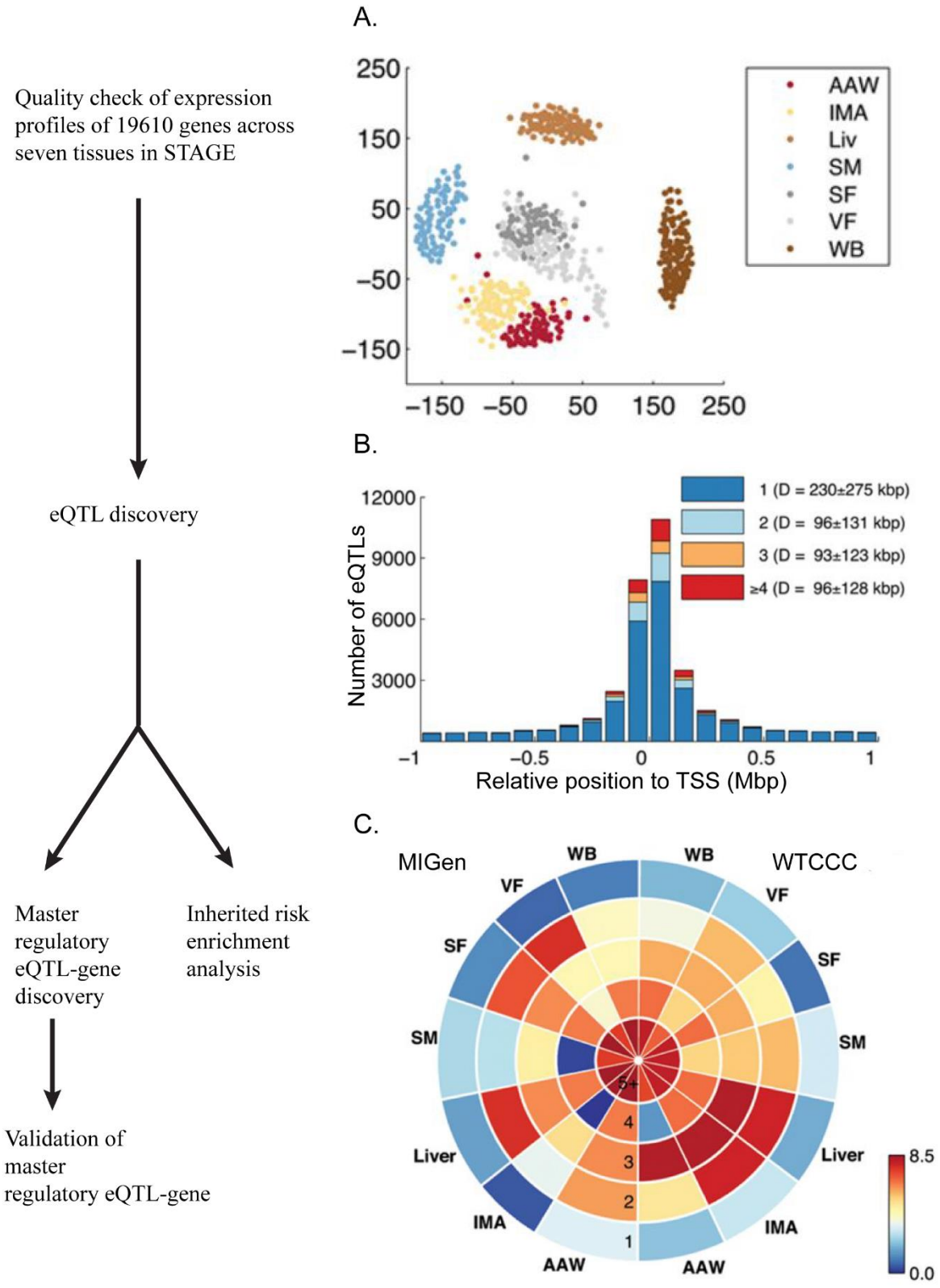

Figure 18: STAGE eQTL and their inherited CAD risk enrichment. ***(A)*** *Sammon plots of global gene expression. Each dot represents each samples expression profiles in each tissue.* ***(B)*** *Distribution of 29,530 cis-eQTLs in terms of their transcription start site (TSS). Dark blue, TS eQTLs; light blue, eQTLs shared by 2 tissues; orange, eQTLs shared by 3 tissues; red, eQTLs shared by >3 tissues.* ***(C)*** *STAGE eQTLs inherited risk enrichment analysis for TS eQTLs (first layer, outer circle) and multi-tissue eQTLs (nth layer from outer circle means shared by n tissues). Color scale satnds for degree of enrichment.* Modified from [71]

In total, we discovered 29,530 *cis*-eQTLs (associated with 6450 unique genes), of which 7429 (about 25% of total) were multi-tissue eQTLs, and 1494 *trans*-eQTLs, of which only 2.9% were multi-tissue eQTLs. **Figure 18B** shows distribution of *cis*-eQTLs in relation to the transcription start site (TSS).

