From the Department of Medicine in Solna,
Karolinska Institutet, Stockholm, Sweden

GENE NETWORKS AND MODULES IN ATHEROSCLEROSIS

Jesper Lundström

Karolinska Institutet

Stockholm 2008

Published by Karolinska Institutet.

ISBN 978-91-7409-191-5

"You can't stop the waves,
but you can learn to surf"

Jon Kabat-Zinn

Dedicated to Maria

Abstract

In this thesis we are using global gene expression profiles to unravel functional gene networks and modules. The focus is atherosclerosis, a disease with manifestations in the artery wall where deposits of lipids accumulate and trigger immune responses causing the development of plaques, which upon rupture can lead to a myocardial infarction or stroke. Atherosclerosis is a complex disease influenced by energy metabolism in multiple organs and by several genetic and environmental risk factors. To meet this complexity, we believe the most appropriate approach is to identify gene networks and modules in patients suffering coronary artery disease as well as a relevant mouse model with human-like dyslipidemia prone to atherosclerosis development. First, we investigate structural properties of the regulatory gene network in yeast, integrating protein–protein interactions with the transcription network resulting in an estimate the effective gene network underlying gene expression data. In this effective gene network, we show evidence of in-hubs and provide a method for predicting in-hubs directly from gene expression data.

In the second study, we used the $Ldlr^{-/-}$ $Apob^{100/100}$ $Mttp^{flox/flox}$ Mx1-*Cre* mouse model to study atherosclerosis development and how this development is effected by plasma cholesterol-lowering. This mouse model has a lipid profile similar to human hyperlipidemia and develops atherosclerosis on a chow diet. Moreover, it contains a genetic switch ($Mttp^{flox/flox}$ Mx1-*Cre*) to turn off the VLDL synthesis in the liver and lowering plasma cholesterol by $> 80\%$. Atherosclerotic lesions progressed slowly at first, then expanded rapidly, and plateaued after advanced lesions formed. Analysis of lesion expression profiles indicated lipid-poor macrophages accumulated prior the rapid expansion of the plaques. When macrophage concentration reached a critical point it was followed by a rapid expansion phase with accelerated foam-cell formation and inflammation, an interpretation also supported by lesion histology. A network of 8 cholesterol-responsive atherosclerosis genes was identified as central to the rapid expansion of the plaques.

Third, in the Stockholm Atherosclerosis Gene Expression (STAGE) study, including 124 well-characterized patients undergoing coronary artery bypass surgery, we measured and analyzed 278 expression profiles from the liver, skeletal muscle, mediastinal fat, and aortic lesion (atherosclerotic artery expression with unaffected arterial wall expression subtracted). Clustering of these gene expression profiles—performed separately in each organ—generated a total of 60 clusters. Two clusters, in aortic lesion (n = 49) and fat (n = 59), related to degree of atherosclerosis. Remarkably, in a validation cohort 27 genes were replicated in a cluster (n = 55) also related to the degree of atherosclerosis. In all three clusters relating to atherosclerosis (i.e., the atherosclerosis module), genes in the transendothelial migration of leukocyte pathway (TEML) were overrepresented and the transcription co-factor LIM-domain binding 2 (LDB2) expressed in lesion macrophages and endothelial cells was identified as a potential regulator of this module.

In the last study, we first identified 2457 cholesterol-responsive genes in the atherosclerotic arterial wall by lowering plasma cholesterol at 10-weeks intervals during atherosclerosis development using the mouse model of Study II. To prioritize the most atherosclerosis-relevant genes among these 2457, we used a list of 1259 genes active during atherogenesis (Study II) together with three global gene networks generated from human atherosclerosis gene expression profiles in study III, public literature mining, and protein-protein interaction data. Using an integrative network approach to identify genes neighboring any of 68 atherosclerosis seed genes, we identified 35 cholesterol-responsive genes that were believed to be highly relevant to atherosclerosis.

Taken together, this thesis provides evidence that systems biological analysis of global gene expression profiles isolated from a wide range of biological specimens can be used to infer functional interactions of genes in modules or networks. The content and architecture of these modules and networks can be used to improve our understanding how complex disorders like atherosclerosis develop.

List of Original Publications

I. Lundström J, Björkegren J, and Tegner J. Evidence of Highly Regulated Genes ("in-hubs") in the Gene Networks of Saccharomyces Cerevisiae *Bioinformatics and Biology Insights 2:313-322, 2008.*

II. Skogsberg J*, Lundström J*, Kovacs A*, Nilsson R, Noori P, Maleki S, Köhler M, Hamsten A, Tegner J and Björkegren J Transcriptional profiling uncovers a network of cholesterol-responsive atherosclerosis target genes. *PLoS Genetics, 4(3):e1000036, 2008.*

III. Hägg S*, Skogsberg J*, Lundström J*, Noori P, Nilsson R, Brinne B, Bradshaw M, Samnegård A, Silveira A, Lockowandt U, Liska J, Konrad P, Takolander R, Rosfors R, Franco-Cereceda A, Ivert T, Hamsten A, Tegner J and Björkegren J Multi-Organ Expression Profiling Uncovers a Module Involving Transendothelial Migration of Leukocytes and the Transcription Co-factor Lim Domain Binding 2 in Coronary Artery Disease: The Stockholm Atherosclerosis Gene Expression (STAGE) Study. *Submitted.*

IV. Lundström J, Nilsson R, Noori P, Skogsberg J, Björkegren J and Tegnér J A integrative systems medicine approach to uncover cholesterol-responsive networks in atherosclerosis. *Submitted.*

List of abbreviations

CABG	Coronary artery bypass grafting
CAD	Coronary artery disease
cDNA	Complementary DNA
cRNA	Complementary RNA
CNV	Copy number variation
CRAG	Cholesterol responsive atherosclerosis gene
CVD	Cardiovascular diseases
DNA	Deoxyribonucleic acid
FDR	False discovery rate
HDL	High density lipoprotein
IMT	Intima-media thickness
LDL	Low density lipoprotein
MAS	Microarray analysis suite
MI	Myocardial infarction
mRNA	Messenger RNA
pI-pC	polyinosinic-polycytidylic acid
RMA	Robust multi-array average
RNA	Ribonucleic acid
SNP	Single nucleotide polymorphism
STAGE	Stockholm atherosclerosis gene expression
TEML	Transendothelial migration of leukocyte pathway
VLDL	Very low density lipoprotein

Contents

1 Introduction

1.1 Background

Up until recently the approach most widely used to investigate the genetic basis of disease has been the candidate gene approach, in one or a few pathway genes are selected and studied in relation to disease. This kind of research has yielded many important medical insights and has led to diagnostics and therapies for a wide range of diseases. Although successful, the candidate gene approach has limitations. In particular, it is of limited value for studying the complex molecular etiology of cancer, cardiovascular disease, neurodegenerative disorders, and other common diseases.

In the last decade, an improved understanding of the genome and its DNA sequence have paved the way for less biased approaches than the candidate gene approach. For instance, many genetic variants that cause so-called single-gene diseases have been identified by determining how DNA markers situated throughout the genome migrate in families with heritable diseases [1, 2]. This success has fueled efforts to use DNA markers and, more recently, dense maps of several hundred thousand single nucleotide polymorphisms (SNPs) to disclose the genetic variations underlying complex diseases [3, 4].

However, pure DNA-based approaches to complex disease are probably not sufficient since the development of complex diseases are reflects environmental influences such as lifestyle choices. It is now becoming increasingly clear that to fully understand the molecular causes of complex diseases, a more holistic approach must be adopted that also takes into consideration functional aspects of the genome [5–8]. The function of the genome is governed by the activity of all genes at a given time, which is a consequence the genetic makeup as well as the environmental pressure at that moment. Thus, by measuring genome activity it is possible to capture both environmental and genetic aspects of disease development.

In parallel with the sequencing of several genomes, including human [9, 10] and mouse [11], new technologies have been developed to study the activity of the genome. The most developed of these technologies is microarray analysis of mRNA concentrations [12, 13]. Since mRNAs are to a large extent inactive messages to encode proteins, great effort

has been expended to develop technologies to measure concentration of proteins [14]. However, proteomic technologies are still far less robust and reliable than gene expression analyses [15].

In this thesis, we performed expression profiling using Affymetrix GeneChips in relevant tissue samples isolated from patients with severe atherosclerosis and in a mouse models with human-like atherosclerosis. In these expression profiles we have used computational methods to identify functionally associated genes of importance to atherosclerosis.

This work has presented a number of challenges. For example, atherosclerosis is a disease in which three major cell types dominate: endothelial cells, smooth muscle cells, and different forms of leukocytes [16]. The relative amounts of these cell types differ from biopsy to biopsy, and in clinical studies, the phenotype of the patients varies. These factors, together with the technical problems associated with using microarrays to generate high-dimensional data, contribute to variation in the mRNA levels that are unrelated to the biological phenomenon under study (i.e., atherosclerosis). We have tried to minimize the effect of the technical noise by using normalization techniques and by choosing reliable sequence-matching procedures [17–19].

As already alluded to, the goal of this thesis was not to isolate novel individual atherosclerosis candidate genes. Rather, we used in-house and previously developed algorithms to identify functionally associated atherosclerosis genes and their interplay in modules and networks. To improve the reliability of this approach, we also used literature mining and databases of gene-gene and protein-protein interactions.

1.2 Cardiovascular diseases and atherosclerosis

Cardiovascular diseases are a collection of disorders that involve the heart or blood vessels, including both arteries and veins. Cardiovascular diseases, the most common cause of mortality globally, caused the death of an estimated 17.5 million people in 2005, most often from myocardial infarction (MI) or stroke [20].

Atherosclerosis—the major cause of cardiovascular disease—can be described as deposits of lipids that accumulate in the intima of the artery wall. The immune response to these lipids leads to the formation of “plaques”. Although plaques can grow large enough

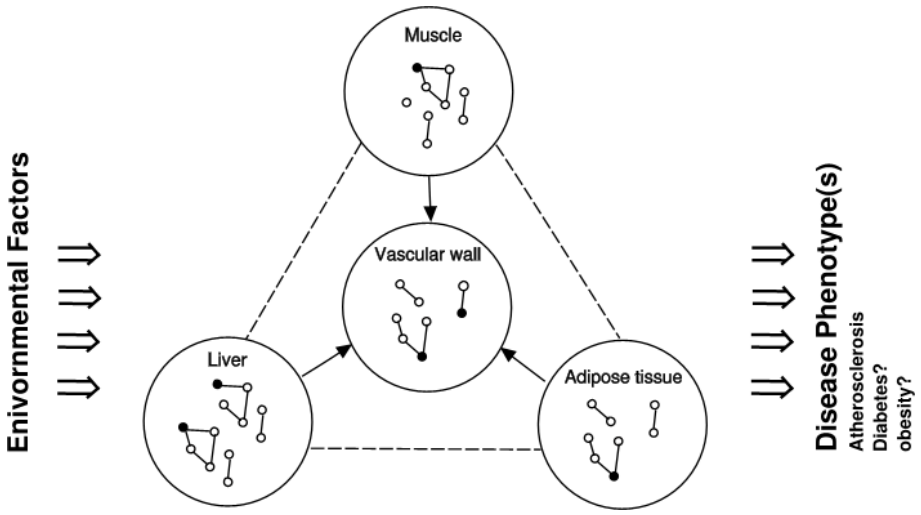

Figure 1: Atheroscerosis is a complex disease with manifestation in the vascular wall and involving multiple other organs (e.g., liver, skeletal muscle and adipose tissue). Each organ is controlled by gene networks in that modulates that organ function and indirect influences atherosclerosis development in the vascular wall. Genes marked are modulated by genetic variation(s) in the population and their function is therefore differnt between individuals. Environmental factors like life-style choices are filtered through this disease system causing disease phenotypes.

to significantly reduce blood flow, the most important complication is the formation of blood clots as a result of a plaque rupture. Such clots can completely block distal blood flow in the coronary arteries, leading to an MI. If part of the clot breaks off, it can travel through the circulation to the fine arteries of the brain, where it can restrict blood flow and cause a stroke.

Atherosclerosis is a progressive, lifelong disease that involves multiple organs and is influenced by several genetic and environmental risk factors [21] (Figure 1). Of particular relevance to atherogenesis is the metabolism of lipids and cholesterol [22] and glucose [23, 24]. Elevated blood glucose levels and insulin resistance are dependent on glucose uptake and metabolism, especially in the skeletal muscle and adipose tissue. Increased fat deposits, in particular abdominal and other visceral fat, lead to increasing levels of circulating free fatty acids, which in turn are believed to have lipotoxic effect on nonadipose tissues, including the pancreas, where insulin is synthesized. States of insulin resistance and increased blood glucose, as in diabetes mellitus and the metabolic syndrome, are associated with increased risk of atherosclerosis [24].

Cholesterol can be synthesized in most cells, but the liver is the main source of plasma cholesterol, which is secreted into the blood stream as very low density lipoprotein (VLDL) particles. These particles are delipidated by lipases anchored to the endothelial surface, leading to the generation of cholesterol-rich low density lipoprotein (LDL) particles that can become trapped in the arterial wall, as explained below, leading to atherosclerosis development. High density lipoprotein (HDL) particles, which are thought to be produced and secreted mainly by the liver and intestine [25, 26], work the opposite way. In a process called reverse cholesterol transport, they unload cholesterol from the atherosclerotic plaques and transport it back to the liver [27]). Thus, to understand the development of atherosclerosis, it is of great interest to monitor the metabolic and regulatory states of the liver, skeletal muscle, and adipose tissues.

