Inherited Risk Enrichment Analysis of gene sets using Genome-wide Association Studies for Coronary Artery Disease

Hassan Foroughi Asl

KTH Datavetenskap
och kommunikation

Degree project in
Computer Science
Second cycle
Stockholm, Sweden 2013

KTH Computer Science
and Communication

Inherited Risk Enrichment Analysis of gene sets using Genome-wide Association Studies for Coronary Artery Disease

HASSAN FOROUGHI ASL

Master's Thesis at CSC
Supervisor:
Lars Arvestad
Johan Björkegren
Examiner: Jens Lagergren

TRITA xxx yyyy-nn

Abstract

Genome-wide association studies (GWAS) has been in the heart of medical research for the last 5 years. These studies seek for common variants in the genome that are linked to risk for common complex diseases (CCDs). Although GWAS has defined a number of interesting genetic loci for a range of CCDs, the current GWAS analysis has limitation such as investigating the DNA variants one-by-one focusing on the most significant DNA variants. As a consequence, most risk variants for CCDs are, in my belief, still hidden in the GWAS data. Herein, I use a method of GWAS analysis that considers risk-enrichment for groups of functionally associated genes defined by for example gene networks, believed to play a role in CCDs.

In this method, a set of expression SNP (single nucleotide polymorphism) was selected from genes which are known to be related to coronary artery disease (CAD) in a way that a single eSNP was chosen for each gene. Then using the data available from the International HapMap Project and a GWAS data available, it is possible to find SNPs which are in strong linkage with the initial set, which we call it expanded set. Depending on the association of the initial set to the CAD, expanded set can show an enrichment score greater or smaller compared to the null distribution set of SNPs with same properties of the expanded set.

In conclusions, CCDs are not a consequence of isolated genetic variants/genes in isolated pathways but instead sets of genetic variants/genes acting in conjunction, cause CAD. Genetic risk enrichment analysis is a fairly simple and straightforward method to determine to what extent a group of functionally associated genetic variants/genes are enriched for a given CCD. In addition, this analysis can perhaps help to decipher some of the 90-85% of risk variation in populations that remains unaccounted.

Acknowledgement

My foremost gratitude goes to my academic supervisor Johan Björkegren at Karolinska Institutet who generously took on supervision of this thesis project and giving me the chance to work on an excellent project and project group. Also, I would like to thank my local supervisor Lars Arvestad who read and improved my thesis, and my examiner, Prof. Jens Lagergren who approved my thesis.

I would like to thank the Department of Computational and Systems Biology, CSC, KTH for granting me the opportunity for this great master's program. I would also like to thank my friends and colleagues Babak Taghavi, Maya Brandi, Måns Magnusson, Roman Valls for making the world more pleasant. I'd would like to also express my gratitude to Josefin Skogsberg for proof reading my thesis and for all those great discussions.

I am also deeply indebted to my beloved Afsaneh, thank you for *everything*.

Last but not least, I especially thank my parents for providing all the things I need to pursue my studies in Sweden, you are the best!

Contents

I Introduction and Methods 1

1 Introduction 3
- 1.1 Background . . . 3
- 1.2 Genome-Wide Association Studies . . . 6
- 1.3 International HapMap Project . . . 6
- 1.4 Atherosclerosis and Coronary Artery Disease . . . 8
  - 1.4.1 Atherosclerosis Module . . . 10

2 Methods 11
- 2.1 Causality and eSNPs . . . 11
  - 2.1.1 LD Blocks and Statisctics . . . 12
  - 2.1.2 Combining GWAS with International HapMap Project 13
- 2.2 Disease Risk Enrichment Analysis Method . . . 13
  - 2.2.1 Enrichment of disease genes . . . 13
  - 2.2.2 Enrichment Score relative to background random SNP sets . . . 14

II Results and Discussion 17

3 Results 19
- 3.1 Validation of Method . . . 19
- 3.2 STAGE cohort and eSNP . . . 20
- 3.3 P-value Distribution and Discussion of Results . . . 21

4 Discussion 25
- 4.1 Interacting genotype with intermediate phenotypes to better understand disease phenotypes . . . 25
- 4.2 Future work and Final words . . . 27

Bibliography 31

Part I

Introduction and Methods

Chapter 1

Introduction

1.1 Background

Our current understanding of common complex diseases (CCDs) has mainly been established from using a candidate gene approach where one or a few genes in a given pathway are studied in relation to a common disease phenotype [31]. This type of reductionist approach has been in the heart of medical research for decades. However, given the level of complexity of gene interactions and regulations that are underlying these types of diseases [30], it is now evident that using the candidate gene approach alone will not give us enough clues to how these diseases can be best battled and treated. In fact, given the current estimates of the number of gene and molecule interactions underlying CCDs [27, 28], we have most likely only been looking at the tip of the iceberg. However, with emerging high-throughput methods such as DNA microarrays and increasingly efficient genome-wide sequencing technologies [22], it is now possible to conduct studies of CCDs from the perspective of the whole-genome generating datasets covering the full scale of molecular activities underlying CCDs.

In principal, DNA stores all information needed to create and maintain nature. In humans, the DNA code contains 3 billion DNA letters consisting of pairwise A-Ts or C-Gs. In most instances, the information in the DNA is activated through DNA transcription into RNA (figure 1.1). The type and number of RNA in a given cell is deciding its molecular task and therefore biological function. Particularly the mRNAs (messenger RNAs) are important since these strings of RNA are translated into proteins. Proteins are the molecular components that execute cellular functions. Although there are major efforts to understand all forms of molecules in the cell (e.g. proteins, metabolites, fatty acids and protein modifications), the technology development for whole-genome reads of cellular DNA and RNA are the most advanced. Since the DNA code is identical in all cells within one individual (besides spontaneous mutations), it is enough to determine the DNA code in one individual once (most commonly being isolated from blood).

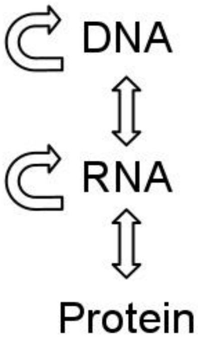


Figure 1.1: Central Dogma of Biology

In contrast, the transcription of DNA into RNA is marked differently in cells belonging to different organs. In fact, the transcription is to a large extent deciding the role of a given cell,tissue or organ. For example, the total RNA profile of a cell that is to become a liver cell is quite different from a cell in the vascular wall despite sharing basic molecular functions. To understand CCDs from a systemic perspective, it is necessary to isolate RNA from the range of organs/tissues believed to underlie the development of the CCD. The rational idea is that the type and concentration of cellular RNA are altered in a normal cell when it becomes part of developing CCDs. However, from the new technologies it is only possible to determine how the types and concentrations of RNA are changing in cells involved in a CCD but it is also necessary to understand how these changes come about.

For this particular purpose, new sophisticated algorithms have been developed enabling to compute gene interactions in networks from genome-wide DNA and RNA datasets. For DNA, there is an estimation of 0.1 % differences between any two individual [35]. Despite representing a small fraction, these DNA variants are believed to underlie most of the differences between humans. Besides functional differences (such as type of skeletal muscle , i.e. fast or slow muscle fibers), there are DNA variants among them that are thought to underlie risks for developing diseases. By searching for these variants using the new high-throughput technologies, it is possible to define those that are coupled to diseases. These variants can be in coding region (sections of the

DNA coding for genes) that are transcribed into mRNA and thereafter translated into proteins (figure 1.1) as well as in non-coding region. The latter is believed to affect risk for diseases by altering gene regulation.