Analysis of inherited CAD-risk analyis revealed that multi-tissue *cis*-eQTLs were more enriched with CAD risk than TS *cis*-eQTLs according to both MIGen and WTCCC GWAS. Among the multi-tissue *cis*-eQTLs, those increasingly shared were increasingly risk enriched (**Figure 18C**).

Then, 42 multi-tissue (shared by 5 or 6 tissues) *cis*-eQTLs were found to be more risk enriched (7.3 -fold, P-value < 7.7×10^{-20} in WTCCC; 2.3-fold, P-value < 3.1×10^{-6} in MIGen and 4.2-fold, P-value < 1.1×10^{-54} in CARDIoGRAM). To assess downstream effects on gene expression governed by these eQTLs, we identified gene sets of correlated genes, which we further analyzed using GO. Twenty-nine highly co-expressed (absolute correlation >0.85) gene sets (each containing 30 genes from different tissues) were associated with 16 of the above 42 multi-tissue *cis*-eQTLs. This 29 multi-tissue gene sets contains 19 unique genes called master regulatory eQTL genes. *G3BP1*, *FLYWCH1*, *PSORS1C3*, and *SNAPIN* were ranked as top four master regulatory eQTL genes based on their association with the functional and biological process and with the atherosclerosis score in STAGE. Finally three (*G3BP1*, *FLYWCH1*, and *PSORS1C3*) of them were validated in the THP-1 foam cell model.

5 DISCUSSION

In this thesis, we used a multifaceted systems genetics approach in which we integrated analyses of gene expression, DNA genotypes, clinical phenotypes, and GWAS datasets. We established a computational pipelines showing how systems genetics can be used to discover and validate eQTLs, regulatory gene networks, and key disease drivers active within and across tissues that are important for a common complex disease, CAD.

Our findings in Paper I provides a preliminary view of the regulatory landscape of causal molecular processes active within and across a majority of tissues believed to be central to advanced CAD. We identified 94 TS modules and 77 CT modules using X-WGCNA, an extension of WGCNA (explained in Paper II). Computationally it was not feasible to consider all genes from the seven STAGE tissues and therefore we only considered the most variant genes from each tissue. Nonetheless, we could still identify TS and CT RGNs that included both established [108] and previously unreported CAD candidate genes in the form of key drivers. These candidate genes participate in diverse molecular processes and established pathways of atherosclerosis, cholesterol and glucose metabolism, and acute inflammation, and were regulated in both TS and CT networks. Importantly, we found that nearly half of the RGNs were evolutionarily conserved, as judged from validation against the HMDP [53]. As proof of concept, in RGN 42, a cross-species-validated, mouse atherosclerosis- and CAD-causal network active in AAW and involving RNA-processing genes, four key drivers (*AIP*, *DRAP1*, *POLR2I*, and *PQBP1*) specifically activated the same network genes and affected THP-1 foam cell formation. The entire RGN 42 was also re-identified in independent gene expression data from both CAD macrophages and carotid lesions.

We also reconstructed a super-network containing all 30 CAD-causal RGNs across all the main CAD tissues. This super-network may prove to be important for understanding CAD because it links a good portion of disease-driving molecular processes known for CAD, including their relation to key metabolic risk factors; it also provides a preliminary overview of the gene regulatory landscape in CAD. Mapping the regulatory framework of complex diseases in this fashion provides a starting point to assess the overall molecular status of individual patients [9]. In fact, more detailed versions of regulatory maps like the one presented here (**Figure 15**, page 28) will likely be required to achieve the goals of precision medicine.