The pathology of atherosclerosis progression in the vascular wall is complex, not least because several cell types are involved. Of particular importance are inflammatory cells, such as leukocytes (e.g., monocytes and T-cells). Early accumulation of LDL particles in the intima of the arterial wall seems to occur as a consequence of mechanical stress to the arterial wall mediated by blood flow, particularly at sites of bifurcations and changes in the direction of blood flow. Within the intima, the LDL particles are modified, most importantly by oxidation. The modified particles are pro-inflammatory and stimulate endothelial cells to increase their expression of adhesion molecules, which induce circulating monocytes and lymphocytes to migrate into the intima. The pro-inflammatory state within the intima causes the monocytes to differentiate into macrophages, which express surface receptors (e.g., *CD36*) that mediate the internalization of modified LDL. These early events lead to the formation of fatty streaks (so named because of their appearance by microscopic inspection of the arterial wall), consisting of so-called foam cells—macrophages in which intracellular lipids, mainly cholesteryl-esters, have accumulated. [16, 28–30]

Later in atherosclerosis development, the continuing uptake of modified LDL particles by macrophages, now at an increased rate, leads to the accumulation of foam cells, which become necrotic and in some cases undergo apoptotic cell death, creating a necrotic core within the plaque. The core is surrounded by a fibrotic cap consisting of smooth muscle cells and fibrin. The smooth muscle cells are not part of the normal intima, but at this

later stage of plaque development they migrate in from the media. Together, the necrotic core and the fibrous cap form the mature plaque. [16, 29]

From a simplistic perspective, there are three types of atherosclerosis research. The first type, clinical studies of patients with atherosclerosis, is highly relevant because human atherosclerosis is indeed what we seek to learn more about. However, studies in humans have limitations. The patients are often in very late stages of the disease, making it difficult to study the early phases of atherogenesis, and in most instances, repeated measurements (i.e., analysis over time) are not feasible. Nor is it possible to perform genetic manipulations or other biological perturbations.

The second type of research uses animal models of atherosclerosis, which allow for both analysis over time and genetic manipulations. However, genetic manipulations (i.e., conditional or permanent knockouts or transgenes) are time consuming, costly, and, in most instances, are not cell specific. Although animal studies of atherosclerosis have been and will continue to be important, they are limited by their lesser biological relevance than human studies.

The third type involves in vitro studies of the cell types affected by atherosclerosis, such as macrophages and endothelial cells. Cell model systems are important for deciphering the true "wiring diagram" of functionally associated genes in atherosclerosis. Many forms of perturbations can be performed much faster and at less expense in cell systems than in animal models. However, cultured cells do not recapitulate the in vivo context of cells active in atherosclerosis, and therefore the results need to be interpreted with caution.

To achieve a balance between disease relevance and the flexibility needed to infer biological networks, we believe it is most appropriate to first identify networks and key nodes in the human and animal model setting and then move to cell model systems to infer the actual web of biological interactions in the network surrounding key nodes [5].

1.3 The structure of biological networks

The sequencing of the human genome revealed that humans have $20,000 - 25,000$ genes [31]—fewer than expected from earlier estimates and comparable to the number of genes in less complex organisms, such as rice, fruit fly, and yeast [32–35]. This would be re-

markable if the complexity of an organism were proportional to the number of genes in its genome. However, the complexity of an organism is probably related to the number of states its genome can assume [36], which is determined by interaction networks of genes, gene products (RNA, proteins), and metabolites [7, 37]. Thus, to understand the functions of the genetic code—including inheritable diseases in humans—we must discover the structure and dynamics of the complex web of interactions between cellular species.

In the last decade, we have seen increased efforts to identify and understand the static structure of cellular networks, including protein-protein interaction [38,39], transcriptional regulation [40–42], and metabolic reaction [43, 44] networks. These networks commonly exhibit a so-called power law degree distribution, whereby most genes have low connectivity, and a few genes—referred to as hubs—are highly connected [45,46]. Hub genes are of particular importance, for example deletion of a hub gene in a protein-protein network is more likely to be lethal than deletion of a non-hub gene [47]. Cellular networks also exhibit a "small world" structure [48], which means the path between any two nodes is short, even in large networks. Moreover, networks are organized in modules ordered in a hierarchical manner [49]. Modules in networks can be loosely defined as groups of nodes with a significantly higher interconnectivity and lower intraconnectivity than in the network as a whole. Finally, some networks are enriched in certain motifs (basic building blocks), which may have a specific function in network information processing [50].

In this thesis, we are studying the state of the cell in terms of mRNA levels a major determinant for these levels have be attained to the transcriptional network [40–42]. However, there are other sources of regulation, including protein interactions, metabolic states, noncoding RNA, DNA accessability (chromatin structure), and translational regulation. We believe that relating gene expression to network structure and dynamics is a promising field, but it must be acknowledged that the underlying gene regulatory network is far more complex than the transcriptional network alone [51] In Section 1.5 we will discuss how mRNA measurements on a global scale can be used to identify structural features, including modules, in the gene network.

1.4 Microarray technology

The term gene expression is used to describe the transcriptional state of the genome in terms of mRNA concentrations at a given time point. There are several methods for measuring mRNA concentrations, including (in chronological order) northern blots, reverse transcription–polymerase chain reaction, microarray analyses, and serial analysis of gene expression. The first two techniques have been used mainly on a small scale to investigate the expression of one or a couple of genes at the same time. Measurement of mRNA levels on a chip was first done in the mid 1990's [13], using a glass slide imprinted with cDNA for 48 genes. Rapid advances in this area have given us the ability to simultaneously measure the expression of tens of thousand of genes simultaneously in a rapid, cost-effective, and fairly reproducible way. There are currently several competing microarray technologies for measuring gene expression, including bead arrays [52], oligonucleotide arrays [12], and cDNA arrays [13].

Microarray technology utilize the ability of an RNA (or single stranded DNA) molecule to hybridize to specific DNA probes with complementary sequence. The microarray has several probe areas containing DNA with specific sequences attached in an array (often a glass or membrane slide). By labeling the sample mRNA with radioactive or fluorescent dye, the amount of hybridized mRNA on each probe area can be detected or estimated with a scanner. The amount of hybridized mRNA in a probe area is proportional to the amount of the mRNA with complementary sequences (and thus gene expression) in the sample.

In the current study, we use Affymetrix oligonucleotide GeneChips [12] to measure global gene expression in samples originating from human and mouse tissue. Before the biological sample is hybridized onto the array, total RNA is isolated, reverse transcribed to cDNA, transcribed to cRNA, labeled with biotin, and fragmented. Finally, a hybridization cocktail with the labeled cRNA fragments is hybridized onto the chip, excess cRNA is washed away, and the hybridized chip is stained and scanned. [53]

The HG U133 Plus 2.0 and MG 430 2.0 GeneChips contain 1,354,896 and 1,004,004 probes, respectively. Each probe contains 25-mer oligonucleotides that match a subsequence of the interrogated mRNA. There are two type of probes: perfect-match probes,

which correspond exactly to the mRNA being interrogated, and mismatch probes, which have an identical sequence except for a substitution at the central base. The mismatch probes are intended to measure nonspecific binding; however, the estimates they provide are not very good [19, 54] and are therefore ignored by some probe summarizing algorithms.

Affymetrix probe sets are defined using Unigene [55] transcripts and most commonly contain 11 probe-pairs, containing one perfect match and one mismatch probe. After the release of Affymetrix probe sequences, researchers could start addressing problems with probe set definitions. Surprisingly, many Affymetrix probe sets on mammalian microarrays did not correspond correctly to the appropriate reference sequence (RefSeq) [56–58]. Moreover, sequence-verified probes (matching RefSeq mRNA) provided more accurate measurements than unverified probes [17, 56]. Inspired by these results, we created our own probe set maps based on RefSeq [58] and Entrez Gene [59] in Study III and IV.

Affymetrix provides a software package, Microarray Analysis Suite (MAS; currently version 5), for preprocessing, normalizing, and summarizing probe intensity values into probe set expression values [60]. However, several other normalization and probe-summarizing techniques may yield "less noisy" expression values [18, 19, 61–64].

Measuring mRNA concentrations does not provide enough resolution to distinguish between the individual regulatory networks (transcription, protein-protein binding, and metabolic networks). Instead, it gives a more coarse-grained picture of an effective gene-to-gene regulatory network. This may explain, at least in part, the problems encountered in validating the results of microarray studies. Measurements on multicellular tissues like vascular wall samples will further complicate the interpretation.

1.5 Identification of genes, modules and networks from whole genome expression profiles

As mentioned previously, whole-genome expression studies generate vast amounts of data, and it is important to extract as much significant biological information as possible. In this section, I will describe three approaches for analyzing gene expression data. First, I will discuss gene-by-gene statistical tests, focusing on those used to identify differentially

expressed genes. Second, I will review clustering approaches for identifying gene modules. Finally, I will discuss methods for inferring and integrating gene networks from expression data.

1.5.1 Gene-by-gene statistical tests

In many studies, the aim has been to identify differentially expressed genes—those with different expression levels between two distinct conditions. In such studies, gene expression under each condition is measured with microarrays, and a statistical test is applied in a gene-by-gene manner to find differentially expressed genes. The major problem in applying standard statistical methods to the microarray setting is that performing multiple tests and controlling the false-positive rate will result significant fraction of false predictions. The family-wise error rate, defined as the probability of having at least one false positive in all tests, has been argued to be too strict and may miss many genes that are, in fact, differentially expressed, resulting in a large fraction false negatives [65]. Statisticians have dealt with this problem by developing methods to control (or estimate) the false discovery rate (FDR)—the expected fraction of false positives—in the predicted set of differentially expressed genes [66–68].

1.5.2 Gene module identification

It is well known that genes and gene products do not act in isolation. Instead, cellular functions are carried out in functional modules [69]. Modules can be identified in the structure of cellular networks and are organized in a hierarchical manner [49, 70]. Moreover, functionally related gene modules can be identified directly from gene expression data through clustering algorithms [71, 72], a process in which genes with similar gene expression profiles are grouped together. Clustering can be applied to observational data alone. Specific perturbations are not required, but meaningful clusters are only obtained for genes that change expression state across the samples.

Clustering algorithms use similarity/dis-similarity measures to compare two data vectors (i.e., gene expression profiles). The most popular ones are based on Euclidean distance, Manhattan distance, Pearson correlation, and Spearman rank correlation. More-

over, several methods for clustering gene expression data are available [72]. Two of the most commonly used are hierarchical clustering and K-means clustering. K-means clustering algorithms divide the gene profiles into k non-overlapping groups (k is supplied by the user). Hierarchical clustering produces a cluster tree appropriate for visualisation purposes. Owing to the nature of gene expression data and the underlying cellular processes, many commonly used clustering methods are not well suited for this problem [73]. Problems that the algorithms and distance measures must accommodate include noisy data measurements with outlier data points, the presence of irrelevant genes, and genes belonging to multiple modules [73].

Several studies have applied standard clustering algorithms to gene expression data to identify gene modules [74–76]. More advanced clustering applications include Segal et al. [77] introduced a gene clustering method to enable the identification of regulatory trees that explain the regulation of each gene cluster. Application of this method to data from 22 different tumor types identified modules generally important for most cancer types, as well as specific modules important in one cancer type [78]. Another interesting method [79, 80] is coupled two-way clustering to identify biologically relevant submatrices in the gene expression data matrix.

1.5.3 Gene network identification

As will be argued throughout this thesis, knowledge of gene networks is instrumental to fully understand biology, cellular function, and complex diseases. Clustering techniques have the ability to capture the modularity in the underlying gene network. However, clustering approaches cannot be used to unravel specific gene-gene edges, as these approaches are unable to differentiate between direct and indirect interactions [81]. Nevertheless, correlation measurements have been used to infer gene networks directly from gene expression data, often using partial correlations [82] to minimize problems with indirect interactions. Other approaches for gene network identification include systems of ordinary differential equations [83–86], bayesian inference [87, 88], and information theoretic models based on mutual information [81, 89].

Because the number of possible edges in a network grows exponentially with the num-

ber of nodes, identifying networks that contain many nodes is not a simple task. The problem becomes even more difficult as the desired resolution increases [37]. To overcome these hurdles, one can limit the search space by including what is known about network structure and individual edges, for example by limiting the in-degree of each node to a small number [83, 84], or integrating information about known network edges.

Several research groups have taken the use of prior knowledge one step further using available network structure in the public domain and integrate it with gene expression data to identify sub structures important to the studied biological process. Luscombe and coworkers [42] take this approach, they use the transcription network in yeast and identify subnetworks active during cell cycle, sporulation, diauxic shift, DNA damage and stress response. In simular study conducted by Lichtenberg et al. [90], they integrated protein–protein interactions and gene expression measurement at different stages of the yeast cell cycle reslting in a dynamic map of protein complexes. In a recent study of breast cancer they were able to identify a disease network from gene expression and various network sources [91]. From this disease network they identified a new breast cancer suseptibility gene*HMMR*, which was also functionally and genetically validated.

2 Aim

The overall aim of this thesis is to use a top-down approach to uncover functional modules and gene networks important to atherosclerosis.

Specific aims of the individual papers:

I. To elucidate the role and relevance of the effective regulatory gene network and investigate whether there are genes with a high in-degree (in-hubs) in that network.

II. To reveal the transcription repertoire of atherosclerosis development and examine the effect of a subacute lowering of plasma cholesterol on lesion development and gene expression.

III. To reveal the transcription repertoire in multiple organs relevant to coronary artery disease (CAD) and to identify modules of functionally associated genes important in atherosclerosis development.

IV. To identify the full repertoire of plasma cholesterol-responsive atherosclerosis genes and their interplay in gene networks.

3 Results and Methods, Study I-IV

In this section, I focus on results and methods from four studies from the perspective of my area of expertise and interest—the analysis of whole-genome expression data. Of note, there are other results in these studies that are interesting and merit discussion. Before surveying each study, I will describe the patient cohorts and the mouse models we used throughout the thesis.

3.1 Human cohorts

3.1.1 The Stockholm Atherosclerosis Gene Expression cohort

The Stockholm Atherosclerosis Gene Expression (STAGE) cohort consists of 114 well-characterized patients who underwent coronary artery bypass grafting (CABG) at Karolinska University Hospital, Solna. In 66 of these patients, we measured whole-genome expression using HG U133 Plus 2.0 (Affymetrix Inc.) in biopsies obtained from liver, adipose tissue, and skeletal muscle during the surgery. In 40 of these 66 patients we also measured atherosclerotic gene expression from aortic root and the mammary artery. From these two expression profiles we defined atherosclerotic tissue gene expression as aortic root expression with mammary artery expression subtracted. All patients were well characterized with anthropometric and biochemical measurements, medical records and history, information on lifestyle factors (e.g., smoking, alcohol consumption, and physical activity), and coronary angiograms. The coronary angiograms were evaluated with quantitative coronary angiography techniques to obtain a surrogate measure of atherosclerosis burden, referred to as the stenosis score.