In this report the idea is to look at DNA genotyping data, which can come from so called genome-wide association studies (GWAS) and it has been in the center of medical research for the last 5 years, revealing disease-coupled DNA variants, or as they are called in the field of genetics, genetic loci, that are highly associated with common complex diseases. GWAS have been paid a lot of attention since they have, in most instances, been performed on several thousand patients suffering disease as well as controls without disease. Thus, the genetic loci defined from whole-genome scans of up to 1 million DNA variants (selected among the 30 million that varies between individuals) are thought to be of high relevance to achieve a better understanding CCDs. Coronary artery disease (CAD), which is the CCD in focus in this thesis, is a disease of the vessel wall underlying myocardial infarction (MI) and 20% of strokes. CAD, MI and stroke are responsible for one third of total mortality rate in 2004 (i.e. 15-20 million deathes per year) and is believed to become one of the deadliest CCD (including infectious diseases like Malaria) this year (predicted to reach 30 million death annually by 2030) [17].

Although GWAS have defined a number of interesting genetic loci for CAD/MI , there are some important shortcomings with the design of how genomic data in GWA datasets conventionally is analyzed. These can be find from GWAS catalogue [10, 9]. The choices of analytic approaches underlies only a minor fraction of the total risk for CCDs have been accounted for (5-15% of total risk variation for most CCDs). In this thesis, I suggest a parallel track to conventional analysis of GWA dataset to reveal a greater portion of risk and thereby the etiologies of CAD/MI than what has been achieved by traditional GWAS analysis alone.

Instead of solely seeking the most significantly related DNA variants associated to CAD/MI by analysis of DNA variants one by one, we suggest to use GWA databases of CAD/MI (such as Wellcome Trust Case Control Consortium [36]) to analyze groups of DNA variants (e.g. single nucleotide polymorphism SNP). These SNPs are defined by groups of functionally related genes associated to CAD/MI, identified from analyzing mRNA data. These groups of genes can simply be lists of differentially expressed genes or gene clusters but in most instances it can be identified by gene networks associated to CAD/MI. From experience of network inference in type II diabetes and obesity [27], gene networks can be defined from genome-wide liver and fat mRNA data isolated from patients suffering these diseases. From these previous examples, networks may harbor important disease mechanisms underlying CAD (or any other CCD). By linking genes in networks to SNPs that affect their expression level (i.e. mRNA levels), also referred to as eSNPs (expression SNPs), we can analyze their combined enrichment of inherited risk for CAD using GWAS data sets. In this fashion we hope that a larger portion of the variations which are responsible for risk of developing CAD can be revealed.

1.2 Genome-Wide Association Studies

Basically, a genome-wide association study (GWAS) is a study in which variations of two groups of cohorts (individuals with a trait (i.e. cases) and individuals without that trait (i.e. controls)) are compared in search for association with a particular disease. GWAS builds on an approach to find disease-linked variants by analyzing the whole genome. When the first GWAS study result was published in 2005, the result was more than 110,000 SNPs linked to complement factor H polymorphism [12]. In fact this was the first GWAS study to use commercial genotyping platforms representing SNPs in genes. Using the commercial genotyping platforms has one downside, it only reflects the common variants and eventually focusing on common variants for CCD [29]. Indeed, the idea behind developing GWAS is based on common disease common variant (CDCV) hypothesis [29], so it should not apply any shortcomings for GWAS analysis.

Later, with availability of more advanced technologies, many GWAS have been performed. A list for recent and as well as older GWAS is gathered at National Human Genome Research Institute GWAS catalog [10] There are several data sets available, but in this thesis, the Wellcome Trust Case Control Consortium (WTCCC) GWAS data has been used [36].

It should be mentioned that although SNPs can explain a fair amount of inherited risk for a disease [14], there are also non-SNP variants, such as: copy number variation (CNV), structural variants, deletion, insertion, etc [35]. In fact the number of non-SNP variants by total number of nucleotides is believed to be more than SNP-variants, but SNP-variants are still considered the most common type of variation in human genome.

1.3 International HapMap Project

Variations in human genome are mostly the result of genomic events, such as recombination and demographic events, therefore a population is neither typical nor genetically exceptional compared to other populations (e.g. Caucasion (CEU) vs Japanese (JPN)) [6]. Any two individuals have 99.9%

identical genome, 0.1% difference means almost one variant per 1,000 bases [15, 4, 37, 35].

I wrote that the SNP is the most common type of DNA variants, an estimation of total human population variation is almost 10 million SNP sites. This means roughly one SNP per 300 bases in DNA and they constitute 90% of total variation in the human population[13] [25]. In this thesis, non-SNP variants are out of scope and the term variant and the word variation will be reserved for SNP-variants only.

Because of historical genomic events, nearby SNPs are associated with each other. Each of these SNPs constitute a form of allele, which in turn is form of a genetic locus [19]. (Figure 1.2). For example, at a SNP site, a particular gene can be C allele or T allele [33]. A set of alleles that is on a single chromosome is called a haplotype [33, 19]. Haplotype formation and changes are caused by mutations and/or recombination. Therefore these haplotype blocks will include certain nearby SNP alleles. These SNPs will have association to each other and tend to co-vary over the same haploblock. In other words, genotypes of a pairs of SNPs are not independent of each other. This non-random association of SNPs is defined as linkage disequilibrium (LD) [24]. LD is very dependant on distance and recombination rate in each population, and it can range from 60 kb (kilo base pairs) up to 200 kb long. Mappings of these associations between pairs of SNPs form block like structures that span throughout the genome. Many genomic events affect the extension of LD blocks. A long LD block can be because of population bottleneck or founder effect, and shorter LD block, on the other hand, can be because of high recombination rate in a population. In LD blocks, there are so called tag SNPs. These tag SNPs are used to determine haploblocks [33].

There are several software packages available to calculate and visualize the LD correlation data. Two examples are Haploview [3] and PLINK (Population-based LINKage analysis) [21]. But there are pre-calculated LD correlation values provided by HapMap, in which Haploview was used in the calculation. An example of LD blocks is illustrated in figure 1.3.

During the two phases of The International Hapmap Project more than 4 million SNPs were genotyped [35]. In phase I, at least one SNP per approximately 5kb was genotyped with a targeted minor allele frequency (MAF) of 0.05 [34]. However, during the second phase there were a few changes to increase the SNP density, higher SNP density means more coverage of the common variations. The result was differences between phase I and II, some of these differences are lower MAF and different LD statistics [35]. Lower MAF is good in the sense that more rare variants will be covered in phase II compared to phase I. It should be noted that in this thesis release 27 of HapMap dataset will be used, which can be found at:

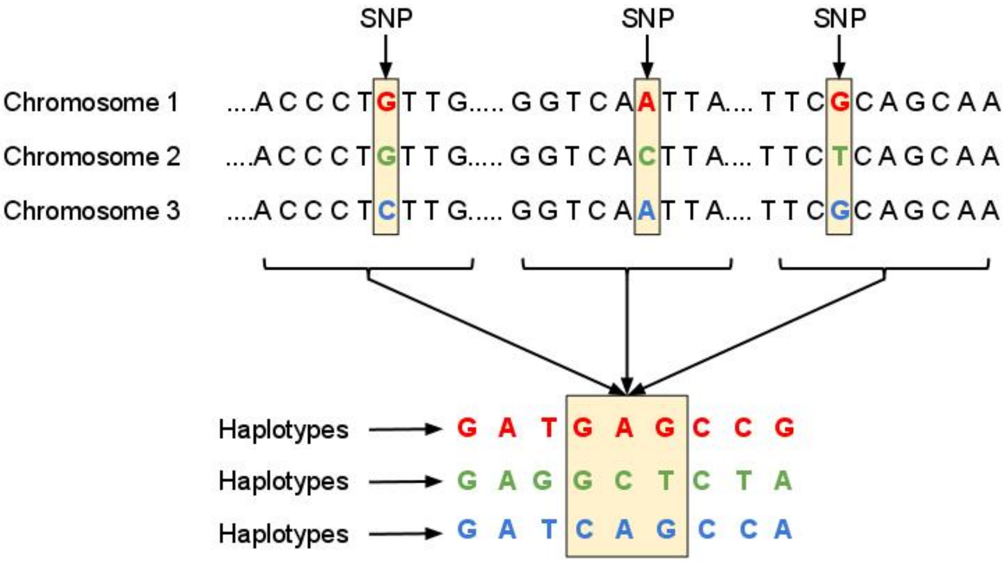

SNP
SNP
SNP
Chromosome 1
....A C C C T G T T G..... G G T C A A T T A.... T T C G C A G C A A
Chromosome 2
....A C C C T G T T G..... G G T C A C T T A.... T T C T C A G C A A
Chromosome 3
....A C C C T C T T G..... G G T C A A T T A.... T T C G C A G C A A
Haplotypes
G A T G A G C C G
Haplotypes
G A G G C T C T A
Haplotypes
G A T C A G C C A

Figure 1.2: Reproduced from International Hapmap Project [33]

`ftp://ftp.ncbi.nlm.nih.gov/hapmap/ld_data/2009-04_rel27/`

From here on the term LD block will be used in the context of 200kb upstream and downstream of each SNP and those SNP that are in this range and are in LD with the the central SNP. This coverage is defined based on LD span in Caucasian which is provided by HapMap.

1.4 Atherosclerosis and Coronary Artery Disease

Atherosclerosis is the underlying cause for CAD/MI and it is a good example of CCD. In this disease, different molecular events, cells, tissues and environmental factors influence its progress [23] [30]. Atherosclerosis affects the arteries and the atherosclerotic lesions are result of accumulated lipids, inflammatory cells and fibrous elements [16]. Large arteries consists of (three) different layers. The top covering layer cells, Endothelial cells (EC), provide a selective permeable barrier for blood and arteries inner layers [16]. In large arteries, specially in branching points, blood turbulence affects the shape of ECs which in turn will change their permeability, which will allow large molecules such as low-density lipoprotein (LDL) pass through tight junctions. This will be the initiation of atherosclerotic lesions in such sites. After passive diffusion through tight junctions, LDL molecules are modified by reactive oxidative species to generate oxidized LDL (oxLDL) [16].

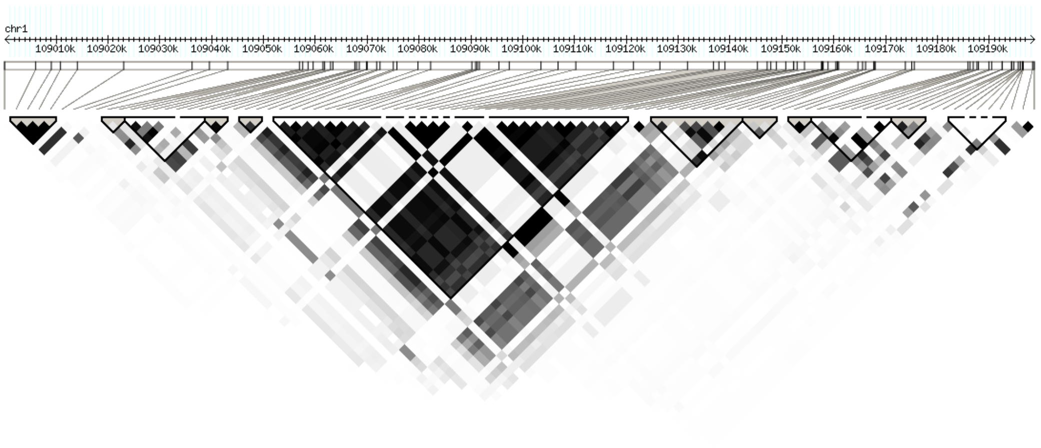


Figure 1.3: An example of LD structure produced by Haploview software [3]

An inflammatory response is then triggered by oxLDL, attracting monocytes and other inflammatory cells [16]. The monocytes differentiate into macrophages in the vessel wall and will therefore take up LDL rapidly and turn into foam cells. After initial inflammatory response and the foam cell formation, smooth muscle cells (SMC) start to migrate from the media [1] and SMC-derived extracellular matrix formation occurs. This will lead to formation of fibrous plaques. In later stages, calcification occurs (similar to bone formation) which will affect the stability of the plaque. Unstable plaques or plaques with thin EC layer with high degrees of lipids and macrophages, tend to rupture frequently at the edges. If rupture happens, it can cause thrombosis, which in turn can result in myocardial infraction (MI) and stroke.

In summary, atherosclerosis is a complex disease. On one hand we mentioned that many different cell types, molecular and non-environmental factors affect the disease development and progress. These non-environmental factors are high LDL, low HDL, family history (inheritance), gender (male), inflammation, etc [2] [30] [7]. On the other hand environmental factors such as high fat diet, smoking, lack of physical activity, etc. exist which can increase the risk for this disease [30].

---

[1] Media is the middle layer of an artery

1.4.1 Atherosclerosis Module

The STockholm Atherosclerosis Gene Expression (STAGE) study was one of the first efforts ever using a systems approach to identify functionally associated genes related to coronary artery disease (CAD) using whole-genome expression profiles from multiple organs [8]. The STAGE study consisted of 114 patients with gene-expression profiles from five CAD related organs. In this study, 128 genes were found to be enriched with genetic risk of CAD and linked to the extent of coronary atherosclerosis (i.e. "coronary stenosis"). This set of genes was therefore referred to as the Atherosclerosis Module (A-module).

eSNP Discovery

Basically, there are two types of eSNPs: cis-acting and trans-acting. They are divided based on physical distance to the gene that they are correlated with. Cis-acting expression SNPs are defined as those situated 1Mb upstream or downstream of the transcription start site of every gene. Cis-acting SNPs, are necessarily those reside on the same chromosome. In contrast trans-acting SNPs are either $>$ 1Mb up- or downstream of any gene on the same chromosome or on different chromosomes than the regulated gene. One immediate explanation for the choice of 1Mb is the total number of genes and size of human genome. The way the eSNPs were chosen to be defined as cis- vs trans-acting eSNPs, are consistent with previous works [28]. Expression SNP calculation is not covered in this report.

Chapter 2

Methods

2.1 Causality and eSNPs

According to the definition of LD explained in the introduction, by having genotyped data for a set of known and common variants related to a particular CCD, with proper data mining, it is possible to expand the eSNP set corresponding to a set of genes in search for correlations with disease SNPs. But a question remains whether they will be causally related to the disease or not.