Besides integrated whole-systems genetics study, it is equally essential to study specific subsystems or development phases of disease. In Paper III, we conducted such studies, and we discovered PCL-responsive genes causal for atherosclerosis regression for three stages of atherosclerosis progression. Here we used only TF genes as prior and the CLR-Pearson method to reconstruct RGNs instead of the more commonly used Bayesian method. We identified three TF RGNs with key drivers that were significantly associated with CAD-related functions like immune response. In the end, we validated key drivers by showing their effects on foam cell formation, a key and continuous disease-driving process at all stages of atherosclerosis development.

Analysis of inherited risk enrichment of module genes using GWAS is a novel way to define causality. We have implemented this approach in Papers I and III. Paper IV originally describes this method and the eQTL discovery in STAGE. STAGE eQTL risk enrichment revealed that multi-tissue (shared by more than one tissue) eQTLs are more enriched in CAD risk than TS eQTLs. This finding is consistent with our module risk enrichment result in Paper I—that CT modules are more risk enriched than TS modules—and is a feasible finding considering the multifactorial and CT nature of CAD.

We made another noticeable observation of this thesis: In Paper I, we discovered RGNs by analyzing a human dataset and validated them in a mouse dataset, whereas in Paper III we discovered PCL-responsive gene sets in mice and validated those genes in human datasets. Thus, it appears both approaches are valid although it seems to us that if possible, it is always preferable to make the initial finding in humans and thereafter to validate the findings in animal models such as mice.

In this thesis, we also extensively discussed and showed how integrated systems genetics analysis can and perhaps must be used to embrace the complexity of common diseases like CAD. A limitation of this thesis is that we only analyzed transcriptomics data along with genotype and clinical phenotype. Proteomics, metabolomics, and other omics data will no doubt complement systems genetics approaches to CAD.

In sum, in this thesis we show that a systems genetics approach on the STAGE GGE study has helped to provide a better understanding of the molecular landscape in CAD and regression of atherosclerosis. It is our hope that the RGNs revealed in this thesis will be proven useful for finding novel therapies and early diagnostics and thereby help to reduce the heavy burden CAD puts on most societies.

6 CONCLUDING REMARKS AND FUTURE WORKS

In this thesis, we showed an application of the NEW biology by using an integrated systems genetics approach to retrieve disease associated gene networks and key drivers. We also developed a cross-tissue weighted gene co-expression network analysis method, called X-WGCNA, and proved that it can reliably capture both TS and CT modules across tissues. Specifically for CAD, my thesis provides a first repository of RGNs and KDs in CAD. In the near future I plan to,

- Apply X-WGCNA to other complex diseases for which multiple tissues have or will be sampled. By updating X-WGCNA from sequential execution to parallel execution, we can minimize its run time and maximize its capacity.
- Apply the X-WGCNA to RNAseq data of the STARNET study
- Map the RGNs to corresponding protein-protein interaction networks.
- Further validate the RGNs and KDs disclosed in this thesis by working together with CAD scientists with in-depth knowledge of established CAD pathways.

7 ACKNOWLEDGEMENTS

I would like to complete my thesis by expressing my gratitude even though it is impossible to express it in text.

As always, at first I am grateful to almighty **ALLAH**, **the god**, for your mercy and blessings.

Thanks to **Karolinska Institutet** to support my entire PhD time.

Johan Björkegren, my main supervisor. I am not sure how I should express my gratitude to you. Thank you very much for give me the opportunity and freedom to work with your group, "cardiovascular genomics" as a PhD student. I feel proud and lucky to get an excellent supervisor like you. You are a great scientist and leader. Besides science I also learned leaderships and caring from you.

Josefin Skogsberg, my co-supervisor, thank you very much for your brilliant teaching capability. I am also thankful for your support and suggestions on many other practical issues.

Tom Michoel, my co-supervisor, you are a great computational biologist, thank you very much for your time and effort to teach how I should think about the logic behind principal. I learned a lot from you. I had most productive time when I visited you in Edinburgh.

Christer Betsholtz, my mentor, unfortunately we didn't have any meeting on my progress or career plan but enjoyed to work with you in "Kidney" project. Thanks for your time and valuable suggestions.

Cardiovascular Genomics group,

KI-lab: **Hassan Foroughi Asl** and **Aranzazu Rossignoli**, it was very nice to have you as group member, as a friend. Thank you very much for enlightening discussions on science and many other practical issues. **Mingmei Shang (**former group member**),** thanks for your friendly discussions and suggestions.

Tartu-lab: **Rajeev Jain, Tiia Tooming (**former lab members**),** and **Katyayani Sukhavasi**, thank you very much for your contribution on my thesis and sharing your knowledge. **Arno Ruusalepp** and **Raili Ermel,** thank you very much for your contribution to build CAD dataset.