3.1.2 Carotid cohort

The carotid cohort consists of 42 patients who underwent carotid surgery at Stockholm South General Hospital. In 25 of the patients, we measured whole-genome expression using HG U133 Plus 2.0 (Affymetrix Inc.) in the carotid lesion removed during surgery. This cohort is as well characterized as the STAGE cohort. Intima-media thickness (IMT) was used as a surrogate measure of atherosclerosis burden.

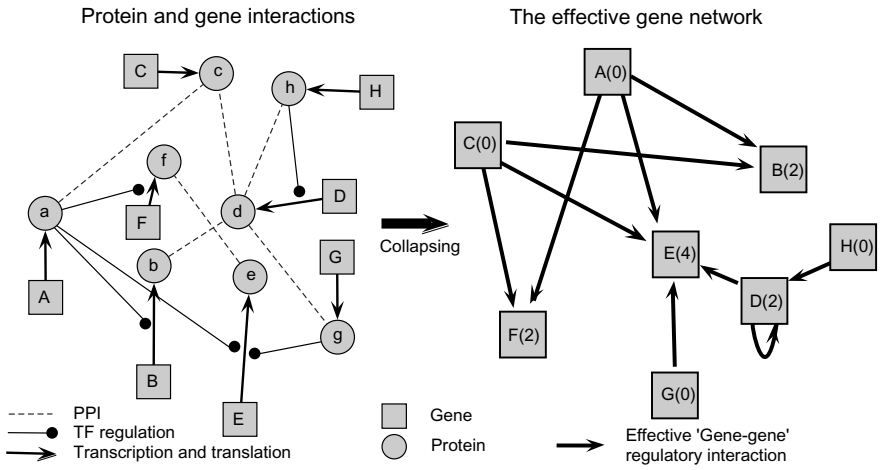

Figure 2: Estimating the effective gene network underlying gene expression (mRNA) by integrating transcription and protein binding edges. The network of interactions between eight hypothetical genes and their respective proteins (left) is collapsed into an effective regulatory gene network (right).

3.2 The Atherosclerosis Mouse Model

We are using the $Ldlr^{-/-}$ $Apob^{100/100}$ $Mttp^{flox/flox}$ Mx1-*Cre* mouse model [92], which develops atherosclerosis on a chow diet. The $Ldlr^{-/-}$ and $Apob^{100/100}$ modifications results in a plasma lipoprotein profile similar to that of patients with familial hypercholesterolemia [92]. The floxed Mttp ($Mttp^{flox/flox}$) and the inducible transgenic expression of Cre recombinase (Mx1-*Cre*) together constitute a genetic switch to turn off the hepatic synthesis of VLDL at a selected time point. This is accomplished by treating mice with polyinosinic-polycytidylic acid (pI-pC) to activate the Mx1 promoter in the liver, leading to Cre recombinase synthesis. Cre recombinase recognises the floxed sites and recombines, causing an ($Mttp^{\Delta/\Delta}$) phenotype and a rapid reduction of plasma cholesterol levels by 80% or more.

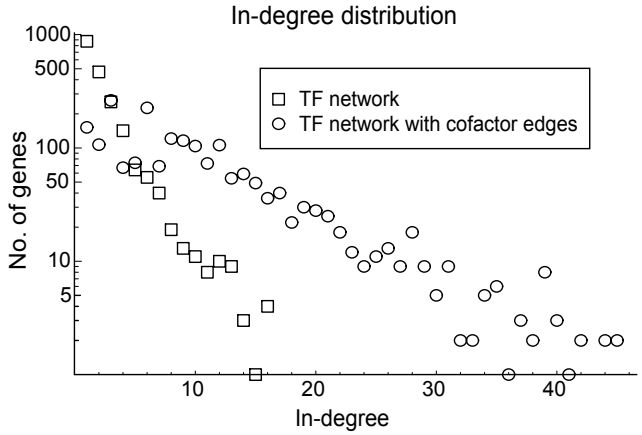

Figure 3: Illustrating the number of incoming regulatory edges in yeast considering the TF network alone (squares) and including proteins binding to TF influencing TF activity (circles).

3.3 In-hubs and the effective gene network (Study I)

In this study, we integrate transcription regulation networks [45] with protein-protein interactions from the Database of Interacting Proteins [93,94] to estimate the effective gene network underlying gene expression data. In the effective gene network, we show evidence of in-hubs and provide a method for predicting in-hubs directly from gene expression data.

The out-degree distribution of the transcription network, it has been suggested, is broad and the in-degree distribution narrow [40]. While this may be true for the transcription network, we are interested in the effective regulatory gene network underlying gene expression data. This network is, as mentioned previously (Section 1.3), influenced not only by the transcription edges but also by networks of interacting proteins, metabolic reactions, and signaling cascades. While it is known that these processes influence gene activity, estimation of the effective regulatory from available network-sources has, to our knowledge, not been addressed before. In study I, we took a first step toward an improved estimation of the effective gene regulatory network by adding transcription co-factor proteins known to bind transcription factors as potential regulators (Figure 2). In this regulatory network underlying gene expression as measured by mRNA, we found evidence of a broader in-degree (Figure 3) with some genes having up to 40 regulators.

In the second part of this study, we present a method that uses gene expression data to separate genes with a high in-degree (in-hubs) from genes with a low in-degree. The rationale behind the method is that genes with a high in-degree are more likely to be affected by random and repeated perturbations of the network. We evaluated the method

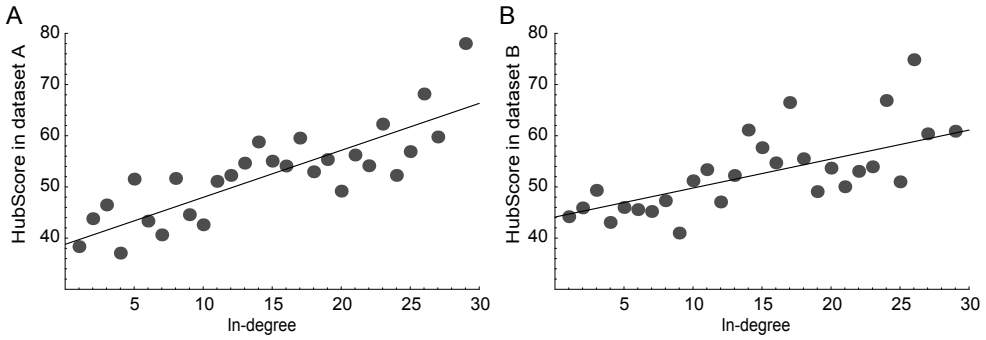

Figure 4: Prediction of genes with high indegree in two independent gene yeast expression datasets The number of regulatory interactions as calculated from the transcription network with co-factor proteins added, shown as a function of the HubScore. (A) 273 gene expression profiles from nonlethal gene deletions [95] and (B) 215 expression profiles generated from yeast cultures with titratable promoters of genes essential for cell survival.

by using computer simulations and a linear model of the regulatory gene network and validated it in two Saccharomyces cerevisiae expression datasets [95, 96] This validation revealed a significant correlation between the method output (HubScore) and in-degree in both datasets (Figure 4).

3.4 Study of atherosclerosis development in mouse model (Study II)

In this study, we used the $Ldlr^{-/-}$ $Apob^{100/100}$ $Mttp^{flox/flox}$ Mx1-*Cre* mouse model (Section 3.2) to follow the development of atherosclerosis in the aorta. The mice were sacrificed at 10, 20, 30, 40, 50, and 60 weeks of age, and the extent and histological appearance of the lesions were carefully determined at each time point. The data were combined with gene expression profiles isolated from the atherosclerotic aortic arch in a parallel set of mice. In this fashion, gene expression changes were coupled to plaque development.

Lesion area during atherosclerosis progression was quantitatively studied by Sudan IV staining of whole aortas harvested from 87 mice evenly distributed over the six time points (Figure 5). Lesion formation progressed slowly at first, reaching an average lesion area of 5% at 30 weeks, and then expanded rapidly to an average lesion area of 12% at 40 weeks ($P < 0.0001$). Thereafter, lesion area reached a plateau at less than 20%.

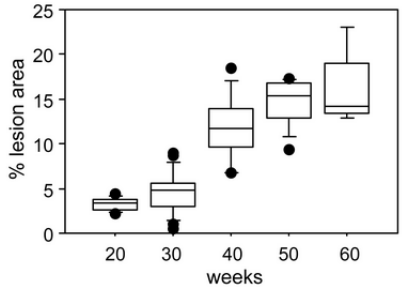

Figure 5: Lesion expansion during atherogenesis. Values are surface lesion areas assessed by Sudan IV staining of pinned-out aortas and given as a percentage of the surface of the entire aorta.

To investigate the transcriptional repertoire of atherosclerosis progression, we isolated total RNA from the atherosclerotic aortic arch (defined as the aorta from the aortic root to the ascending aorta at the 3rd rib) from 32 mice (4–7 mice per time point) and obtained global gene expression profiles with Affymetrix GeneChips (MG 430 2.0). The data were analyzed with an empirical Bayes test [97] ($FDR < 0.05$) to detect genes that were differentially expressed between any of the time points 10, 20, 30, 40, 50, and 60 weeks. In all, 1259 genes met this criterion and were therefore considered to be active during atherogenesis. The expression profiles of this set of genes were clustered into four gene groups using K-medoids clustering [98, 99] (Figure 6).

The two most interesting clusters in terms of what is already known about atherosclerosis were cluster 1 and 3. Cluster 1 consisted of 293 genes that were expressed at a low level at the early time points, were activated at 30 week,s and remained active throughout the 60 week time point (Figure 6). Text mining [100] revealed that genes in cluster 1 were associated with atherosclerosis (36%) and macrophages (44%). The latter association was also reflected in functional analysis performed in DAVID [101], which showed enrichment in immune and inflammatory activities.

Cluster 3 contained 331 genes, of which 27% were previously associated with atherosclerosis (Figure 6). Genes in this cluster were activated at 30 weeks and were deactivated at 40 weeks, when lesion area began to expand rapidly, as shown in Figure 5. Gene-annotation enrichment analysis with DAVID [101] suggested that a majority of these genes were involved in lipid metabolism.

To control for changes in cellular composition of the atherosclerotic lesion over time,

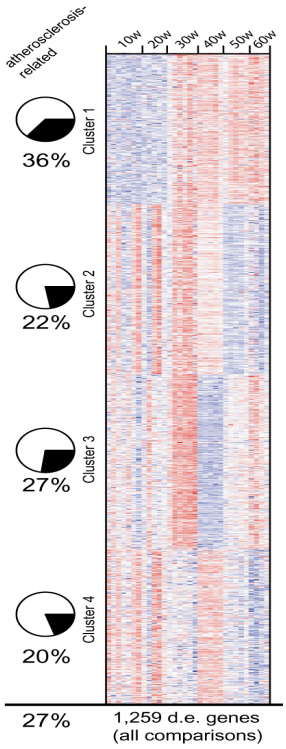

Figure 6: Genes active during atherogenesis. Heat map of 1259 genes differentially expressed in at least one pair-wise time-point comparison ($FDR < 0.05$). Expression levels are color coded red indicating high expression and blueindicating low expression. Each column represents mRNA levels in one mouse at 10 to 60 weeks of age (n = 5 to 7 per time point) sorted by time. Each row represents mRNA levels for one gene clustered according to expression similarity into four clusters. Pie charts show percentages of genes related to atherosclerosis in each cluster.

the mRNA levels of cell-type specific markers were assessed at different time-points. By averaging the mRNA levels of five to 11 markers per cell type (i.e., endothelial cells, monocytes/macrophages, smooth muscle cells, and T-cells), we estimated the relative contribution of these major atherosclerosis cell types over time (Figure 7). To our surprise, the contributions were fairly stable over time, the exception being the relative content of lesion macrophages, which increased significantly between 20 and 30 weeks, just before the rapid expansion of the lesions (Figure 5).

From the combined histological and expression profile analyses, it was clear that the 30-week time point was critical for atherosclerosis development in these mice. In brief, these analyses showed that up until 30 weeks, macrophages accumulated in the arterial wall. At 30 weeks, this accumulation seemed to have reached a critical point, leading to formation of small but well-defined plaques. At 30 weeks, lesion inflammation and immune reactions were strongly activated (cluster 1, Figure 6) and remained activated throughout the study period of 60 weeks. This inflammation may be a trigger for the rapid increase in plaque size between 30 and 40 weeks (Figure 5). The increase in plaque size is primarily caused by an enlargement of existing macrophages due to cholesterol-ester accumulation.

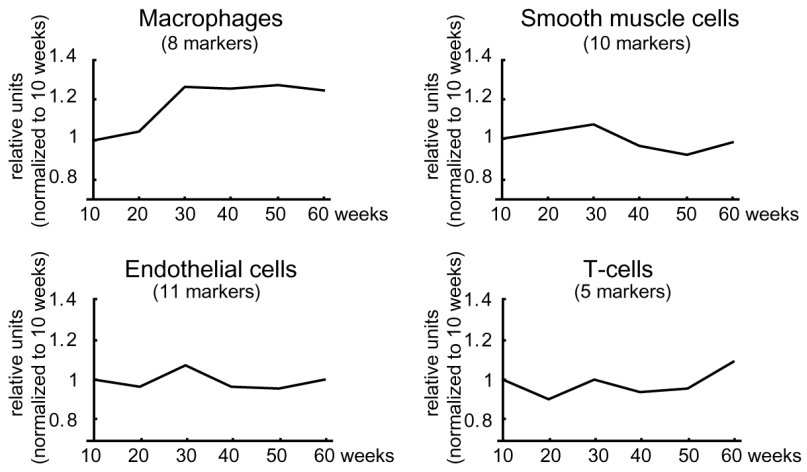

Figure 7: Relative expression levels of cell-specific markers of four atherosclerosis cell types. The number of markers per cell type is indicated. The only statistically significant increase was in the number of foam cells, which increased by 30% between 20 and 30 weeks (P¡0.001) and remained elevated at 60 weeks.

We therefore chose to lower plasma cholesterol at the 30-week time point, using the inducible transgene (Mx-1 *Cre*) and floxed Mttp alleles in the mouse model. Cholesterol levels were decreased by 80% or more at 30 weeks, and this reduction completely prevented the rapid expansion of the lesions between 30 and 40 weeks of age (see paper II). This dramatic effect suggests that cholesterol-lowering drugs such as statins may have an even more potent effect than has already been documented [102] if administrated early to patients at risk for premature atherosclerosis.