Causality refers to the relationship between a locus and traits, in such a way that a variation in locus can affect expression of a gene or genetic locus which in turn contributes to the complex trait [27]. Schadt et al 2008 [28], proposed different models for how these can be related to each other. Causal model is the model where a DNA locus (L) affects the gene expression levels (R) which in turn affects a complex trait (C) , in other words "simplest causal model is: L acts on C through transcript R" [27]. However, a parallel reactive model was also suggested (i.e. "R is modulated by C").

Using these models, Schadt showed that by combining whole genome DNA variation data (i.e. global SNP profiles) with gene expression (i.e. in RNA expression levels) isolated from a patient cohort with a common complex trait (C), it is possible to "decipher" those loci that are casually related to the trait. However, the approach explained above needs the construction of gene co-expression networks that is beyond the scope of this thesis. Instead I will use LD statistics to analyze the relatedness of given set of eSNP to a common complex trait (C), which is in our case is CAD, and present a measure to be able to compare different eSNP sets from different tissue sample from different cohorts.

2.1.1 LD Blocks and Statisctics

I said that LD blocks in human genome are not continuous. This lack of continuity in LD block is mainly because of recombination hot spots [34]. For this reason some SNPs can be in comparably smaller blocks than other SNPs, and it will limit the number of SNPs they are in LD with. But the length of LD block is one factor. The other factor is the strength of correlation between two SNPs in the same LD.

There are some insights in the nature of the segmentation and block-like structures of LD. First, it will make it possible to define a tag SNP[1] for each haplotype and take it as a representative for the other SNPs under certain conditions (LD statistics). Second, is the possibility to find other common variants related to a disease by finding disease SNPs that are in a strong correlation with SNPs previously identified as to be related to the disease (i.e. eSNPs). This can help to identify novel loci associated with the disease. Last but not least, the strength of correlation between SNPs in the same LD block can help us define certain proxy SNPs [34], this is not the same as tag SNP. Tag SNP represent the whole block, but proxy SNPs are just representing those SNPs that they are in strong correlation with.

When it comes to the statistics of LD, there are several measures available for LD statistics[11], here we choose r^2 ($0 < \mathrm{r}^2 < 1$). Since LD blocks are disrupted by historical events, it is also reflected in r^2 measures (e.g. recombination rates [5]).r^2 is the square of correlation coefficient between two SNPs, shows a distorted pattern in an LD block with decay towards the furhter SNPs [34, 11]. The values for it are independent of physical distance, and adjusted for allele frequency.

It seems r^2 is a good measure when looking for SNPs to represent other SNPs. Because in the results that was shown in the International Hapmap Project's last two published articles they show that by having certain number of SNPs at a $r^2 \geq 0.8$, it is possible to capture all common SNPs at minor allele frequency (MAF) ≥ 0.05 (i.e. tag SNPs) [34, 35]. This number for Caucasian population (CEU) is approximately 552,000, this means capturing a small subset of SNPs can act as surrogates for a large set of SNPs. They defined the term proxy SNP as a SNP which shows a strong correlation with one or several other SNPs. The term perfect proxy refers to the complete LD between one proxy SNP and others, i.e. $r^2 = 1.0$ [33, 34]. Here the term expanded SNP list will be used to refer to those SNPs that are in LD to any of the proxy SNPs at at defined r^2, the value of r^2 can change depending on the type of analysis. This will be further clarified in the following chapters.

---

[1] Tag SNPs set is the minimum number of SNPs to identify a haplotype

2.1.2 Combining GWAS with International HapMap Project

GWA studies are important because they make it possible to identify the most significant loci associated with a CCD. However, they fail to define the majority of the risk for disease. In the description for LD statistics, we discussed the proxy SNPs and expanded SNP list. If we have a set of SNPs that are functionally associated with a disease through genes and genetic loci (i.e. eSNPs), we can use the LD data available in HapMap, to expand our list to bigger set of SNPs that are in strong LD with significant disease SNPs (dSNP) represented by a GWAS.

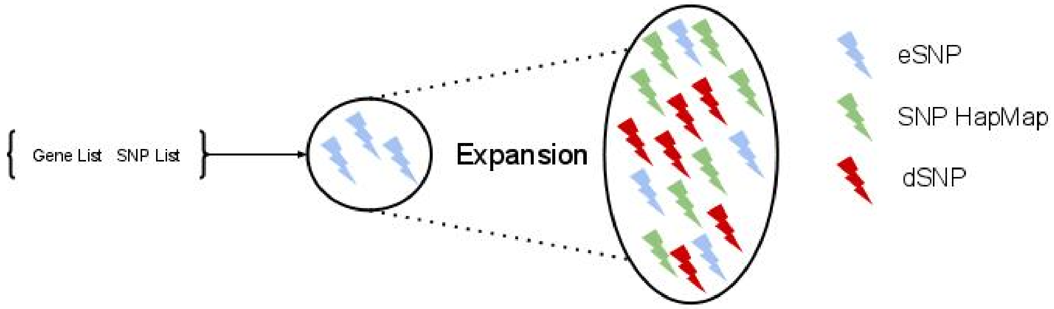


Figure 2.1: Combining GWAS and HapMap

It is clear that some eSNPs that don't fall into the coding region of a gene, which is the case for most of the eSNPs, are instead located in introns or intergenic loci. GWAS studies can capture only a subset of the variations in the DNA[32, 38]. Some of the uncaptured variation may be important for CCD. Also there could be rare variants ($MAF < 0.05\%$) that remain undetected due to the density of SNPs on the microarrays. Using LD measures (r^2 in this thesis) provided by HapMap, helps linking those SNPs that are not present on the microarray used by different GWAS.

2.2 Disease Risk Enrichment Analysis Method

2.2.1 Enrichment of disease genes

So far we have been discussing how SNPs can be correlated to each other, and how eSNPs can represent the genes they regulate and eventually how these eSNPs can be correlated to CCDs.

Based on these principles there are only a few steps for enrichment analysis of inherited risk for CCDs:

1. **Gene Selection:** Selection refers to choosing a set of genes that are under investigation to be in association with disease (an example is a set of genes defined by a network). One example can be the data available by genetics of gene expression studies (like the STAGE dataset).

2. **eSNP calculation:** In this step eSNPs are calculated for the selected gene list using Genetics of Gene Expression (GGE) studies data, where SNPs for each gene is selected based on their allelic association with the mRNA levels, to define eSNPs.

3. **Expansion:** Expansion refers to expanding the set of eSNPs using LD measures and HapMap. After this step, the eSNP set is referred to as expanded SNP set.

4. **Enrichment:** The expanded SNP set is examined for their enrichment of SNPs associated with CCD according to GWAS dataset.

There are some criteria that needs to be considered here. First of all, there are cases that either one SNP is in LD with several eSNPs or one eSNP is in LD with multiple SNPs. If this happens, the strongest correlation will be chosen (i.e. stongest LD value for SNP-eSNP pairs). Secondly, there could be cases where some eSNPs are in LD with other eSNPs, and it is not inevitable, since eSNPs by definition are those SNPs that affect the expression of genes. In addition there could be multiple eSNPs for the same gene [39], in these cases the best (strongest) eSNP is choosen. Alternatively, there could be several genes that share same eSNP, however this will not affect the enrichment analysis. Because after step two, the list that can affect the enrichment is the list of eSNPs and not genes.