MSSM-lab: **Eric E. Schadt**, thank you very much for your suggestions on computational pipeline at the beginning of my PhD studies and it was great to closely work with you in a bioinformatics workshop at MSSM. **Ariella Cohain**, **Yifan Mo**, and **Oscar Franzen**, thanks for your valuable suggestions on solving different problems and lively discussions on STARNET project. **Chiara Giannarelli**, thanks for giving me the opportunity to work with your project. **Saboor Hekmaty,** thanks for your companion during MSSM and Edinburgh visit.

Vascular Biology, I am thankful to all colleagues at vascular biology. **Ulf Eriksson**, thanks for your leadership at VB. **Sebastian Lewandowski**, thanks for discussions on ALS data

analysis. **Miyuki Katayama** (former VB member), thanks for sharing your knowledge and data on kidney diseases.

I am thankful to all administrative persons. **Chad Tunell, Gizella Bengtsson,** and **Alessandra Nanni** thank you very much for your excellent support and suggestions on different IT and administrative issues.

Leducq consortium: I am thankful to all Leducq members, special thanks to **Mete Civelek** and **Aldons J. Lusis** for sharing your data and knowledge.

I am also thankful to all co-authors of my publications, published during PhD studies.

Outside lab:

I am grateful to my brother-in-law, **Rafiqul Hyder**, thank you very much for your support and motivation to move in the field of "computational biology and bioinformatics".

I am also thankful to **Jeanette Hellgren Kotaleski**, my former supervisor. Thanks for your support, directions, and inspiration.

Riad vai, I am grateful to you for sharing information regarding research in computational neuroscience.

I am thankful to all of my Bangladeshi friends and their families in Sweden. **Saiful vai, Shabbir vai, Mushfiq vai, Emon vai** thanks for your companion at KI. **Noman vai-Adina apu, Tawhid vai, Raju vai, Ribon vai, Shohagh vai, Masum vai, Babu vai, Shahid vai, Rekha vabi** and your families, thank you very much for staying in touch as a good family friend. **Bahktiar vai** and **Sheuli apa** thanks for being like a local guardian.

I am thankful to all of my friends. **Harun, Rifat, Foni, Mamun, Holy, Tajul, Topon, Shaibal**, **Masud Mama**, thanks for staying in touch as a good friend.

I am thankful to all of my relatives for their love and caring, special thanks to **Sumon vaia** – is a great tutor.

Family members:

I am grateful to my in-laws (**Boro vaia - Bulbuli vabi, Boro apu - Opu vaia, Ritu vabi, Shimul apu - Jahangir vai, Shishir apu - Sheikh vaia, Sohel vai and Shuvo**) for their love, support and inspirations. Special thanks to **Ma** (my **mother-in-law**), **Shumi Apu,** and **Tisa** for their regular caring about progress of my study. Missing you a lot **Baba** (my **father-in-law**), I still can feel your inspiration for higher study, thanks ….

I am thankful to my brother-**Habib Ahmed**, and his wife **Salma Akhter (Poly)**, my sister-**Kamrun Nahar (Happy)**, and her husband **Aminul Islam (Helim)** for their continuous support and love.

My parents, **Abba-Amma,** I only can say I am exist only because of your love, sacrifice and Doa. **Tahmina Akhter (Jisa)**, my lovely wife, you are a true game changer, I can do anything with you but nothing without you. **Zayef Nur Ahmed,** my son, amazing gift from almighty and greatest achievement of my life.

I would like to finish by giving my love and Doa to all of our sweet and cute babies.

8 REFERENCES

1. Mendis S, Puska P, and e. Norrving B, *Global Atlas on Cardiovascular Disease Prevention and Control.* World Health Organization, Geneva, 2011.

2. Mensah, G.A., et al., *The global burden of cardiovascular diseases, 1990-2010.* Glob heart, 2014. **9**(1): p. 183-4.

3. GBD 2013 Mortality and Causes of Death Collaborators, *Global, regional, and national age-sex specific all-cause and cause-specific mortality for 240 causes of death, 1990-2013: a systematic analysis for the Global Burden of Disease Study 2013.* The Lancet, 2015. **385**(9963): p. 117-171.