Gene expression profiling of mice with high and low plasma cholesterol at 30 weeks identified 38 genes that were responsive to cholesterol lowering. ($FDR < 0.05$) . Some of these genes were perturbed—using silencing interference RNA (siRNA)—in a cell culture of THP-1 macrophages incubated with acetylated LDL. Gene expression profiling of the perturbed cell cultures and reverse engineering [84] resulted in a network of eight genes important for foam cell formation and the rapid-expansion phase of lesions in these mice (see Figure 4 in Paper II).

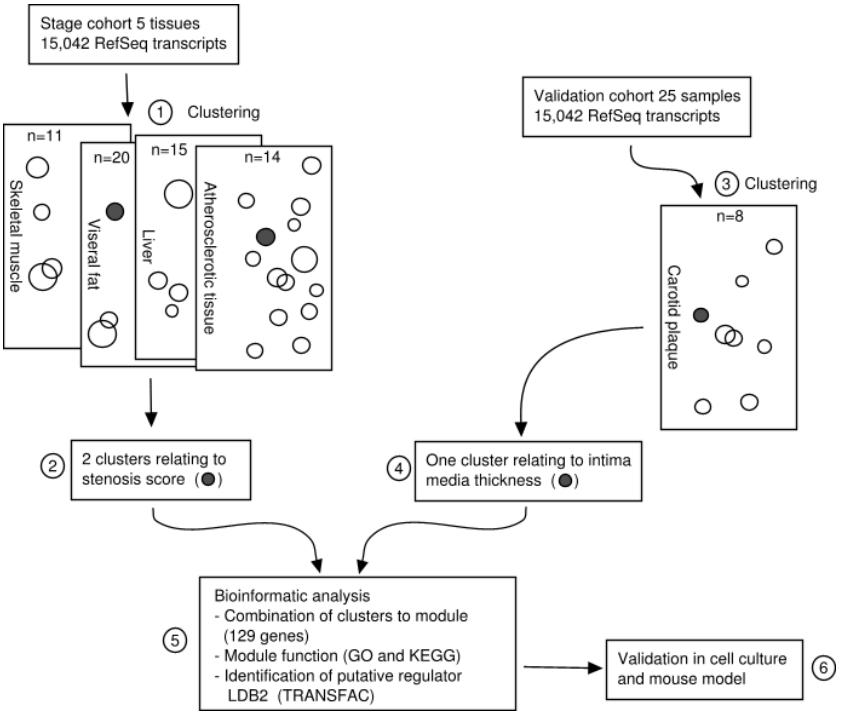

Figure 8: A principle scheme of the analytical steps performed in Study III. (1) Sixty-six gene profiles (15,042 RefSeq each) from the liver, skeletal muscle and mediastinal fat and 40 from the atherosclerotic tissue were clustered with a coupled two-way approach. First, the RefSeq expression profiles were clustered, separately in each tissue, according to pairwise Spearman rank correlation, resulting in 11–20 gene clusters for each tissue. Each gene cluster was then clustered dividing the patients into two groups according to mRNA levels only in the cluster genes. (2) Two gene clusters that divided the patients according to the degree of coronary stenosis were further analyzed. (3) To validate the atherosclerosis-related clusters identified in the STAGE cohort, the clustering procedure was applied to a validation cohort containing 25 carotid stenosis patients with gene expression profiles of lesion gene expression. Instead of degree of coronary stenosis, IMT was used to define clusters relevant to coronary artery disease. (4) The first clustering step resulted in 8 gene clusters, only one divided the patient according to IMT. (5) Bioinformatic analysis revealed a significant overlap between all three clusters and similar functional annotation, indicating all clusters are from the same functional module. *LDB2* is identified as a putative regulator. (6) Functional validation of *LDB2* as a key regulator of the module.

3.5 Identifying a gene-module relevant to atherosclerosis severity in CABG patients (Study III)

In this study, we analyze whole-genome expression profiles from the STAGE study (Section 3.1.1) using a clustering approach inspired by Getz et al. [79], which identifies functional gene modules important to disease development, in our case atherosclerosis.

The outline of the data analysis is shown in Figure 8. First, we identified correlated mRNA levels by grouping genes in 11 to 20 gene-clusters per tissue with a super-paramagnetic clustering algorithm [103, 104], using the absolute value of Spearman rank correlation as the similarity measure. Among the benefits of this method, is that it does not assume a underlying data distribution and that it allows each data point to belong to more than one cluster. For each gene-cluster containing up to 1000 genes, we clustered the patients into two groups based on the mRNA levels of the genes in that cluster.

Two gene clusters divided the patients into groups with statistically significant difference in stenosis score. One cluster was identified from the expression profiles of the atherosclerotic tissue ($n = 49$ genes, $P = 0.008$) and the other from the profiles in mediastinal fat ($n = 59$ genes, $P = 0.00015$), see Figure 9. Interestingly, seven gene were present in both clusters, which is highly unlikely to happen by chance ($P < 10^{-9}$).

To validate these results, we also measured and analyzed whole-genome expression profiles of carotid biopsies from 25 patients that underwent carotid surgery (see Section 3.1.2). Of eight gene-clusters identified, one ($n = 55$ genes; Figure 10) divided the patients into two groups that differed significantly in IMT score ($P = 0.04$). Remarkably, the overlap between this gene cluster and the two clusters identified in mediastinal fat and atherosclerotic aorta in the CABG patients included 16 and 17 genes, respectively ($P < 10^{-26}$, $P < 10^{-29}$).

We believe clustering analyses have uncovered a module of 129 genes relevant to atherosclerosis development, 28 with evidence from expression measurements in two or three different tissues. Gene set enrichment annotation [101] suggests that this module and its 129 genes are likely to be involved in the KEGG pathway transendothelial migration of leukocytes ($P < 10^{-6}$).

The only regulatory gene present in all three clusters was the transcription co-factor

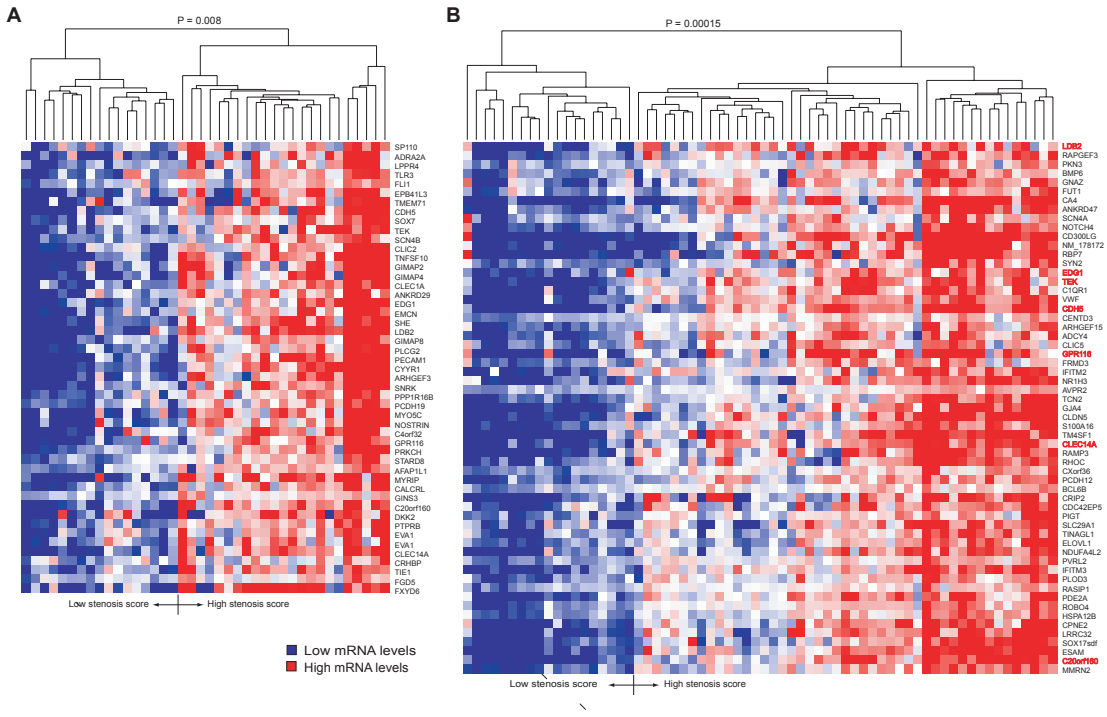

Figure 9: Heat maps of two clusters related to stenosis score in coronary artery bypass graft (CAGB) patient expression profiles. Columns represent individual patients, and rows represent RefSeq transcripts. Levels of mRNA are represented as a color; brighter blue indicates lower mRNA levels, and brighter red higher mRNA levels. (A) Heat map indicating the mRNA levels of 49 genes in atherosclerotic tissue belonging to the one cluster out of 14 that related to the stenosis score ($P = 0.008$). (B) Heat map indicating the mRNA levels of 59 genes in medistinal fat belonging to the one cluster out of 20 that related to the stenosis score ($P = 0.00015$). Highlighted genes were also found in the cluster shown in panel (A).

LIM-domain binding 2 (*LDB2*). To investigate the role of LDB2 as a potential regulator of this module, we identified transcription factors that interact with LDB2 and matched their binding site sequences from TRANSFAC (v11.2) [105] with the upstream sequences of the 129 genes in the module. We found that 122 of the genes could theoretically be regulated by LDB2. Furthermore, cell culture and immunohistochemical analyses revealed that *LDB2* is expressed in two key cell types of atherosclerosis—endothelial cells and macrophages. Finally, we examined the mRNA levels of 10 genes central to transendothelial migration of leukocytes in the arterial wall of *Ldb2* knockout and wildtype mice. Eight genes were differentially expressed; the difference in expression levels of five of the genes was statistically significant ($p < 0.05$).

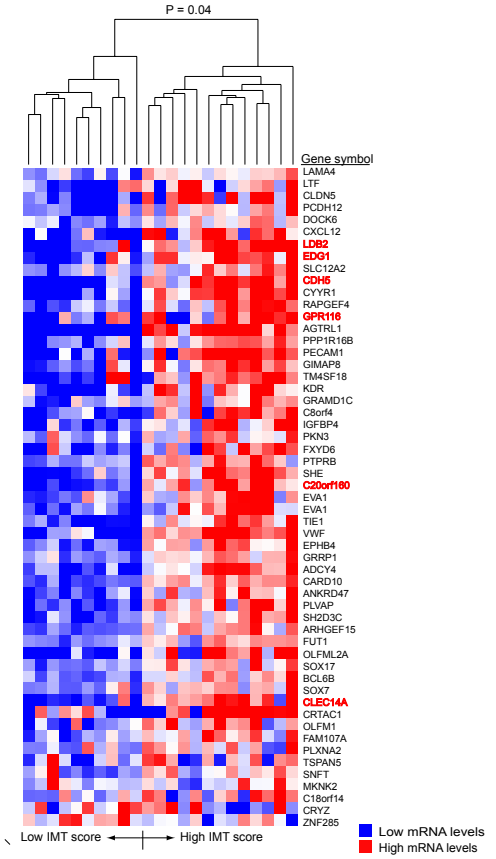

Figure 10: Heat map indicating the mRNA levels of 55 atherosclerosis genes belonging to the one cluster out of 8 that related to IMT ($P = 0.038$) in the validation cohort including 25 carotid plaques. Highlighted in red are genes also identified in both clusters shown in Figure 9.

3.6 A data-integrative approach to uncover cholesterol-responsive networks of atherosclerosis genes (Study IV)

In this study, we identified a subnetwork of cholesterol-responsive genes involved in atherogenesis by prioritizing among genes that were differentially expressed in the atherosclerotic arterial wall in response to plasma cholesterol lowering (i.e., cholesterol-responsive atherosclerosis genes, CRAGs) using in-house gene expression datasets from human and mouse, protein binding [106] and literature mining data.

In the first step, we lowered plasma cholesterol by 80% or more (by pI-pC injection as described in Section 3.2) in the mice at five time points (20, 30, 40, 50, and 60 weeks) $Ldlr^{-/-}$ $Apob^{100/100}$ $Mttp^{flox/flox}$ Mx1-*Cre*; saline-injected littermate mice without cholesterol lowering served as controls. One week after plasma cholesterol lowering, the mice

were sacrificed, and the aortic arch was isolated for subsequent RNA isolation and global gene expression analyses with Affymetrix GeneChip MG 430 2.0. Comparison of the expression profiles in the two groups at each time point with an empirical Bayes statistical test identified 2457 CRAGs ($FDR < 0.1$).

In comparing two sets of gene expression profiles (or any other genome measure), the lowest P value is often used to identify the "most significant" genes—those most relevant from a biological perspective. However, although the subset of differentially expressed genes defined by a given FDR in most instances is relevant, using the individual P-values is most likely not the optimal way to rank and prioritize biologically relevant genes.

In this study, we used another approach, which we call the integrative network approach. We prioritized CRAGs according to their position relative to atherosclerosis seed genes (see definition below) in three global gene networks. To further prioritize these genes, we used the set of 1259 genes found to be active during atherogenesis in Study II.

The first network, the expression network, was inferred using first order partial correlations [82] from the human expression compendium of 40 samples of atherosclerotic aorta in study III. The second network, the protein–protein interaction network, was downloaded from the Human Protein Reference Database (HPRD) [106]. The third network, the literature network, was derived by connecting genes associated with significantly overlapping article sets ($P < 10^{-5}$) in the Entrez Gene database [59].

Atherosclerosis seed genes were identified by searching PubMed for articles associated with the search terms "(atherosclerosis and cholesterol) OR foam cells". This search identified 18,002 articles, which were then linked to genes in the Entrez Gene database. Sixty-eight genes were linked to these 18,002 articles more often than expected by chance ($P < 0.05$). These genes were considered to be atherosclerosis seed genes.

Next, we identified which of the 2457 CRAGs were neighbors (not nececarily first neighbors) to seed genes in at least two of the three global networks. Of 387 CRAGs that met this criterion, 35 were among the 1259 genes active during atherogenesis in Study II (Figure 6). These 35 CRAGs were considered the top-ranked genes based on this integrative network approach. Seven of the top-ranked 35 CRAGs were also atherosclerosis seed genes: *CXCL16, ICAM1, LDLRAP1, PPARA, PPARG, SCARB1*, and *SREBF2*. One of the top-ranked CRAGs was *PECAM1*, which was not an atherosclerosis seed

gene but is of known importance in atherosclerosis progression and endothelial activation. Interestingly, *PECAM1* is included in the module of 129 genes related to the Kegg pathway of transendothelial migration identified in study III (Figures 9A and 10).