2.2.2 Enrichment Score relative to background random SNP sets

In order to assign an enrichment score for each set of genes and corresponding eSNPs, the number of SNPs with p-value < 0.05 was compared to the number of SNPs with p-value < 0.05 in a competitive null hypothesis[2]. To construct the competetive null hypothesis 10,000 random sets SNPs, equal to the number of eSNP sets, are generated. Then the expansion and enrichment step is applied to each of the sets. To achieve a higher accuracy in this method, it is best to mimic the initial SNP set as much as possible. In general the selection of SNPs in random set are based on: (i) similar LD structures as of initial SNP set (i.e. if two SNPs are in LD with each other, no matter how strong this LD is, the random set should also at least contain two SNPs that are in LD with eachother), (ii) the chromosomal distribution of initial SNP set (iii) same minor allele frequency[3] cut-off, and (iv) last but not least, number of SNPs in random set should match as of initial SNP set.

---

[2] Competitive null hypothesis here is that random set of SNPs will at most have the same number of SNPs with p-value < 0.05 as the non-random set.

After applying these criteria there is no guarantee that the final set (after step four) will be the same number of SNPs as the final set for initial SNP set. To over come this, the enrichment score is defined as below:

$$F = \frac{\frac{N_{H1(p<0.05)}}{N_{H1}}}{\frac{N_{H0(p<0.05)}}{N_{H0}}} \tag{2.1}$$

F stands for fold enrichment, N_{H1} and N_{H0} are the number of SNPs in the enriched set for initial SNP set and random set, respectively. $N_{H1(p<0.05)}$ and $N_{H0(p<0.05)}$ are the number of SNPs in enriched set at p-value < 0.05 for initial SNP set and random set, respectively. The above equation is applied to each of random sets, and as a result 10,000 F values are generated. The final F value will be the mean value of all of them.

---

[3] Minor allele frequency (MAF) refers to lowest possible allele frequency of a SNP in a given population

Part II

Results and Discussion

Chapter 3

Results

3.1 Validation of Method

To investigate the efficiency of the method, a list of SNPs known to be related to CAD was first chosen. Thus, this list represent positive controls to confirm the validity of the method. This list was generated from GWAS catalog [9, 10], by selecting all main categories that were related to CAD (i.e. coronary heart disease, myocardial infarction). GWAS catalog list is consisting of 123 SNPs. Then inherited risk-enrichment analysis explained in Methods (chapter 2) was applied to this set of SNPs. What we are expecting here is a high enrichment score. As expected, high enrichment scores for this set of CAD associated SNPs was observed (tables3.1). Summary of expansion and enrichment steps are in tables 3.2,3.3.

	Number of SNPs	
SNP set	Initial size	Expanded Set
GWAS Catalog	123	36291
A-module 1 Mb	1247	71389
eSNP STAGE	448	29082

Table 3.1: Number of SNPs in initial sets and corresponding expanded sets. After selecting the SNPs/eSNPs for each set, each of them was expanded according to step three in the method described in section 2.2.1.

SNP set	$r^2 =$	0	0.6	0.7	0.8	0.9
		Number of disease associated SNPs				
GWAS Catalog		5576	311	253	199	166
A-module 1 Mb		10171	1250	1102	933	809
eSNP STAGE		4226	521	484	462	451

Table 3.2: **Number of SNPs in each set at different r^2 thresholds according to one GWAS dataset and each set's corresponding expanded list (step three in 2.2.1).** This was done based on step four in the method described in section 2.2.1. Given the definition of LD, the ones that show higher F value at higher r^2 thresholds, will be the ones include SNPs with higher inherited risk.

SNP set	$r^2 =$	0	0.6	0.7	0.8	0.9
		Enrichment Score (F)				
GWAS Catalog		1.72	8.11	8.62	7.91	7.76
A-module 1 Mbp		1.12	0.92	0.99	1.09	1.06
eSNP STAGE		1.25	1.58	1.59	1.63	1.67

Table 3.3: **Fold-enrichment score of disease association for each set.** This score is calculated according to the method described in 2.2.2. Scores with $F > 1$ represent an existing association to disease, and the higher is F, the stronger will be the association.

3.2 STAGE cohort and eSNP

Next the inherited risk enrichment procedure was applied to data generated from STAGE dataset. As it was discussed in the introduction, A-module consists of 128 genes, which are believed to be related to atherosclerosis. Thus, this list was used to perform the inherited risk-enrichment for a group of SNPs postulated to be associated with CAD [8]. From this a list of 128 genes, a list of 1247 tag SNPs within 1 Mbp of each gene was selected. A second set consisting of eSNPs, calculated from the STAGE cohort [8], was also tested. Since A-module has been shown to be associated with CAD [8], the enrichment results for the associated SNPs and/or the eSNPs should possibly confirm that. these eSNPs should also confirm that. The summary of the results are in tables 3.1, 3.2, 3.3.

As shown in tables 3.1 - 3.3, eSNPs for the A-module genes were enriched for inherited risk of CAD according to the GWAS cohort of WTCCC. In contrast, SNPs selected for the same genes (A-module genes) based on proximity alone (i.e. ± 1Mb) were not (i.e. 1.06 at $r^2 \geq 0.9$). These results show the importance of choosing accurate SNP (i.e. eSNPs) to represent groups of genes that are under investigation.

3.3 P-value Distribution and Discussion of Results

We saw that the selection of SNPs can affect the risk enrichment results, the stronger the association between SNPs and genes (i.e. eSNPs), the higher the enrichment score (F) will be. High values of F reflect higher association of initial SNP/gene set with disease. So the set of highly associated SNPs (e.g. gathered from previous GWAS studies) should show an increased risk enrichment for CAD. This can be observed in a histogram of the p-value distribution (figures 3.1, 3.2, 3.3), the more risk-enriched the more shifted to the left.

In figures 3.1, 3.2, 3.3, the number of SNPs in each bar is normalized based on a total number of SNPs in each set. A clear shift to the left, i.e. lower p-values, is an alternative measure for the association of the set with the disease.

By increasing the threshold for r^2 (i.e. $0.0 \rightarrow 0.9$), we capture those genotypes that are in stronger LD with disease SNPs. The value for r^2 is, in this way, depending on how strong the correlation is with disease. Differences in genotype platforms, and as well as differences in type of patient cohorts and microarray, are the factors that decide the choice of r^2. However, it is also important to mention that if by increasing r^2 threshold greater than 0.0, a decline in F is observed, it represents a poor (or no) association to the current GWAS, specifically, as disease risk model and to disease in general.

In conclusion, there are some limitations that needs to be considered whilst performing disease-risk enrichment analysis. However, with current advances in genotyping and gene expression technologies (i.e. gain from microarrays to sequencing) huge amounts of data leading to increasing matches between platforms will lead to that these limitations will become negligible.