4. Murray, C.J.L., et al., *Disability-adjusted life years (DALYs) for 291 diseases and injuries in 21 regions, 1990-2010: a systematic analysis for the Global Burden of Disease Study 2010.* The Lancet, 2012. **380**(9859): p. 2197-2223.

5. Lozano, R., et al., *Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: a systematic analysis for the Global Burden of Disease Study 2010.* The Lancet, 2012. **380**(9859): p. 2095-2128.

6. Samani, N.J. and D.P. de Bono, *Prevention of coronary heart disease with pravastatin.* N. Engl. J. Med., 1996. **334**(20): p. 1333-4.

7. Bangalore, S., et al., *beta-Blocker use and clinical outcomes in stable outpatients with and without coronary artery disease.* JAMA., 2012. **308**(13): p. 1340-9.

8. Barabasi, A.L., N. Gulbahce, and J. Loscalzo, *Network medicine: a network-based approach to human disease.* Nat. Rev. Genet., 2011. **12**(1): p. 56-68.

9. Bjorkegren, J.L., et al., *Genome-wide significant loci: how important are they?: Systems genetics to understand heritability of coronary artery disease and other common complex disorders.* J. Am. Coll. Cardiol., 2015. **65**(8): p. 830-845.

10. Civelek, M. and A.J. Lusis, *Systems genetics approaches to understand complex traits.* Nat. Rev. Genet., 2014. **15**(1): p. 34-48.

11. Schadt, E.E., *Molecular networks as sensors and drivers of common human diseases.* Nature, 2009. **461**(7261): p. 218-23.

12. Schadt, E.E. and J.L. Bjorkegren, *NEW: network-enabled wisdom in biology, medicine, and health care.* Sci. Transl. Med., 2012. **4**(115): p. 115rv1.

13. MacLellan, W.R., Y. Wang, and A.J. Lusis, *Systems-based approaches to cardiovascular disease.* Nat Rev Cardiol, 2012. **9**(3): p. 172-184.

14. Deloukas, P., et al., *Large-scale association analysis identifies new risk loci for coronary artery disease.* Nat. Genet., 2013. **45**(1): p. 25-33.

15. Peden, J.F. and M. Farrall, *Thirty-five common variants for coronary artery disease: the fruits of much collaborative labour.* Hum Mol Genet, 2011. **20**(R2): p. R198-205.

16. Barabasi, A.L. and Z.N. Oltvai, *Network biology: understanding the cell's functional organization.* Nat. Rev. Genet., 2004. **5**(2): p. 101-13.

17. Lusis, A.J., *Atherosclerosis.* Nature, 2000. **407**(6801): p. 233-41.

18. Lusis, A.J., *Genetics of atherosclerosis.* Trends in Genetics, 2012. **28**(6): p. 267-275.

19. Hansson, G.K., *Inflammation, atherosclerosis, and coronary artery disease.* N Engl J Med, 2005. **352**(16): p. 1685-95.

20. Tegnér, J., J. Skogsberg, and J. Björkegren, *Thematic review series: Systems Biology Approaches to Metabolic and Cardiovascular Disorders. Multi-organ whole-genome measurements and reverse engineering to uncover gene networks underlying complex traits.* Journal of Lipid Research, 2007. **48**(2): p. 267-277.

21. Janine, P., et al., *Cardiovascular Disease and Dyslipidemia: Beyond LDL.* Current Pharmaceutical Design, 2011. **17**(9): p. 861-870.

22. Sowers, J.R., M. Epstein, and E.D. Frohlich, *Diabetes, Hypertension, and Cardiovascular Disease: An Update.* Hypertension, 2001. **37**(4): p. 1053-1059.

23. Mayer, B., J. Erdmann, and H. Schunkert, *Genetics and heritability of coronary artery disease and myocardial infarction.* Clinical Research in Cardiology, 2007. **96**(1): p. 1-7.

24. Reaven, G.M., *Insulin Resistance: the Link Between Obesity and Cardiovascular Disease.* Medical Clinics of North America, 2011. **95**(5): p. 875-892.

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

26. Wikipedia, *Tunica intima in Wikipedia.* Online accessed 15th June 2016, https://en.wikipedia.org/wiki/Tunica_intima.

27. Leiva, E., et al., *Role of Oxidized LDL in Atherosclerosis.* 2015.

28. Koenig, W. and N. Khuseyinova, *Biomarkers of Atherosclerotic Plaque Instability and Rupture.* Arteriosclerosis, Thrombosis, and Vascular Biology, 2007. **27**: p. 15-26.