To learn more about the 35 top-ranked genes, we constructed a subnetwork based on these genes and their first neighbors, according to their edges in any of the three global networks. This network, shown in Figure 12, contains 947 nodes, of which 26 are atherosclerosis seed genes, 219 are CRAGs (i.e., part of the 2457 genes), and 124 are active during atherogenesis (i.e., part of the 1259 genes). Using DAVID [101] to conduct gene-annotation enrichment analysis, we found that this gene subnetwork is most significantly associated with the disease class cardiovascular disease ($FDR = 0,0002$). The network also included 26 of 47 genes in the notch signaling pathway ($FDR < 10^{-8}$) and 58 of 199 genes in the focal adhesion pathway ($FDR < 10^{-8}$).

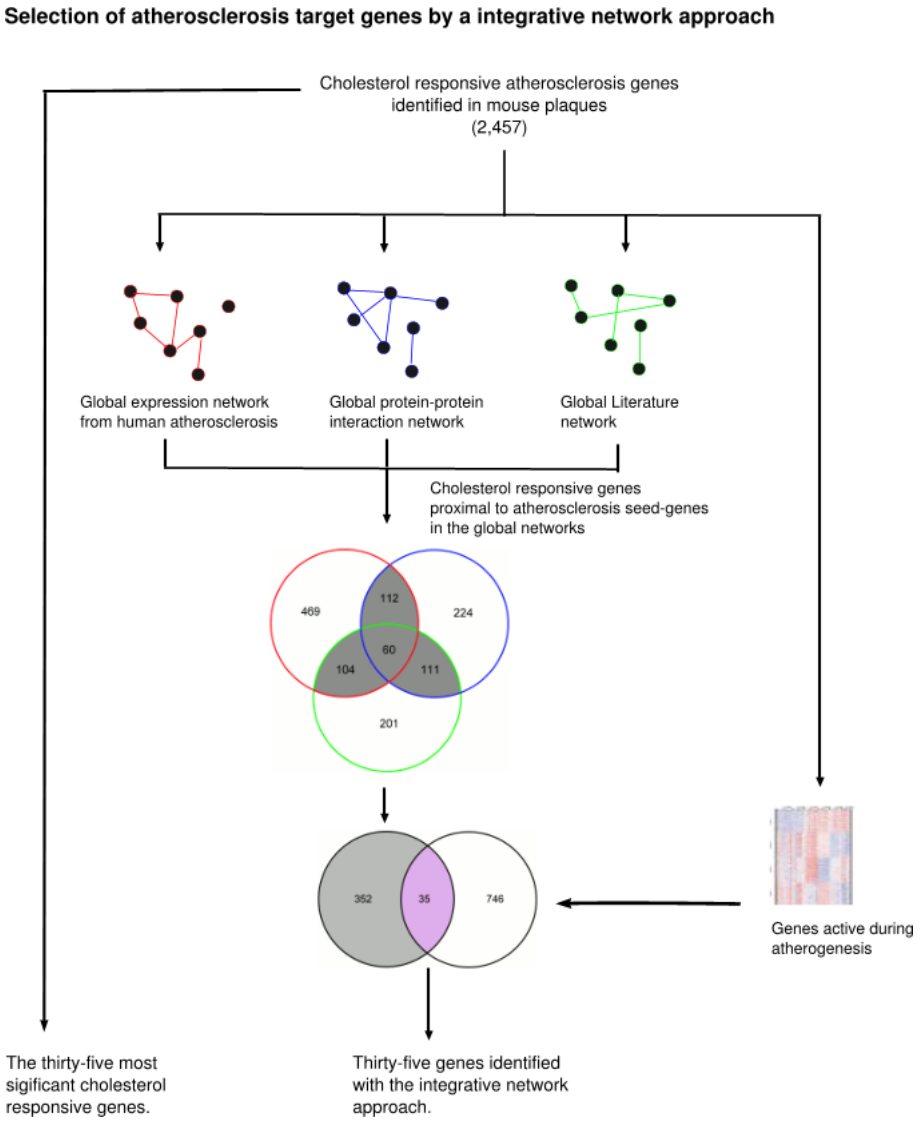

Figure 11: First, 2457 cholesterol-responsive atherosclerosis genes ($FDR < 0.1$) were identified by comparing mRNA levels in atherosclerotic aortas from saline-treated control mice with aortas from mice sacrificed 1 week after induction of a genetic switch ($Mttp^{\Delta/\Delta}$) to lower plasma cholesterol by $> 80\%$. Second, three global networks of genes were generated (see Materials and Methods for details) from a compendium of human atherosclerosis gene expression profiels (study III) (red), protein-protein-binding data (blue), and literature mining(green), respectively. The global networks were then used to prioritize the 2,457 cholesterol-responsive atherosclerosis genes by their location with respect to 68 atherosclerosis seed genes. Three-hundred and eighty-seven genes were identified in at least two of the global networks (gray area in the first Venn diagram). Thirty-five of these genes had previously been identified as active during atherogenesis in study II (second Venn diagram, pink area). These 35 genes are referred to as the top-ranked genes from the integrative network approach (see Paper IV Table 1). A set of 35 of the statistically most significant of 2457 differentially expressed genes was used as a reference (see PaperIV Table 2).

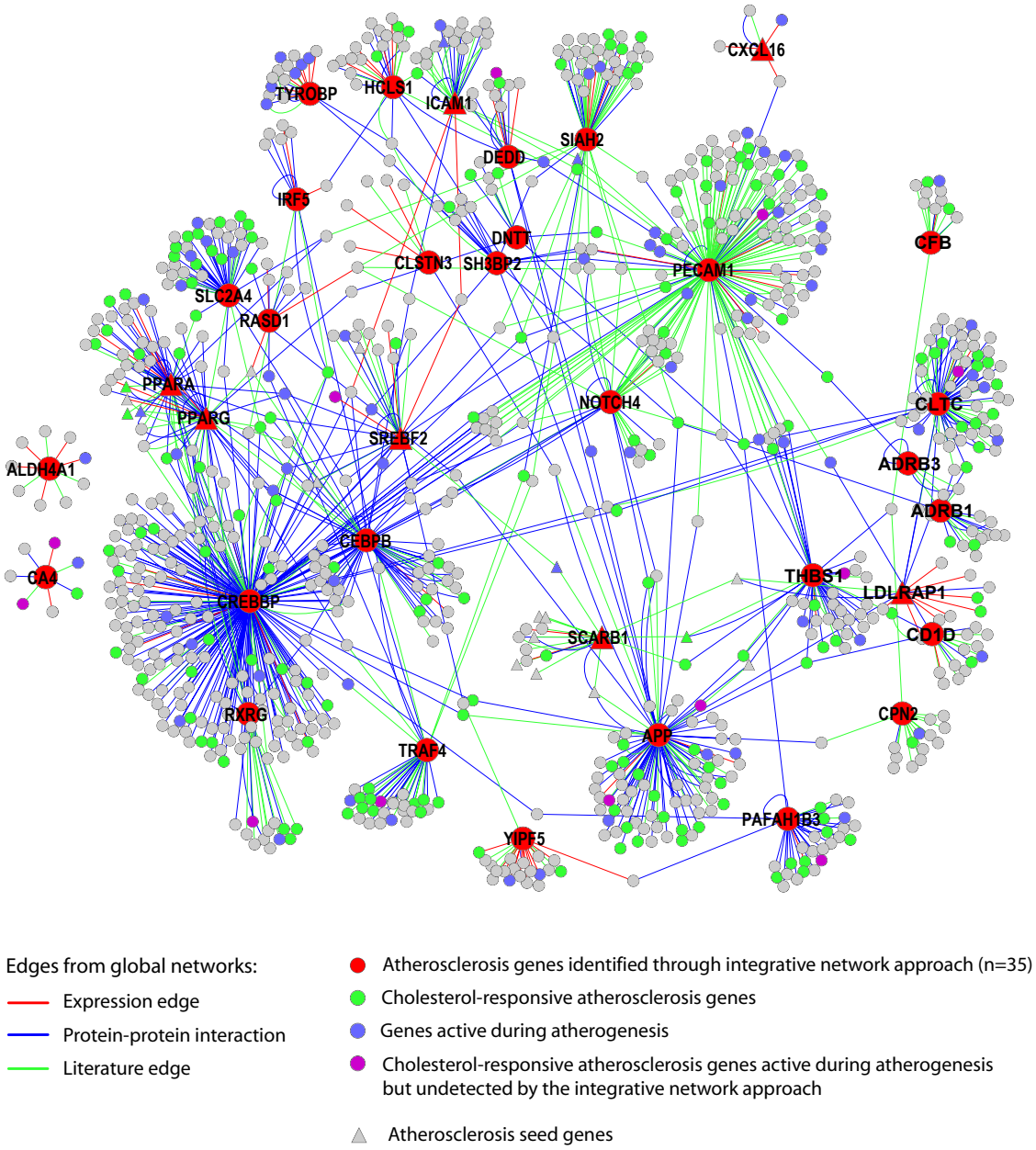

Figure 12: A network of 35 top-ranked genes as identified by the network integrative approach and their first neighbours are shown. This network contained 943 nodes and 1221 edges. Nodes indicated with a triangle are atherosclerosis seed genes. The colour code of the edges indicates which of the three global networks described in Figure 11 the edge were derived from. Two (CA4 and ALDH4A1) of the 35 top-ranked genes were not part of the main connected network because they were found to be more than two edges away from any other top-ranked gene. The remaining 33 top-ranked genes were in the main connected module, and 13 were directly connected with each other.

3.7 Methodological Considerations

3.7.1 Data Storage

Gene expression profiling on a global scale generates vast amounts of data, handling databases is therefore a central and important task. For instance, it is necessary to store the data in a structured way to avoid any risk of "mislabeling". Several laboratory information management systems are available for this purpose. However, we chose to design a custom database schema for storing the in-house generated gene expression data in a MySQL database (http://www.mysql.com) using the InnoDB storage engine (http://www.innodb.com/). In this database, we have also integrated public data sources, for example RefSeq mRNA transcript sequences [58] and the Entrez Gene [59] and Gene Ontology [107] databases.

3.7.2 Probe set Definition of Affymetrix Genechips

Affymetrix probe sets are defined from UniGene [108] transcripts and most commonly contain 11 probe pairs. Inspired by [56,57], we decided to define custom probe sets based on matching the probe sequences to high-quality RefSeq transcripts in Study III. After removing all cross-hybridizing probes, we recovered 14,699[1] probe sets RefSeq probe sets on the HG U133 Plus 2.0. Reducing the number of probe sets from 54,675 to 14,699[1], may seem drastic but we are still including measurements from 33%[1] of perfect-match probes after removal of unreliable and cross-hybridizing probe sets.

Sometimes, multiple RefSeq transcripts corresponding to the same gene have a large sequence similarity; therefore, many or all probes match both probe sets. In Study IV, we changed the probe set definition to accommodate this. Here we define a probe set for each gene based on all probes matching any RefSeq transcript of that gene, resulting in 17,014 probe sets utilizing 39% of the perfect-match probes on the HG U133 Plus 2.0.

[1]This figure is recalculated using a newer version of RefSeq as compared to Study III

3.7.3 Low level processing of Affymetrix Genechip data

In Study II, we used the standard protocol in MAS version 5.0 [60], which includes global scaling and probe set summarization. We then averaged probe set signals corresponding to the same gene to give a gene signal. Before identifying differentially expressed genes between two states, we normalized the samples with Loess [109] to remove intensity bias.

Studies like [18, 19] show that their methods outperform MAS 5.0 by reducing noise and improving specificity and sensitivity in detecting differential expression. For this reason, we changed our preprocessing strategy by applying quantile normalization [18] and summarizing the normalized probe signals with robust multiarray analysis [19] in the later Study III and IV.

3.7.4 Identification of differentially expressed genes

As discussed in Section 1.5.1, normal P-values are not appropriate in a multiple testing setting. To identify differentially expressed genes, we estimated the FDR with an empirical Bayes method developed by [97]. We used this approach in all studies except for Study I, where we performed differential testing without adjustments for multiple testing as a part of the Hubdetector method. In Study IV, we tested several clusters for partitions relevant to atherosclerosis measurement using Benjamini–Hochberg FDR correction [66]

Another important analytical issue is that genes with low variance sometimes show strong statistical significance, which, in most instances, is rather meaningless because the differences in mRNA levels are too small. In Study IV, we acknowledged this and used a t-statistic modified by adding a constant "fudge factor" s_0 to the denominator [67, 68]. The fudge factor we used was the 90th percentile of gene-specific standard deviation distribution as suggested by [68].

3.7.5 Clustering

We used clustering algorithms in study II and study III. In study II, we identified genes responsive to a change between the time points before we clustered the genes, thereby avoiding the problem of including a large set of uninformative and "noisy" genes into

the clustering algorithm (see Kerr et al. [73]). The accual clustering was performed with a k-medoid clustering algorithm [98, 99], which is similar to k-means clustering faster. In Study III, we instead used a more unbiased method, in which genes and samples are clustered with a two-way approach [79,80] without first identifying differentially expressed genes. The original two-way approach clusters the samples and genes iteratively; however, we used a "light version" that includes only one iteration. In a larger cohort it could be interesting to continue the iteration further.

3.7.6 Literature mining

The massive amount of published research makes it extremely difficult go through articles manually to identify gene functions and relationships among genes. This problem has prompted a new field of research—automated literature and text mining [110–113].

We used automated literature mining in both study II and IV. However, the techniques differed. In study II, we used a text mining algorithm to search for gene names and symbols in the article abstracts [100]. In study IV, we used the article-to-gene links in the Entrez Gene database [59].

3.7.7 Functional analysis of gene-sets

In all of the studies, we needed to annotate the gene-sets resulting from our analysis. For this purpose, we commonly used gene-annotation enrichment analysis, in most cases with the DAVID tool [101]. Gene-annotation enrichment analysis is performed by computing the probability of drawing the observed number of genes with a specific annotation (e.g., a GO category or a KEGG pathway) from a set of background genes. A hypergeometric distribution is used to make this computation. One problem with this approach is again multiple testing, here further complicated by relatedness of functional cetegories.