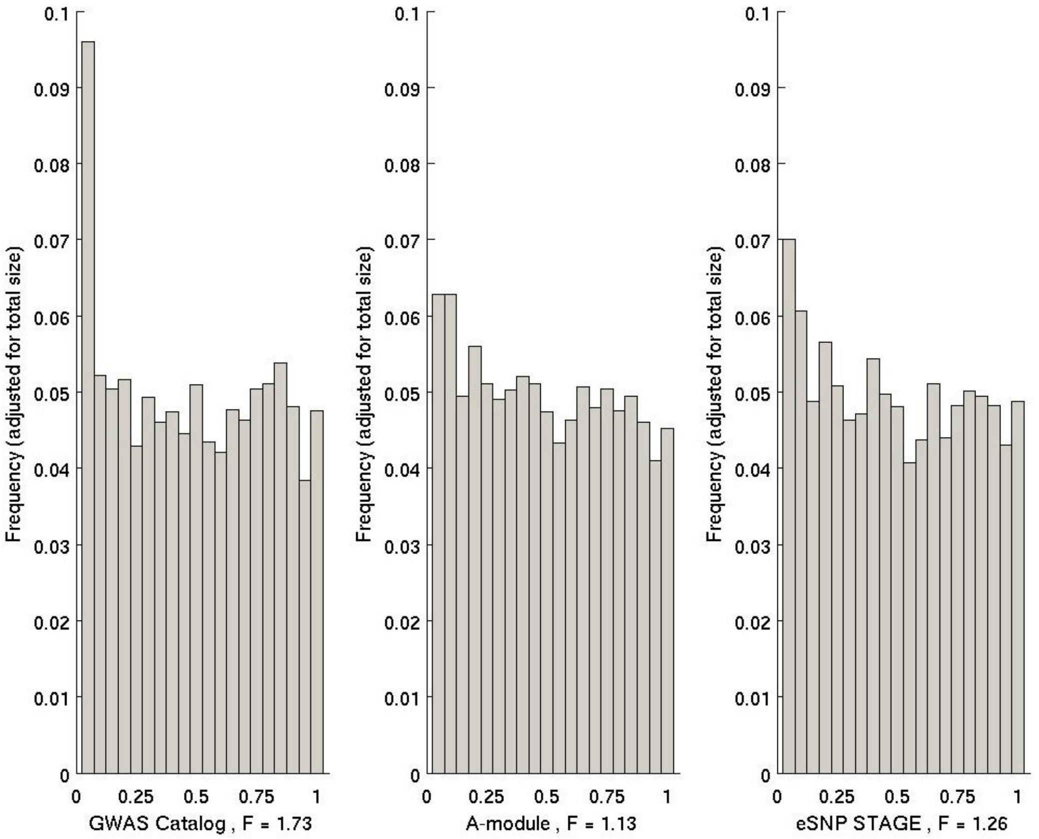
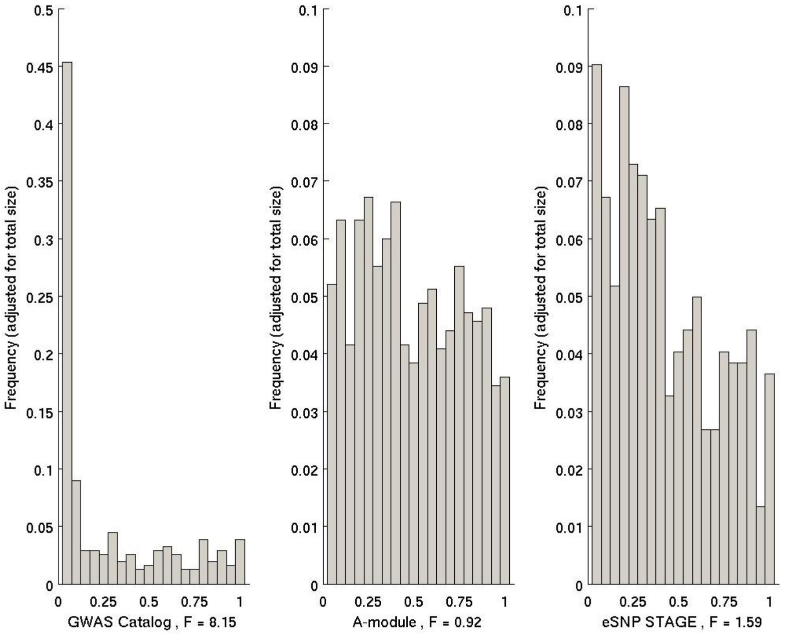
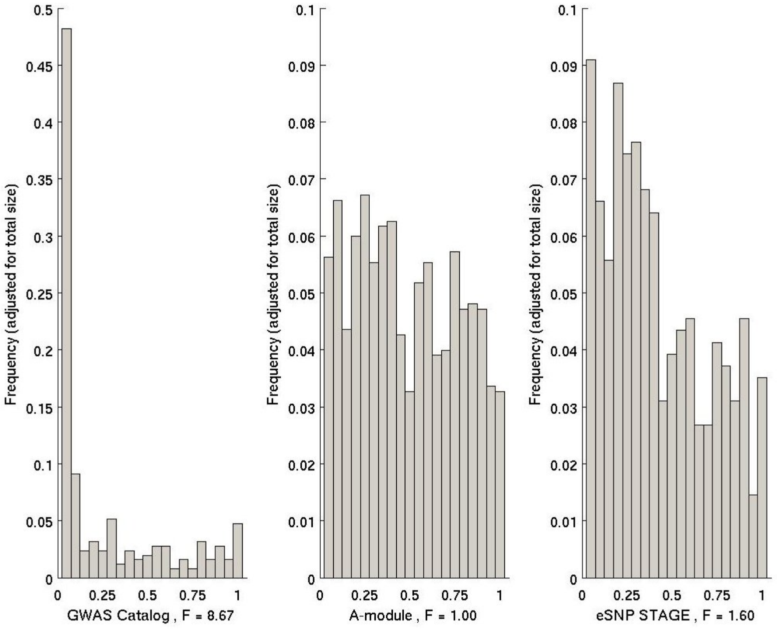


Figure 3.1: **P-value distribution at $r^2 \geq 0$.** Notice the shift to the lower p-values is in accordance with the higher values of F. The shift to the left, however, is not an accurate measure.