29. Lusis, A.J. and J.N. Weiss, *Cardiovascular Networks: Systems-Based Approaches to Cardiovascular Disease.* Circulation, 2010. **121**(1): p. 157-170.

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

31. Schadt, E.E., *Molecular networks as sensors and drivers of common human diseases.* Nature, 2009. **461**(7261): p. 218-223.

32. MB, M. and T. YW, *Basic concepts of microarrays and potential applications in clinical.* Clin Microbiol Rev. 2009 Oct;22(4):611-33. doi: 10.1128/CMR.00019-09., 2009(1098-6618 (Electronic)): p. 611-33.

33. 5 5Russo, G., C. Zegar, and A. Giordano, *Advantages and limitations of microarray technology in human cancer.* Oncogene, 2003. **22**(42): p. 6497-6507.

34. Metzker, M.L., *Sequencing technologies [mdash] the next generation.* Nat Rev Genet, 2010. **11**(1): p. 31-46.

35. Schuster, S.C., *Next-generation sequencing transforms today's biology.* Nat Meth, 2008. **5**(1): p. 16-18.

36. Ohashi, H., et al., *Next-Generation Technologies for Multiomics Approaches Including Interactome Sequencing*. Vol. 2015. BioMed Research International. 9.

37. Illumina, *An Introduction to Next-Generation Sequencing Technology.* Online accessed 24th June 2016, http://www.illumina.com/content/dam/illumina-marketing/documents/products/illumina_sequencing_introduction.pdf.

38. Wang, Z., M. Gerstein, and M. Snyder, *RNA-Seq: a revolutionary tool for transcriptomics.* Nature reviews. Genetics, 2009. **10**(1): p. 57-63.

39. Neto, E.C., et al., *Causal graphical models in systems genetics: a unified framework for joint inference of causal network and genetic architecture for correlated phenotypes.* The annals of applied statistics, 2010. **4**(1): p. 320-339.

40. Vidal, M., M.E. Cusick, and A.-L. Barabási, *Interactome Networks and Human Disease.* Cell, 2011. **144**(6): p. 986-998.

41. Ravasz, E., et al., *Hierarchical organization of modularity in metabolic networks.* Science, 2002. **297**(5586): p. 1551-5.

42. Jeong, H., et al., *The large-scale organization of metabolic networks.* Nature, 2000(0028-0836 (Print)).

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

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

45. Stuart, J.M., et al., *A Gene-Coexpression Network for Global Discovery of Conserved Genetic Modules.* Science, 2003. **302**(5643): p. 249-255.

46. Madar, A., et al., *DREAM3: Network Inference Using Dynamic Context Likelihood of Relatedness and the Inferelator.* PLoS ONE, 2010. **5**(3): p. e9803.

47. Friedman, N., *Inferring cellular networks using probabilistic graphical models.* Science, 2004. **303**(5659): p. 799-805.

48. Nir Friedman, Moises Goldszmidt, and A. Wyner, *Data Analysis with Bayesian Networks: A Bootstrap Approach.* Cornell University Library, 2013. **arXiv:1301.6695v1**.

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

50. Heckerman, D., *A Tutorial on Learning With Bayesian Networks.* Microsoft Research, 1998.

51. Bonnet, E., L. Calzone, and T. Michoel, *Integrative Multi-omics Module Network Inference with Lemon-Tree.* PLoS Comput Biol, 2015. **11**(2): p. e1003983.

52. Brem, R.B., et al., *Genetic Dissection of Transcriptional Regulation in Budding Yeast.* Science, 2002. **296**(5568): p. 752-755.

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

54. Lappalainen, T., et al., *Transcriptome and genome sequencing uncovers functional variation in humans.* Nature, 2013. **501**(7468): p. 506-511.

55. van Nas, A., et al., *Expression Quantitative Trait Loci: Replication, Tissue- and Sex-Specificity in Mice.* Genetics, 2010. **185**(3): p. 1059-1068.

56. Orozco, L.D., et al., *Unraveling inflammatory responses using systems genetics and gene-environment interactions in macrophages.* Cell, 2012. **151**(3): p. 658-70.

57. Romanoski, C.E., et al., *Systems Genetics Analysis of Gene-by-Environment Interactions in Human Cells.* The American Journal of Human Genetics, 2010. **86**(3): p. 399-410.