4 Discussion

In this section, I decided to focus on cell-type heterogeneity, regulation of gene activity and network–expression integration which I believe are three critical issues in the thesis.

4.1 Measuring expression in heterogeneous samples

In Studies II, III and IV, we analyzing gene expression by measuring RNA levels in tissue samples from mice and human patients. To some degree, all such samples contain multiple cell types. Thus, it is impossible to attribute changes in mRNA expression levels to any particular cell type on the basis of expression data alone. In fact, the expression profiles reflect not only gene activity within the cells but also the cellular composition of the tissues. In such cases, interpretation of the gene expression data is more problematic than in studies of homogeneous cells (e.g., cultured cells). However, culturing the atherosclerotic cells of interest instead causes another problem—the cultured cells have been removed from their natural environment, which alters their transcriptional patterns and reduces disease relevance.

To measure cell-type-specific gene expression from a heterogeneous biopsy, one could use laser microdissection techniques [114] to collect specific cells for further analysis (e.g., measuring RNA levels). However, for three reasons, we elected not to use this interesting technology. First, although several cell types are important in atherogenesis, expression profiling of whole lesions is still useful for detecting meaningful biological processes. For instance, with our approach, we captured cellular interplay, as reflected in the leukocyte transendothelial migration module we identified that involves genes from both leukocytes and endothelial cells (see Paper III, Figure 3B). Second, at least 500 cells are needed to isolate enough RNA for microarray expression profiling—a labor-intensive task if cells are to be isolated one by one using laser microdissection [115]. Also, one may question the usefulness of this technique since within one atherosclerosis cell-type there are many subtypes. For instance from histological examination, it is clear that cell-type like smooth muscle cells come in many shapes and sizes, and those differences are most likely are reflected in their transcriptional repertoire.

In Study III, we measured global gene expression in the aortic root, which con-

tains both normal tissue and diseased tissue. Thus, the expression profiles also reflect nonatherosclerotic vascular expression. To remove this vascular expression, we used the internal mammary artery from the same patient as a control, as this vessel exhibits little or no atherosclerosis [116].

In studying atherogenesis in the $Ldlr^{-/-}$ $Apob^{100/100}$ $Mttp^{flox/flox}$ Mx1-*Cre* mouse model in Study II, we expected that the cell composition would change as atherosclerosis progressed between time points, which was also confirmed by the histological investigation. However, by measuring the mRNA levels of cell-type specific markers, were able to predict the accumulation of macrophages before the rapid expansion of plaque area. Moreover, in Study II and IV, we studied how plasma cholesterol lowering affects transcription, aiming to identify cholesterol-responsive genes. In these experiments, we wanted to avoid identifying gene expression changes due to differences in cellular composition between the mice with low plasma cholesterol and the control mice. Therefore, in an additional set of experiments, we looked for changes in cellular composition for 2 weeks after cholesterol levels were lowered. No changes in lesion size or cellular marker concentrations were observed (see Paper II Figure 3).

4.2 Regulation of gene activity

In studies of gene expression, it is of key importance that mRNA levels reflect actual functional gene activity. Several regulatory mechanisms have been proposed, for example transcription, translation, chromatin remodeling, nuclear territories, and regulatory roles for noncoding RNA. Although the mRNA level is clearly an important predictor of gene activity and function, there have been some debate about the extent to which the regulatory processes is captured by gene expression measurements. In a study of *E. Coli* and *S. Cerevisiae* gene expression, the correlation between transcriptional regulators and their target genes was in many cases insignificant [117], possibly because the activity of many transcription factors is regulated at a post-transcriptional level. Or transcription factor concentration—both on mRNA and protein level—is stable over time, with transcription activity being modulated through the binding of small molecules and proteins. We argued in Study I for an extended interpretation of gene regulatory networks that includes

transcriptional co-factor proteins. Since co-factors can obviously function as active modulators of transcription factors, it would be interesting to investigate whether co-factor proteins are co-regulated with their target genes.

The low correlation between transcription factor and target gene has another implication—that reverse engineering schemes may not be able uncover the entire physical transcription network without the use of perturbation techniques. However, it is also important to recognize other reasons for low correlation, including low expression of transcription factors (under detection level) [118] and nonlinear regulation effects (e.g., time delay between expression of transcription factors and effects on their target genes' expression [119].

In contrast, target genes regulated by the same transcription factor or the same set of transcription factors often exhibit a correlated gene expression pattern [117,119], suggesting that gene expression levels predict functional activity. Thus, clustering techniques are likely to identify target genes in a co-regulated module, while transcriptional regulators may be missed. These possibilities are consistent with our observations in Study III, in which cluster analysis identified the transcriptional co-factor LDB2 but not its partner transcription factor.

4.3 Integration of expression profiling and global network structures

In Study IV, we integrated data from global networks with microarray measurements in atherosclerosis-prone mice during atherosclerosis development and after cholesterol lowering. Rather than use P-values to prioritize differentially expressed cholesterol-responsive genes based on their statistical significance, we utilized a list of atherosclerosis seed genes and prioritized differentially expressed genes based on whether they are neighbors to seed genes in three global networks. This approach resulted in a gene set that was more relevant to our current knowledge of atherosclerosis because it had more homogeneous functions, as judged from Gene Ontology and pathway analyses, than the differentially expressed cholesterol-responsive genes that were most statistically significant. This approach has two significant strengths: it expands from what is currently known about the atheroscle-

rosis process, and it results in a set of genes that can be put into a more relevant context (i.e., into one or several subnetworks of interactions).

However, these two strengths are related to two weaknesses of the approach. First, the approach relies on a list of seed genes of known important in atherosclerosis. Thus, priority is given to genes associated with pathways and processes in atherosclerosis that are already known rather than to those previously unidentified. However, this is appropriate, since one major task of network approaches is identify new genes playing key roles in known atherosclerosis processes and thus completing the parts lists. Second, the approach relies on global networks—in the current study, protein-protein binding, literature mining, and atherosclerosis expression networks. Genes not present in these networks will obviously not be identified, which leads to bias toward genes that have, in general, been more well studied (not limited to known atherosclerosis genes). This bias is most pronounced in the literature network and to some extent in the protein-protein binding network.

The integrative network approach (Study IV) raises other issues we intend to address—for example, the extent to which the three global networks contribute to the list of prioritized genes and how this list would be altered by excluding one or several of the global networks from the integration process. To address this question, we could integrate combinations of these three global networks with comparisons of plasma cholesterol-responsive genes (n=2457) and with the genes identified as active in atherosclerosis (n =1257, Study II) to generate alternative gene lists.

5 Future Perspectives

5.1 Integrating gene and protein expression measurements

In the last decade, the increasing use of global gene expression profiling (i.e., global mRNA levels) has yielded several interesting results and insights into biological processes and human diseases (for example [95, 120–122]). However, the most important role of an RNA molecule is to serve as a messenger for protein synthesis. Thus, one intriguing idea is to increase network resolution generated from gene expression studies by adding information from global measurements of protein expression.

However, protein expression measurements are not nearly as well developed as global mRNA measurements. Multidimensional liquid chromatography combined with tandem mass spectrometry can be used to measure the concentrations of up to 1000 proteins in a single sample [123]. With further development, this technology might accommodate global measurement of protein expression [15]. Another technology, protein micoarrays, can be used to measure protein expression in a parallel fashion. However, for at least two reasons, this technology has not matured as fast as DNA microarray technology. First, DNA microarrays exploit the ability of single-stranded nucleic acid sequences to hybridize onto a complementary sequence, but it is much more difficult to identify compounds that bind to a specific protein from the sequence alone. Second, RNA samples can readily be amplified by polymerase chain reaction (PCR), but no easy protocol for protein amplification exists, and thus the measurement technology will have to be much more sensitive.

One common type of protein microarray is the antibody array, in which a slide is printed with protein-specific antibodies. Moreover, there are attempts to create antibodies for all human proteins. A recent study, for example, described a library containing 5067 "gene-centric" antibodies covering ∼25% of the protein-coding human genes [124].

Although several studies have measured concentrations of multiple proteins on a chip [125] or with tandem mass spectrometry [126], it is not currently possible to measure protein expression on a truly global scale. However, if technical advances were to make such measurements possible, it would definitely be highly interesting to combine global protein expression measurement with gene expression measurements to provide a more

detailed view of gene regulation.

Protein expression measurements could also provide new opportunities for the reverse engineering community and enable researchers to moving beyond the effective gene network (see Section 4.2). It would be very interesting to determine whether transcription factor protein levels correlated with the mRNA expression of their target genes (see also Section 4.2). Moreover, it would be interesting to cluster protein expression data and compare it to cluster results from mRNA expression data. It is likely these data sources will reveal different aspects of the investigated processes, as a consequence extending Study III in this thesis with protein expression measurements may enable a more complete identification of gene modules important to atherosclerosis severity.

Last, genome-wide expression studies like in this thesis and by others generate an increasingly robust list of atherosclerosis candidate genes. In a few years from now, we may have end up with a list in the thousands of relatively well-established atherosclerosis genes. Clearly, developing custom made protein analysis platforms focusing on these genes will by-pass some of the problems inherent with whole-genome proteomic approaches.

5.2 Integrating gene expression with genotyping

Single nucleotide polymorphisms (SNPs) are mutations in which one specific DNA base is substituted in the genome of at least 1% of the human population. SNPs are important in human diseases [4] and can be interrogated on a large scale by using genotyping arrays[1], which currently allow the detection of up to 1,000,000 SNPs from the same sample in parallel. Copy number variation (CNV)—insertions, deletions, and multiplications of DNA segments—are another potentially important source of genetic variability with implications for human disease [127]. For example, having multiple copies of the *CCL3L1* gene reduces susceptibility to HIV infection [128]. Luckily, copy number variation can also be identified using commercially available SNP arrays [129].

Genetic variants, like SNPs and CNVs described above, will cause differences in regulatory properties and/or changes in actual gene sequence that are reflected in gene expression profiles and physical protein properties. In some cases, a mutation in one gene

[1]The two major array manufacturers are Illumina and Affymetrix

will cause changes leading to a higher-order phenotype (e.g., sickle-cell anemia and cystic fibrosis), while in other cases, the interplay of several genetic changes leads to a higher-order phenotype. Traditional approaches—focusing on mapping genotypes to higher-order phenotypes—have had trouble unraveling complex phenotypes such as atherosclerosis. Gene expression may serve as an intermediate step between genotype and complex phenotype. In early studies by Brem et al. [130] and Schadt et al. [131], gene expression patterns were shown to be highly hereditable; moreover, a large number of genetic loci affecting gene expression—referred to as expression quantitative loci or eQTL—were identified in yeast mouse, maize, and human. Schadt and coworkers also used eQTLs and gene expression to link five genomic regions that were important in defining the fat-pad-mass trait in these mice, which would not have been possible using traditional techniques [131].

In more recent studies, this approach has been applied to a range of settings to identify potential susceptibility genes for several complex traits, including obesity, diabetes, atherosclerosis, and neuronal function, in mice and in human subjects [132–136].

In the light of these results, it would be interesting to genotype patients in the STAGE study (see section 3.1.1) using a global SNP array. The benefit of the STAGE cohort is that we have multiple expression profiles for the same gene in up to five tissues, which would enable us to identify similarities and differences in the genetic architecture in those tissues. The combined expression genotype data would, for instance, give us the opportunity to find genomic regions associated with the module shown to be related to atherosclerosis severity in Study III. In a small-scale study involving a handful selected SNPs, we could, using statistical and bioinformatic methods, show evidence that one of these SNPs is responsible for regulation in this module. This SNP have also been further validated and found to cause myocardial infarction or atherosclerosis in the Swedish population of three independent cohorts [137–139].

6 Concluding Remarks

This thesis provides evidence that analysis of global gene expression profiles isolated from a wide range of biological specimens can be used to infer functional interactions of genes in modules or networks. The content and structure of these modules and networks can be used to improve our understanding how complex disorders like atherosclerosis develop.

It is hard to predict the most efficient path to a more complete understanding of complex diseases. I believe in depth investigation of candidate genes will be important in the future but only as a complement to global approaches. Many things will be learnt from combing different genomic strategies bringing their different strengths and weaknesses to the the same table. By doing this we can get a course grained picture of the disease process at different levels, giving us the opportunity to find new disease relevant relationships.

Acknowledgements

This research has been performed at the Computational Medicine group, Department of Medicine at Karolinska Institutet. There are several people who have contributed directly or indirectly to this thesis. In particular I wish to thank:

My supervisors Johan Björkegren and Jesper Tegnér for introducing me to the computational medicine and atherosclerosis research field and also for supporting me and not loosing faith in me when I choose alterative paths. Johan for being creative, intelligent and having a positive and easy going attitude to science and life in general. Jesper for beeing open minded and willing to discuss all sorts of ideas about science, philosophy, and totally unrelated matters like the stock market.

My supervisor Josefin, for beeing enthusiastic, knowledgable, and eager to explain. And also for friendly talks and advice in scientific and non-scientific matters.

All members of Computational Medicine group, Sara, Peri, Shoreh, Olivia, and also all previous group members for good collaboration, nice lunch and coffey breaks. This thesis would not have been completed without you.

All members of the atherosclerosis research unit for providing a good scientific environment.

Mika and Michael, for intersting discussions about reverse engineering schemes and cell regulation and for fruitful collaboration.

My mum and dad for always being supportive without interfering with my life.

My two children, Hampus and Molly, for being the best kids and for distracting my attention away from thesis writing and to more important things.

I also want to thank Maria for *everything*. You are my true love.

This research has been supported by the Swedish Knowledge Foundation through the Industrial PhD programme in Medical Bioinformatics at the Strategy and Development Office at Karolinska Institutet. The thesis has been proof read and edited by Stephen Ordaway.

References

[1] Riordan, J. R., Rommens, J. M., Kerem, B., Alon, N., Rozmahel, R., Grzelczak, Z., Zielenski, J., Lok, S., Plavsic, N., and Chou, J. L. (1989). Identification of the cystic fibrosis gene: cloning and characterization of complementary DNA. Science *245*, 1066–73.