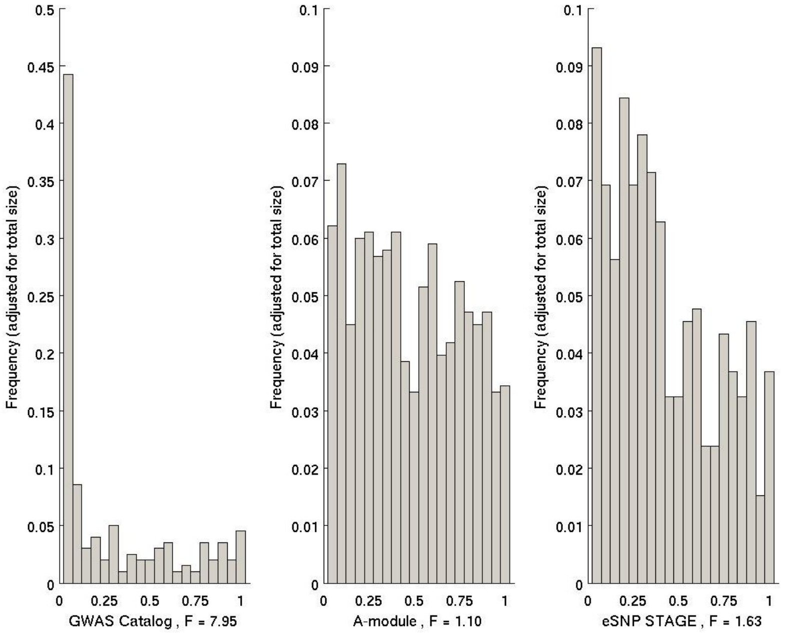


(a) $r^2 \geq 0.6$

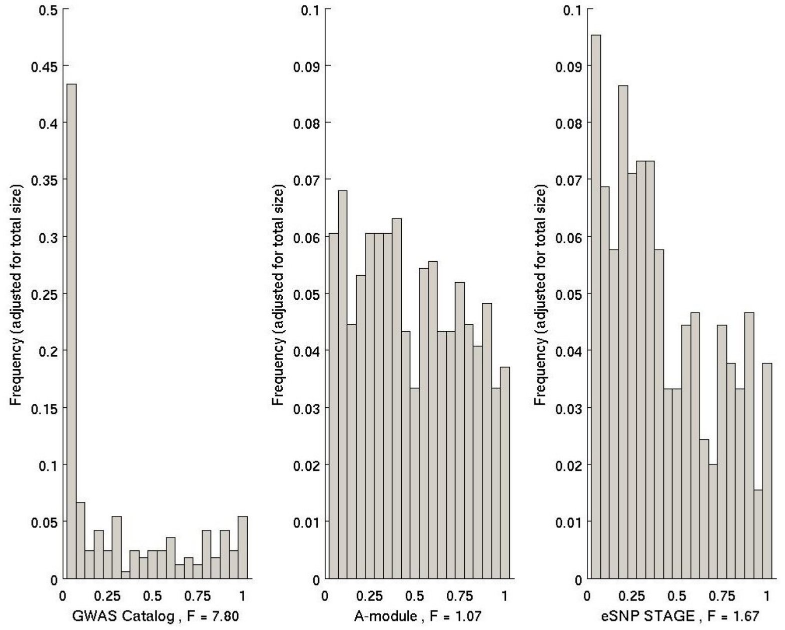


(b) $r^2 \geq 0.7$

Figure 3.2: **P-value distribution at different r^2 thresholds.** The general shift to the more significant p-values is not strong anymore in the A-module set. Note the different scale of Y-axis in the left most plots



(a) $r^2 \geq 0.8$



(b) $r^2 \geq 0.9$

Figure 3.3: **P-value distribution at different r^2 thresholds.** The general shift to the more significant p-values is not clear anymore in the A-module set. Note the different scale of Y-axis in the left most plots

Chapter 4

Discussion

4.1 Interacting genotype with intermediate phenotypes to better understand disease phenotypes

Inherited risk enrichment analysis of gene sets believed to be important for CCDs are making new use of the GWAS datasets. The proposed method can be used for any candidate gene list, independent of how this list has been generated (i.e. from disease networks, gene clusters or lists of differentially expressed genes either from human or mouse models where disease driving genes have been deleted (i.e. knockout mice)).

The method can also be used to determine whether a group of genes are causally linked to disease or not (i.e. not reactively linked genes; altering their expression levels secondary to disease development, see chapter 2.1). Reactive genes will not be risk-enriched whereas causal genes are expected to. The extent of risk-enrichment can help to rank gene lists based on their degree of risk enrichments and thus relevance for disease. When several different CCD-type of GWAS datasets are available, the risk-enrichment analysis can help to show what type of CCD a given gene list is most relevant.

In order to perform risk-enrichment analysis, access to some type of data is essential. First, it is vital to access the entire datasets of GWAS (not just the top hits which are available at GWAS catalog [10, 9]). To exclude bias and cover lack of power from one GWAS several GWAS per CCD (or at least two) is preferable [1, 20]. Next, and as demonstrated, when assigning SNPs to a given gene list, the best way is to use actual eSNPs. Now, to compute eSNPs, GGE datasets are required. Preferably, these datasets should also have been retained from the same CCD as the GWAS. One cannot rule out that the repertoire of eSNPs is changing (at least in part) with different CCDs. This notion is supported by the fact the majority of eSNPs that we have identified (unpublished data) are tissue-specific (only affecting mRNA of a gene in one tissue). We interpret this as the local environment in a given tissue is affecting the combinations of DNA binding proteins and transcription factors which in turn will alter which DNA sites (i.e. SNPs) are having an effects on transcription or not. Since CCDs at least in some tissues are changing this microenvironment [26], it is likely that some eSNPs are CCD-specific as well.

The current computation challenges in choosing random sets to assess the enrichment score (F) and eSNP calculation demand access to larger computing power in form of computer clusteres or some cloud solution. However, NIH is building a repertoire of tissue mRNA profiles with associated DNA genotype profiles that can be used to define SNPs (i.e. eSNPs) for a set of genes to enable risk-enrichment analysis. If eSNPs are not available, SNPs within 400kb up to 1 Mb can be selected based on HapMap, as described. Using this less selective SNP-selection approach may, however, lead to false negative results and strong enrichments shall also be considered with caution. Thus, it is preferable to use eSNPs.

The risk-enrichment analysis stands in great contrast to the traditional analysis of GWAS where SNPs (and other genetic loci) are analyzed one-by-one. This has lead to a vast multiple testing problem where the level of significance for "true" hits is very high ($P<10^{-8}$). Surely many of the disease SNPs in the range of 10^{-8} - 10^{-3} may also be true disease loci. Also, and more importantly, many risk variants with relative low disease-association if analyzed one-by-one may very well when analyzed as part of a group of functionally linked genes (and corresponding eSNPs) contribute strongly to disease [27]. Using risk-enrichment analysis, these combinations of risk-variants can be revealed allowing CCDs to be analyzed from the perspective of groups of genes linked to disease in their capacity of taking part in common disease process. In our view, this is highly desirable.

A current drawback in the risk-enrichment analysis is the existence of many different genotyping array densities and platforms. Some of these are not sufficiently large to actually be truly "genome-wide". For problems of this nature, the international HapMap project, who identified approximately 4 Million SNP for the Caucasian (CEU) population [34, 35] provides the help needed enabling to expand the initial eSNP set to cover a larger portion of disease linked SNPs, as described previously.

The same problems also concern the GWAS datasets. However, in our case since the WTCCC dataset is based on the same population panel[36] but screened less number of SNPs (n = 500 000), it can actually be viewed as a subset of the HapMap dataset. However, caution need to be taken to match

GGE and GWAS datasets as to technical platform used (i.e. microarrays) and the ethnicity of the cohorts screened (making sure that ethnicities are the same in GGE and GWAS).

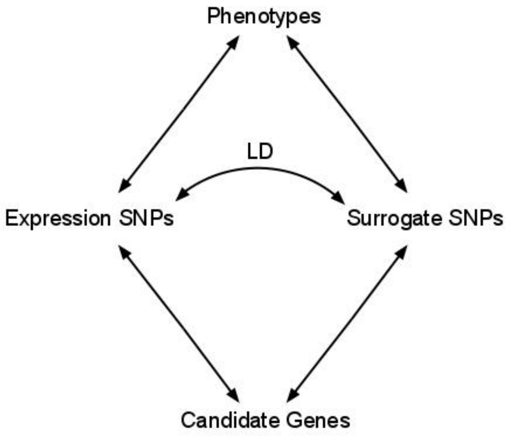


Figure 4.1: From candidate genes to phenotypes using eSNPs and surrogate SNPs.

In the longer perspective, GWAS is expected to be screened using DNA sequencing. A larger cohort of CAD patients consisting of 700 CAD patients is intended to be screened with DNA and RNA sequencing to establish a data intensive GGE cohort. Although, this will present with a substantial data-handling challenge, the need for expanding sets of genes/eSNPs using surrogate SNPs to explain observed phenotype (figure 4.1) will be largely avoided when full details of entire genomes and transcriptomes are available.