58. Breitling, R., et al., *Genetical Genomics: Spotlight on QTL Hotspots.* PLoS Genet, 2008. **4**(10): p. e1000232.

59. Lewis Cm Fau - Knight, J. and J. Knight, *Introduction to genetic association studies.* Cold Spring Harbor Protocols, 2012. **3**(1559-6095 (Electronic)): p. 297-306.

60. GWAS-studies, *Accessed on 25th June 2016 from https://ghr.nlm.nih.gov/primer/genomicresearch/gwastudies.*

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

62. Hindorff LA, M.J.E.B.I., Wise A, Junkins HA, Hall PN, Klemm AK, and Manolio TA, *A Catalog of Published Genome-Wide Association Studies. Available at: www.genome.gov/gwastudies.* 2012.

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

64. Monks, S.A., et al., *Genetic inheritance of gene expression in human cell lines.* Am J Hum Genet, 2004. **75**(6): p. 1094-105.

65. Schadt, E.E., et al., *Genetics of gene expression surveyed in maize, mouse and man.* Nature, 2003. **422**(6929): p. 297-302.

66. Consortium, G.T., *Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans.* Science, 2015. **348**(6235): p. 648-60.

67. Ardlie, K.G., et al., *The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans.* Science, 2015. **348**(6235): p. 648-660.

68. Hägg, S., et al., *Multi-organ expression profiling uncovers a gene module in coronary artery disease involving transendothelial migration of leukocytes and LIM domain binding 2: the Stockholm Atherosclerosis Gene Expression (STAGE) study.* PLoS genetics, 2009. **5**: p. e1000754.

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

70. Oscar Franzén, et al., *Cardiometabolic Risk Loci Share Downstream Cis- and Trans-Gene Regulation Across Tissues and Diseases. The Stockholm-Tartu Atherosclerosis Reverse Network Engineering Task (STARNET) Study.* Accepted in Science, 2016.

71. Foroughi Asl, H., et al., *Expression quantitative trait Loci acting across multiple tissues are enriched in inherited risk for coronary artery disease.* Circ Cardiovasc Genet, 2015. **8**(2): p. 305-15.

72. Austen, W.G., et al., *A reporting system on patients evaluated for coronary artery disease. Report of the Ad Hoc Committee for Grading of Coronary Artery Disease, Council on Cardiovascular Surgery, American Heart Association.* Circulation, 1975. **51**(4 Suppl): p. 5-40.

73. Nicholls, S.J., et al., *Effect of diabetes on progression of coronary atherosclerosis and arterial remodeling: a pooled analysis of 5 intravascular ultrasound trials.* 2008(1558-3597 (Electronic)).

74. Kovacic, J.C., et al., *The Relationships Between Cardiovascular Disease and Diabetes: Focus on Pathogenesis.* Endocrinology and Metabolism Clinics of North America, 2014. **43**(1): p. 41-57.

75. Ridker, P.M., et al., *C-reactive protein and other markers of inflammation in the prediction of cardiovascular disease in women.* The New England Journal of Medicine, 2000(0028-4793 (Print)).

76. Ridker, P.M., et al., *Comparison of C-reactive protein and low-density lipoprotein cholesterol levels in the prediction of first cardiovascular events.* The New England Journal of Medicine, 2002(1533-4406 (Electronic)).

77. Shang, M.M., et al., *Lim domain binding 2: a key driver of transendothelial migration of leukocytes and atherosclerosis.* Arterioscler Thromb Vasc Biol, 2014. **34**(9): p. 2068-77.

78. Stengel, D., et al., *Inhibition of LPL Expression in Human Monocyte–Derived Macrophages Is Dependent on LDL Oxidation State: A Key Role for Lysophosphatidylcholine.* Arteriosclerosis, Thrombosis, and Vascular Biology, 1998. **18**(7): p. 1172-1180.

79. Lusis, A.J., et al., *The Hybrid Mouse Diversity Panel: a resource for systems genetics analyses of metabolic and cardiovascular traits.* Journal of Lipid Research, 2016. **57(6)**(1539-7262 (Electronic)): p. 925-942.

80. Lieu, H.D., et al., *Eliminating atherogenesis in mice by switching off hepatic lipoprotein secretion.* Circulation, 2003(1524-4539 (Electronic)).

81. Bjorkegren, J.L., et al., *Plasma cholesterol-induced lesion networks activated before regression of early, mature, and advanced atherosclerosis.* PLoS Genet., 2014. **10**(2): p. e1004201.