[2] Rommens, J. M., Zengerling, S., Burns, J., Melmer, G., Kerem, B. S., Plavsic, N., Zsiga, M., Kennedy, D., Markiewicz, D., and Rozmahel, R. (1988). Identification and regional localization of DNA markers on chromosome 7 for the cloning of the cystic fibrosis gene. Am J Hum Genet *43*, 645–63.

[3] McPherson, R., Pertsemlidis, A., Kavaslar, N., Stewart, A., Roberts, R., Cox, D. R., Hinds, D. A., Pennacchio, L. A., Tybjaerg-Hansen, A., Folsom, A. R., et al. (2007). A common allele on chromosome 9 associated with coronary heart disease. Science *316*, 1488–91.

[4] Helgadottir, A., Thorleifsson, G., Manolescu, A., Gretarsdottir, S., Blondal, T., Jonasdottir, A., Jonasdottir, A., Sigurdsson, A., Baker, A., Palsson, A., et al. (2007). A common variant on chromosome 9p21 affects the risk of myocardial infarction. Science *316*, 1491–3.

[5] Tegner, J., Skogsberg, J., and Björkegren, J. (2007). Multi-organ whole-genome measurements and reverse engineering to uncover gene networks underlying complex traits. Journal of Lipid Research *48*, 267–277.

[6] Ideker, T., Galitski, T., and Hood, L. (2001). A new approach to decoding life: systems biology. Annu Rev Genomics Hum Genet *2*, 343–72.

[7] Kitano, H. (2002). Systems biology: a brief overview. Science *295*, 1662–4.

[8] Ehrenberg, M., Elf, J., Aurell, E., Sandberg, R., and Tegner, J. (2003). Systems biology is taking off. Genome Res *13*, 2377–80.

[9] Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860–921.

[10] Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., et al. (2001). The sequence of the human genome. Science *291*, 1304–51.

[11] Waterston, R. H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J. F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P., et al. (2002). Initial sequencing and comparative analysis of the mouse genome. Nature *420*, 520–62.

[12] Lockhart, D. J., Dong, H., Byrne, M. C., Follettie, M. T., Gallo, M. V., Chee, M. S., Mittmann, M., Wang, C., Kobayashi, M., Horton, H., et al. (1996). Expression monitoring by hybridization to high-density oligonucleotide arrays. Nat Biotechnol *14*, 1675–80.

[13] Schena, M., Shalon, D., Davis, R. W., and Brown, P. O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science *270*, 467–70.

[14] Ginsburg, G. S., Donahue, M. P., and Newby, L. K. (2005). Prospects for personalized cardiovascular medicine: the impact of genomics. J Am Coll Cardiol *46*, 1615–27.

[15] MacBeath, G. (2002). Protein microarrays and proteomics. Nat Genet *32 Suppl*, 526–32.

[16] Lusis, A. J., Mar, R., and Pajukanta, P. (2004). Genetics of atherosclerosis. Annu Rev Genomics Hum Genet. *5*, 189–218.

[17] Mecham, B. H., Klus, G. T., Strovel, J., Augustus, M., Byrne, D., Bozso, P., Wetmore, D. Z., Mariani, T. J., Kohane, I. S., and Szallasi, Z. (2004). Sequence-matched probes produce increased cross-platform consistency and more reproducible biological results in microarray-based gene expression measurements. Nucleic Acids Res *32*, e74.

[18] Bolstad, B. M., Irizarry, R. A., Astrand, M., and Speed, T. P. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics *19*, 185–93.

[19] Irizarry, R. A., Bolstad, B. M., Collin, F., Cope, L. M., Hobbs, B., and Speed, T. P. (2003). Summaries of Affymetrix GeneChip probe level data. Nucleic Acids Res *31*, e15.

[20] World Health Organization http://www.who.int/mediacentre/factsheets/fs317/en/. (2007). Fact sheet N^{o}317: Cardiovascular diseases.

[21] Ghazalpour, A., Doss, S., Yang, X., Aten, J., Toomey, E. M., Nas, A. V., Wang, S., Drake, T. A., and Lusis, A. J. (2004). Thematic review series: The pathogenesis of atherosclerosis. Toward a biological network for atherosclerosis. J Lipid Res *45*, 1793–805.

[22] Cohn, J. S., Wat, E., Kamili, A., and Tandy, S. (2008). Dietary phospholipids, hepatic lipid metabolism and cardiovascular disease. Curr Opin Lipidol *19*, 257–62.

[23] Balkau, B., Hu, G., Qiao, Q., Tuomilehto, J., Borch-Johnsen, K., and Pyorala, K. (2004). Prediction of the risk of cardiovascular mortality using a score that includes glucose as a risk factor. The DECODE Study. Diabetologia *47*, 2118–28.

[24] Nigro, J., Osman, N., Dart, A. M., and Little, P. J. (2006). Insulin resistance and atherosclerosis. Endocr Rev *27*, 242–59.

[25] Yokoyama, S. (2000). Release of cellular cholesterol: molecular mechanism for cholesterol homeostasis in cells and in the body. Biochim Biophys Acta. *1529*, 231–44.

[26] Tsujita, M., Wu, C.-A., Abe-Dohmae, S., Usui, S., Okazaki, M., and Yokoyama, S. (2005). On the hepatic mechanism of HDL assembly by the ABCA1/apoA-I pathway. J Lipid Res *46*, 154–62.

[27] Maxfield, F. R. and Tabas, I. (2005). Role of cholesterol and lipid organization in disease. Nature *438*, 612–21.

[28] Libby, P. (2002). Inflammation in atherosclerosis. Nature *420*, 868–74.

[29] Hansson, G. K. (2005). Inflammation, atherosclerosis, and coronary artery disease. N Engl J Med *352*, 1685–95.

[30] Packard, R. R. S. and Libby, P. (2008). Inflammation in atherosclerosis: from vascular biology to biomarker discovery and risk prediction. Clin Chem *54*, 24–38.

[31] International Human Genome Sequencing Consortium. (2004). Finishing the euchromatic sequence of the human genome. Nature *431*, 931–45.

[32] Goff, S. A., Ricke, D., Lan, T.-H., Presting, G., Wang, R., Dunn, M., Glazebrook, J., Sessions, A., Oeller, P., Varma, H., et al. (2002). A draft sequence of the rice genome (Oryza sativa L. ssp. japonica). Science *296*, 92–100.

[33] Yu, J., Hu, S., Wang, J., Wong, G. K.-S., Li, S., Liu, B., Deng, Y., Dai, L., Zhou, Y., Zhang, X., et al. (2002). A draft sequence of the rice genome (Oryza sativa L. ssp. indica). Science *296*, 79–92.

[34] Adams, M. D., Celniker, S. E., Holt, R. A., Evans, C. A., Gocayne, J. D., Amanatides, P. G., Scherer, S. E., Li, P. W., Hoskins, R. A., Galle, R. F., et al. (2000). The genome sequence of Drosophila melanogaster. Science *287*, 2185–95.

[35] Goffeau, A., Barrell, B. G., Bussey, H., Davis, R. W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J. D., Jacq, C., Johnston, M., et al. (1996). Life with 6000 genes. Science *274*, 546, 563–7.

[36] Claverie, J. M. (2001). Gene number. What if there are only 30,000 human genes? Science *291*, 1255–7.

[37] Tegnér, J. and Björkegren, J. (2007). Perturbations to uncover gene networks. Trends Genet *23*, 34–41.

[38] Uetz, P., Giot, L., Cagney, G., Mansfield, T. A., Judson, R. S., Knight, J. R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P., et al. (2000). A comprehensive analysis of protein-protein interactions in saccharomyces cerevisiae. Nature *403*, 623–627.

[39] Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M., and Sakaki, Y. (2001). A comprehensive two-hybrid analysis to explore the yeast protein interactome. Proc Natl Acad Sci U S A *98*, 4569–74.

[40] Lee, T. I., Rinaldi, N. J., Robert, F., Odom, D. T., Bar-Joseph, Z., Gerber, G. K., Hannett, N. M., Harbison, C. T., Thompson, C. M., Simon, I., et al. (2002). Transcriptional regulatory networks in saccharomyces cerevisiae. Science *298*, 799–804.

[41] Horak, C., Luscombe, N., Bertone, J. Q. P., Piccirrillo, S., Gerstein, M., and Snyder, M. (2002). Complex transcriptional circuitry at the g1/s transition in saccharomyces cerevisiae. Genes Dev *16*, 3017–33.

[42] Luscombe, N., Babu, M., Yu, H., Snyder, M., Teichmann, S., and Gerstein, M. (2004). Genomic analysis of regulatory network dynamics reveals large topological changes. Nature *431*, 308–312.

[43] Kanehisa, M., Goto, S., Hattori, M., Aoki-Kinoshita, K. F., Itoh, M., Kawashima, S., Katayama, T., Araki, M., and Hirakawa, M. (2006). From genomics to chemical genomics: new developments in KEGG. Nucleic Acids Res *34*, D354–7.

[44] Duarte, N. C., Becker, S. A., Jamshidi, N., Thiele, I., Mo, M. L., Vo, T. D., Srivas, R., and Palsson, B. O. (2007). Global reconstruction of the human metabolic network based on genomic and bibliomic data. Proc Natl Acad Sci U S A *104*, 1777–82.

[45] MacIsaac, K., Wang, T., Gordon, D., Gifford, D., Stormo, G., and Fraenkel, E. (2006). An improved map of conserved regulatory sites for saccharomyces cerevisiae. BMC Bioinformatics *7*, 113.

[46] Barabàsi, A.-L. and Albert, R. (1999). Emergence of scaling in random networks. Science *286*, 509–12.

[47] Jeong, H., Mason, S. P., Barabasi, A. L., and Oltvai, Z. N. (2001). Lethality and centrality in protein networks. Nature *411*, 41–2.

[48] Barabási, A. and Oltvai, Z. (2004). Network biology: understanding the cell's functional organization. Nat Rev Genet *5*, 101–113.

[49] Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N., and Barabasi, A. L. (2002). Hierarchical organization of modularity in metabolic networks. Science *297*, 1551–5.

[50] Shen-Orr, S. S., Milo, R., Mangan, S., and Alon, U. (2002). Network motifs in the transcriptional regulation network of Escherichia coli. Nat Genet *31*, 64–8.

[51] Bansal, M., Belcastro, V., Ambesi-Impiombato, A., and di Bernardo, D. (2007). How to infer gene networks from expression profiles. Mol Syst Biol *3*, 78.

[52] Kuhn, K., Baker, S. C., Chudin, E., Lieu, M.-H., Oeser, S., Bennett, H., Rigault, P., Barker, D., McDaniel, T. K., and Chee, M. S. (2004). A novel, high-performance random array platform for quantitative gene expression profiling. Genome Res *14*, 2347–56.

[53] Affymetrix. (2004). GeneChip® Expression Analysis Technical Manual.

[54] Naef, F., Hacker, C. R., Patil, N., and Magnasco, M. (2002). Empirical characterization of the expression ratio noise structure in high-density oligonucleotide arrays. Genome Biol *3*, RESEARCH0018.

[55] Boguski, M. S. and Schuler, G. D. (1995). ESTablishing a human transcript map. Nat Genet *10*, 369–71.

[56] Mecham, B. H., Wetmore, D. Z., Szallasi, Z., Sadovsky, Y., Kohane, I., and Mariani, T. J. (2004). Increased measurement accuracy for sequence-verified microarray probes. Physiol Genomics *18*, 308–15.

[57] Gautier, L., Moller, M., Friis-Hansen, L., and Knudsen, S. (2004). Alternative mapping of probes to genes for Affymetrix chips. BMC Bioinformatics *5*, 111.

[58] Pruitt, K. D., Tatusova, T., and Maglott, D. R. (2007). NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res *35*, D61–5.

[59] Maglott, D., Ostell, J., Pruitt, K. D., and Tatusova, T. (2005). Entrez Gene: gene-centered information at NCBI. Nucleic Acids Res *33*, D54–8.

[60] Affymetrix. GeneChip® Expression Analysis Data Analysis Fundamentals.

[61] Li, C. and Wong, W. H. (2001). Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. Proc Natl Acad Sci U S A *98*, 31–6.

[62] Zhang, L., Miles, M. F., and Aldape, K. D. (2003). A model of molecular interactions on short oligonucleotide microarrays. Nat Biotechnol *21*, 818–21.

[63] Wu, Z., Irizarry, R. A., Gentleman, R., Murillo, F. M., and Spencer, F. (2004). A model based background adjustment for oligonucleotide expression arrays. Journal of the American Statistical Association *99*, 909–917.

[64] Affymetrix http://www.affymetrix.com/support/technical/technotes/plier_technote.pdf. (2005). Guide to Probe Logarithmic Intensity Error (PLIER) Estimation.

[65] Wit, E. and McClure, J. (2004). Statistics for Microarrays: Design, Analysis and Inference. (John Wiley & Sons, Ltd).

[66] Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Statist. Soc. pp. 289–300.

[67] Tusher, V. G., Tibshirani, R., and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci U S A *98*, 5116–21.

[68] Efron, B., Tibshirani, R., Storey, J., and Tusher, V. (2001). Empirical bayes analysis of microarray experiment. J Am Stat Assoc *96*, 1151–1160.

[69] Hartwell, L. H., Hopfield, J. J., Leibler, S., and Murray, A. W. (1999). From molecular to modular cell biology. Nature *402*, C47–52.

[70] Ravasz, E. and Barabasi, A.-L. (2003). Hierarchical organization in complex networks. Phys Rev E Stat Nonlin Soft Matter Phys *67*, 026112.

[71] D'haeseleer, P., Liang, S., and Somogyi, R. (2000). Genetic network inference: from co-expression clustering to reverse engineering. Bioinformatics *16*, 707–26.

[72] Asyali, M. H., Colak, D., Demirkaya, O., and Inan, M. S. (2006). Gene expression profile classification: A review. Current Bioinformatics *1*, 55–73.

[73] Kerr, G., Ruskin, H. J., Crane, M., and Doolan, P. (2008). Techniques for clustering gene expression data. Comput Biol Med *38*, 283–93.

[74] Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci U S A *95*, 14863–8.

[75] Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E. S., and Golub, T. R. (1999). Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. Proc Natl Acad Sci U S A *96*, 2907–12.

[76] Tavazoie, S., Hughes, J. D., Campbell, M. J., Cho, R. J., and Church, G. M. (1999). Systematic determination of genetic network architecture. Nat Genet *22*, 281–5.

[77] Segal, E., Shapira, M., Regev, A., Pe'er, D., Botstein, D., Koller, D., and Friedman, N. (2003). Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. Nat Genet *34*, 166–76.

[78] Segal, E., Friedman, N., Kaminski, N., Regev, A., and Koller, D. (2005). From signatures to models: understanding cancer using microarrays. Nat Genet *37 Suppl*, S38–45.

[79] Getz, G., Levine, E., and Domany, E. (2000). Coupled two-way clustering analysis of gene microarray data. Proc Natl Acad Sci U S A *97*, 12079–84.

[80] Getz, G., Gal, H., Kela, I., Notterman, D. A., and Domany, E. (2003). Coupled two-way clustering analysis of breast cancer and colon cancer gene expression data. Bioinformatics *19*, 1079–89.

[81] Margolin, A. A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Favera, R. D., and Califano, A. (2006). ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. BMC Bioinformatics *7 Suppl 1*, S7.

[82] de la Fuente, A., Bing, N., Hoeschele, I., and Mendes, P. (2004). Discovery of meaningful associations in genomic data using partial correlation coefficients. Bioinformatics *20*, 3565–74.

[83] Yeung, M. K., Tegner, J., and Collins, J. J. (2002). Reverse engineering gene networks using singular value decomposition and robust regression. Proc Natl Acad Sci U S A *99*, 6163–6168.

[84] Tegnér, J., Yeung, M. K., Hasty, J., and Collins, J. J. (2003). Reverse engineering gene networks: Integrating genetic perturbations with dynamical modeling. Proc Natl Acad Sci U S A *100*, 5944–5949.

[85] Gustafsson, M., Hornquist, M., and Lombardi, A. (2005). Constructing and analyzing a large-scale gene-to-gene regulatory network–lasso-constrained inference and biological validation. IEEE/ACM Trans Comput Biol Bioinform *2*, 254–61.

[86] Gustafsson, M., Hörnquist, M., Lundström, J., Björkegren, J., and Tegnér, J. (2008). Reverse engineering of gene networks with lasso and non-linear basis functions. Manuscript.

[87] Friedman, N., Linial, M., Nachman, I., and Pe'er, D. (2000). Using Bayesian networks to analyze expression data. J Comput Biol *7*, 601–20.

[88] Pena, J. M., Bjorkegren, J., and Tegner, J. (2005). Growing Bayesian network models of gene networks from seed genes. Bioinformatics *21 Suppl 2*, ii224–9.

[89] Butte, A. J. and Kohane, I. S. (2000). Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. Pac Symp Biocomput pp. 418–29.

[90] de Lichtenberg, U., Jensen, L. J., Brunak, S., and Bork, P. (2005). Dynamic complex formation during the yeast cell cycle. Science *307*, 724–7.

[91] Pujana, M. A., Han, J.-D. J., Starita, L. M., Stevens, K. N., Tewari, M., Ahn, J. S., Rennert, G., Moreno, V., Kirchhoff, T., Gold, B., et al. (2007). Network modeling links breast cancer susceptibility and centrosome dysfunction. Nat Genet *39*, 1338–49.

[92] Lieu, H. D., Withycombe, S. K., Walker, Q., Rong, J. X., Walzem, R. L., Wong, J. S., Hamilton, R. L., Fisher, E. A., and Young, S. G. (2003). Eliminating atherogenesis in mice by switching off hepatic lipoprotein secretion. Circulation *107*, 1315–21.

[93] Xenarios, I., Rice, D., Salwinski, L., Baron, M., and Marcotte, E. (2000). Dip: the database of interacting proteins. Nucleic Acids Res. *28*, 289–91.

[94] Deane, C., Salwiński, L., Xenarios, I., and Eisenberg, D. (2002). Protein interactions: two methods for assessment of the reliability of high throughput observations. Mol Cell Proteomics *1*, 349–56.

[95] Hughes, T. R., Marton, M. J., Jones, A. R., Roberts, C. J., Stoughton, R., Armour, C. D., Bennett, H. A., Coffey, E., Dai, H., He, Y. D., et al. (2000). Functional discovery via a compendium of expression profiles. Cell *102*, 109–126.

[96] Mnaimneh, S., Davierwala, A., Haynes, J., Moffat, J., Peng, W., Zhang, W., Yang, X., Pootoolal, J., Chua, G., Lopez, A., et al. (2004). Exploration of essential gene functions via titratable promoter alleles. Cell *118*, 31–44.

[97] Efron, B. and Tibshirani, R. (2002). Empirical bayes methods and false discovery rates for microarrays. Genet Epidemiol *23*, 70–86.

[98] Wolfram Research, I. (2003). Mathematica Edition: Version 5.1. (Champaign, Illinois: Wolfram Research, Inc.).

[99] Kaufman, L. and Rousseeuw, P. (1990). Finding Groups in Data: An Introduction to Cluster Analysis. (New York: John Wiley & Sons).

[100] Strandberg, P. E. (2005). On text mining to identify gene networks with a special reference to cardiovascular disease. Master's thesis Linköping University.

[101] Dennis, G. J., Sherman, B. T., Hosack, D. A., Yang, J., Gao, W., Lane, H. C., and Lempicki, R. A. (2003). DAVID: Database for Annotation, Visualization, and Integrated Discovery. Genome Biol *4*, P3.

[102] Helfand, M., Carson, S., and Kelley, C. (2006). Drug class review on hmg-coa reductase inhibitors (statins). http://www.ohsu.edu/drugeffectiveness/reports/final.cfm.

[103] Blatt, M., Wiseman, S., and Domany, E. (1996). Superparamagnetic clustering of data. Phys. Rev. Lett. *76*, 3251–3254.

[104] Tetko, I. V., Facius, A., Ruepp, A., and Mewes, H.-W. (2005). Super paramagnetic clustering of protein sequences. BMC Bioinformatics *6*, 82.

[105] Matys, V., Kel-Margoulis, O. V., Fricke, E., Liebich, I., Land, S., Barre-Dirrie, A., Reuter, I., Chekmenev, D., Krull, M., Hornischer, K., et al. (2006). TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. Nucleic Acids Res *34*, D108–10.

[106] Mishra, G. R., Suresh, M., Kumaran, K., Kannabiran, N., Suresh, S., Bala, P., Shivakumar, K., Anuradha, N., Reddy, R., Raghavan, T. M., et al. (2006). Human protein reference database–2006 update. Nucleic Acids Res *34*, D411–4.

[107] Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., et al. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet *25*, 25–9.

[108] Schuler, G. D. (1997). Pieces of the puzzle: expressed sequence tags and the catalog of human genes. J Mol Med *75*, 694–8.

[109] Cleveland, W. (1979). Robust locally weighted regression and smoothing scatterplots. J Am Statist Assoc *74*, 829–836.

[110] Zweigenbaum, P., Demner-Fushman, D., Yu, H., and Cohen, K. B. (2007). Frontiers of biomedical text mining: current progress. Brief Bioinform *8*, 358–75.

[111] Zhou, D. and He, Y. (2008). Extracting interactions between proteins from the literature. J Biomed Inform *41*, 393–407.

[112] Jimeno, A., Jimenez-Ruiz, E., Lee, V., Gaudan, S., Berlanga, R., and Rebholz-Schuhmann, D. (2008). Assessment of disease named entity recognition on a corpus of annotated sentences. BMC Bioinformatics *9 Suppl 3*, S3.

[113] Kim, J.-J., Pezik, P., and Rebholz-Schuhmann, D. (2008). MedEvi: retrieving textual evidence of relations between biomedical concepts from Medline. Bioinformatics *24*, 1410–2.

[114] Fink, L., Kwapiszewska, G., Wilhelm, J., and Bohle, R. M. (2006). Laser-microdissection for cell type- and compartment-specific analyses on genomic and proteomic level. Exp Toxicol Pathol *57 Suppl 2*, 25–9.

[115] Fink, L., Kohlhoff, S., Stein, M. M., Hanze, J., Weissmann, N., Rose, F., Akkayagil, E., Manz, D., Grimminger, F., Seeger, W., et al. (2002). cDNA array hybridization after laser-assisted microdissection from nonneoplastic tissue. Am J Pathol *160*, 81–90.

[116] Sims, F. H. (1983). A comparison of coronary and internal mammary arteries and implications of the results in the etiology of arteriosclerosis. Am Heart J *105*, 560–6.

[117] Herrgard, M. J., Covert, M. W., and Palsson, B. O. (2003). Reconciling gene expression data with known genome-scale regulatory network structures. Genome Res *13*, 2423–34.

[118] Kong, Y. M., Macdonald, R. J., Wen, X., Yang, P., Barbera, V. M., and Swift, G. H. (2006). A comprehensive survey of DNA-binding transcription factor gene expression in human fetal and adult organs. Gene Expr Patterns *6*, 678–86.

[119] Yu, H., Luscombe, N. M., Qian, J., and Gerstein, M. (2003). Genomic analysis of gene expression relationships in transcriptional regulatory networks. Trends Genet *19*, 422–7.

[120] Spellman, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., Brown, P. O., Botstein, D., and Futcher, B. (1998). Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. Mol Biol Cell *9*, 3273–97.

[121] Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., et al. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science *286*, 531–7.

[122] Quackenbush, J. (2006). Microarray analysis and tumor classification. N Engl J Med *354*, 2463–72.

[123] Wong, J. W. H., Sullivan, M. J., and Cagney, G. (2008). Computational methods for the comparative quantification of proteins in label-free LCn-MS experiments. Brief Bioinform *9*, 156–65.

[124] Berglund, L., Björling, E., Oksvold, P., Fagerberg, L., Asplund, A., Szigyarto, C. A.-K., Persson, A., Ottosson, J., Wernerus, H., Nilsson, P., et al. (2008). A gene-centric human protein atlas for expression profiles based on antibodies. Mol Cell Proteomics *7*, 2019–2027.

[125] Haab, B. B., Dunham, M. J., and Brown, P. O. (2001). Protein microarrays for highly parallel detection and quantitation of specific proteins and antibodies in complex solutions. Genome Biol *2*, RESEARCH0004.

[126] Guo, Y., Xiao, P., Lei, S., Deng, F., Xiao, G. G., Liu, Y., Chen, X., Li, L., Wu, S., Chen, Y., et al. (2008). How is mRNA expression predictive for protein expression? A correlation study on human circulating monocytes. Acta Biochim Biophys Sin (Shanghai) *40*, 426–36.

[127] Redon, R., Ishikawa, S., Fitch, K. R., Feuk, L., Perry, G. H., Andrews, T. D., Fiegler, H., Shapero, M. H., Carson, A. R., Chen, W., et al. (2006). Global variation in copy number in the human genome. Nature *444*, 444–54.

[128] Gonzalez, E., Kulkarni, H., Bolivar, H., Mangano, A., Sanchez, R., Catano, G., Nibbs, R. J., Freedman, B. I., Quinones, M. P., Bamshad, M. J., et al. (2005). The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility. Science *307*, 1434–40.

[129] Komura, D., Shen, F., Ishikawa, S., Fitch, K. R., Chen, W., Zhang, J., Liu, G., Ihara, S., Nakamura, H., Hurles, M. E., et al. (2006). Genome-wide detection of human copy number variations using high-density DNA oligonucleotide arrays. Genome Res *16*, 1575–84.

[130] Brem, R. B., Yvert, G., Clinton, R., and Kruglyak, L. (2002). Genetic dissection of transcriptional regulation in budding yeast. Science *296*, 752–5.

[131] Schadt, E. E., Monks, S. A., Drake, T. A., Lusis, A. J., Che, N., Colinayo, V., Ruff, T. G., Milligan, S. B., Lamb, J. R., Cavet, G., et al. (2003). Genetics of gene expression surveyed in maize, mouse and man. Nature *422*, 297–302.

[132] Emilsson, V., Thorleifsson, G., Zhang, B., Leonardson, A. S., Zink, F., Zhu, J., Carlson, S., Helgason, A., Walters, G. B., Gunnarsdottir, S., et al. (2008). Genetics of gene expression and its effect on disease. Nature *452*, 423–8.

[133] Chen, Y., Zhu, J., Lum, P. Y., Yang, X., Pinto, S., MacNeil, D. J., Zhang, C., Lamb, J., Edwards, S., Sieberts, S. K., et al. (2008). Variations in DNA elucidate molecular networks that cause disease. Nature *452*, 429–35.

[134] Chesler, E. J., Lu, L., Shou, S., Qu, Y., Gu, J., Wang, J., Hsu, H. C., Mountz, J. D., Baldwin, N. E., Langston, M. A., et al. (2005). Complex trait analysis of gene expression uncovers polygenic and pleiotropic networks that modulate nervous system function. Nat Genet *37*, 233–42.

[135] Hubner, N., Wallace, C. A., Zimdahl, H., Petretto, E., Schulz, H., Maciver, F., Mueller, M., Hummel, O., Monti, J., Zidek, V., et al. (2005). Integrated transcriptional profiling and linkage analysis for identification of genes underlying disease. Nat Genet *37*, 243–53.

[136] Schadt, E. E., Molony, C., Chudin, E., Hao, K., Yang, X., Lum, P. Y., Kasarskis, A., Zhang, B., Wang, S., Suver, C., et al. (2008). Mapping the genetic architecture of gene expression in human liver. PLoS Biol *6*, e107.

[137] Leander, K., Hallqvist, J., Reuterwall, C., Ahlbom, A., and de Faire, U. (2001). Family history of coronary heart disease, a strong risk factor for myocardial infarction interacting with other cardiovascular risk factors: results from the Stockholm Heart Epidemiology Program (SHEEP). Epidemiology *12*, 215–21.

[138] Samnegard, A., Silveira, A., Lundman, P., Boquist, S., Odeberg, J., Hulthe, J., McPheat, W., Tornvall, P., Bergstrand, L., Ericsson, C.-G., et al. (2005). Serum matrix metalloproteinase-3 concentration is influenced by MMP-3 -1612 5A/6A promoter genotype and associated with myocardial infarction. J Intern Med *258*, 411–9.

[139] Farrall, M., Green, F. R., Peden, J. F., Olsson, P. G., Clarke, R., Hellenius, M.-L., Rust, S., Lagercrantz, J., Franzosi, M. G., Schulte, H., et al. (2006). Genome-wide mapping of susceptibility to coronary artery disease identifies a novel replicated locus on chromosome 17. PLoS Genet *2*, e72.