82. *Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls.* Nature, 2007. **447**(7145): p. 661-78.

83. Consortium, M.I.G., *Genome-wide association of early-onset myocardial infarction with single nucleotide polymorphisms and copy number variants.* Nat Genet, 2009. **41**(3): p. 334-341.

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

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

86. Soranzo, N., et al., *Common variants at 10 genomic loci influence hemoglobin A(1)(C) levels via glycemic and nonglycemic pathways.* Diabetes, 2010. **59**(12): p. 3229-39.

87. Strawbridge, R.J., et al., *Genome-wide association identifies nine common variants associated with fasting proinsulin levels and provides new insights into the pathophysiology of type 2 diabetes.* Diabetes, 2011. **60**(10): p. 2624-34.

88. Burdett T (EBI), H.P.N., Hasting E (EBI) Hindorff LA (NHGRI), Junkins HA (NHGRI), Klemm AK (NHGRI), MacArthur J (EBI), Manolio TA (NHGRI), Morales J (EBI), Parkinson H (EBI) and Welter D (EBI). *The NHGRI-EBI Catalog of published genome-wide association studies. Available at: www.genome.gov/gwastudies. Accessed: October 2014.*

89. Faith, J.J., et al., *Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles.* PLoS Biol, 2007. **5**(1): p. e8.

90. Margolin, A.A., et al., *ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context.* 2006(1471-2105 (Electronic)).

91. Langfelder, P., B. Zhang, and S. Horvath, *Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R.* Bioinformatics, 2008. **24**(5): p. 719-20.

92. Maere, S., K. Heymans, and M. Kuiper, *BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks.* Bioinformatics, 2005. **21**(16): p. 3448-9.

93. Benjamini, Y. and Y. Hochberg, *Controlling the false discovery rate - a practical and powerful approach to multiple testing.* Journal of the Royal Statistical Society Series B-Methodological, 1995. **57**(1): p. 289-300.

94. Ashburner, M., et al., *Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.* Nat Genet, 2000. **25**(1): p. 25-9.

95. The Gene Ontology, C., *Gene Ontology Consortium: going forward.* Nucleic Acids Research, 2015. **43**(D1): p. D1049-D1056.

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

97. Hwang, D., et al., *A data integration methodology for systems biology.* Proc. Natl. Acad. Sci. U S A, 2005. **102**(48): p. 17296-301.

98. Storey, J.D., *A direct approach to false discovery rates.* Journal of the Royal Statistical Society Series B-Methodological, 2002. **64.3** p. 479-498.

99. Zhong, H., et al., *Integrating pathway analysis and genetics of gene expression for genome-wide association studies.* Am J Hum Genet, 2010. **86**(4): p. 581-91.

100. Peng, G., et al., *Gene and pathway-based second-wave analysis of genome-wide association studies.* Eur J Hum Genet., 2010. **18(1):111-7.**(- 1476-5438 (Electronic)): p. - 111-7.

101. Koller, D. and J. Friedman, *Probabilistic Graphical Models: Principles and Techniques*. 2009: The MIT Press.

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

103. Schmidt, M., A. Niculescu-Mizil, and K. Murphy, *Learning graphical model structure using L1-regularization paths.* AAAI'07 Proceedings of the 22nd National Conference on Artificial Intelligence, 2007. **2**: p. 1278–1283.

104. Zhang, B. and J. Zhu, *Identification of key causal regulators in gene networks.* Proceedings of the World Congress on Engineering 2013, 2013. **II**: p. 1309-1312.

105. Skogsberg, J., et al., *Transcriptional profiling uncovers a network of cholesterol-responsive atherosclerosis target genes.* PLoS. Genet., 2008. **4**(3): p. e1000036.

106. Kugiyama, K., et al., *Suppression of atherosclerotic changes in cholesterol-fed rabbits treated with an oral inhibitor of neutral endopeptidase 24.11 (EC 3.4.24.11).* Arterioscler. Thromb. Vasc. Biol., 1996. **16**(8): p. 1080-7.

107. Qi, J., et al., *kruX: matrix-based non-parametric eQTL discovery.* BMC Bioinformatics, 2014. **15**: p. 11.

108. Brænne, I., et al., *Prediction of Causal Candidate Genes in Coronary Artery Disease Loci.* Arteriosclerosis, Thrombosis, and Vascular Biology, 2015. **35**(10): p. 2207-2217.