4.2 Future work and Final words

At the time when GWAS were designed, inheritance for CCDs was believed to follow a similar pattern as Mendelian rare diseases. It is now evident that this was a false assumption. Instead CCDs are influenced by many genetic loci affecting the expression of many genes as well as many disease processes active in specific tissues as well as across tissues shifting over time. The missing 85% of the heritability that is not explained by common variants already identified by GWAS (15% of total risk variation in most instances), I believe remain “hidden” in the GWAS datasets among the less significant disease-associated SNPs (as well as other variants). Applying the proposed risk-enrichment analysis, the portion of these that are indeed contributing to CCD risk can be identified as part of groups of functionally linked SNPs/genes. Such groups can efficiently be investigated for their risk-enrichments as proposed in this thesis using CAD as the example disease. Previous studies of Diabetes Mellitus Type 2 and obesity show the viability of this approach [27]. In addition, the proposed risk-enrichment analysis holds promises to make better use of the large GWAS investments made over the last 5 years.

Of note, risk inheritance is also increasingly believed to be mediated by epigenetic alterations to particularly DNA (e.g. DNA methylation) and CCD inheritance may also be mediated through non-RNA intermediate phenotypes like proteins and metabolites. These sources of risk inheritance and risk mediators need also to be considered in the risk-enrichment analysis if we are aiming to map the entire variation of risk for CCDs in different populations. Currently, there is no strategy for this using the method presented in this thesis but as GWAS are being complemented with epigenetic screens, the principals of the risk-enrichment analysis can be applied also to non-DNA genotypes as well as non-RNA intermediate phenotypes.

After this thesis, I am pursuing PhD studies in the same laboratory with the overall goal to map and rank groups of functionally-associated genes in CAD based on their risk enrichments using the WTCCC ([36]) and MI-Gen ([18]) GWAS datasets together the STAGE GGE cohort to define eSNPs. In addition, I will seek access to additional GWAS datasets and imputation methods to improve coverage [38]. During these 4 years, the sequence data of the expanded STAGE cohort (from 124 patients to 700) will become available. This will, in our belief, serve to enable us mapping a large portion of the entire landscape of inheritable risk for CAD.

List of Figures

1.1 Central Dogma of Biology 4
1.2 Reproduced from International Hapmap Project [33] 8
1.3 An example of LD structure produced by Haploview software [3] . 9

2.1 Combining GWAS and HapMap 13

3.1 **P-value distribution at $r^2 \geq 0$.** Notice the shift to the lower p-values is in accordance with the higher values of F. The shift to the left, however, is not an accurate measure. 22
3.2 **P-value distribution at different r^2 thresholds.** The general shift to the more significant p-values is not strong anymore in the A-module set. Note the different scale of Y-axis in the left most plots 23
3.3 **P-value distribution at different r^2 thresholds.** The general shift to the more significant p-values is not clear anymore in the A-module set. Note the different scale of Y-axis in the left most plots 24

4.1 From candidate genes to phenotypes using eSNPs and surrogate SNPs. 27

Bibliography

[1] David Altshuler and Mark Daly. Guilt beyond a reasonable doubt. *Nature genetics*, 39(7):813–5, July 2007.

[2] G. Assmann, P. Cullen, F. Jossa, B. Lewis, and M. Mancini. Coronary Heart Disease: Reducing the Risk : The Scientific Background to Primary and Secondary Prevention of Coronary Heart DiseaseA Worldwide View. *Arteriosclerosis, Thrombosis, and Vascular Biology*, 19(8):1819–1824, August 1999.

[3] J C Barrett, B Fry, J Maller, and M J Daly. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics (Oxford, England)*, 21(2):263–5, January 2005.

[4] M Cargill, D Altshuler, J Ireland, P Sklar, K Ardlie, N Patil, N Shaw, C R Lane, E P Lim, N Kalyanaraman, J Nemesh, L Ziaugra, L Friedland, A Rolfe, J Warrington, R Lipshutz, G Q Daley, and E S Lander. Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nature genetics*, 22(3):231–8, July 1999.

[5] M J Daly, J D Rioux, S F Schaffner, T J Hudson, and E S Lander. High-resolution haplotype structure in the human genome. *Nature genetics*, 29(2):229–32, October 2001.

[6] Morris W Foster and Richard R Sharp. Race, ethnicity, and genomics: social classifications as proxies of biological heterogeneity. *Genome research*, 12(6):844–50, June 2002.

[7] U Goldbourt and H N Neufeld. Genetic aspects of arteriosclerosis. *Arteriosclerosis (Dallas, Tex.)*, 6(4):357–77, 1986.

[8] Sara Hägg, Josefin Skogsberg, Jesper Lundström, Peri Noori, Roland Nilsson, Hua Zhong, Shohreh Maleki, Ming-Mei Shang, Björn Brinne, Maria Bradshaw, Vladimir B Bajic, Ann Samneg, Angela Silveira, Lee M Kaplan, Bruna Gigante, Karin Leander, Ulf de Faire, Stefan Rosfors,

Ulf Lockowandt, Jan Liska, Peter Konrad, Rabbe Takolander, Anders Franco-Cereceda, Eric E Schadt, Torbjörn Ivert, Anders Hamsten, Jesper Tegnér, and Johan Björkegren. Multi-organ expression profiling uncovers a gene module in coronary artery disease involving transendothelial migration of leukocytes and LIM domain binding 2: the Stockholm Atherosclerosis Gene Expression (STAGE) study. *PLoS genetics*, 5(12):e1000754, December 2009.

[9] Lucia A Hindorff, Praveen Sethupathy, Heather A Junkins, Erin M Ramos, Jayashri P Mehta, Francis S Collins, and Teri A Manolio. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences of the United States of America*, 106(23):9362–7, June 2009.

[10] Hindorff LA, MacArthur J (European Bioinformatics Institute), Wise A, Junkins HA, Hall PN, Klemm AK and Manolio TA. A Catalog of Published Genome-Wide Association Studies. Available at: www.genome.gov/gwastudies. Accessed January 1, 2012.

[11] L.B. Jorde. Linkage Disequilibrium and the Search for Complex Disease Genes. *Genome Research*, 10(10):1435–1444, October 2000.

[12] Robert J Klein, Caroline Zeiss, Emily Y Chew, Jen-Yue Tsai, Richard S Sackler, Chad Haynes, Alice K Henning, John Paul SanGiovanni, Shrikant M Mane, Susan T Mayne, Michael B Bracken, Frederick L Ferris, Jurg Ott, Colin Barnstable, and Josephine Hoh. Complement factor H polymorphism in age-related macular degeneration. *Science (New York, N.Y.)*, 308(5720):385–9, April 2005.

[13] L Kruglyak and D A Nickerson. Variation is the spice of life. *Nature genetics*, 27(3):234–6, March 2001.

[14] Chee Seng Ku, En Yun Loy, Yudi Pawitan, and Kee Seng Chia. The pursuit of genome-wide association studies: where are we now? *Journal of human genetics*, 55(4):195–206, April 2010.

[15] W H Li and L A Sadler. Low nucleotide diversity in man. *Genetics*, 129(2):513–23, October 1991.

[16] A J Lusis. Atherosclerosis. *Nature*, 407(6801):233–41, September 2000.

[17] Colin D Mathers and Dejan Loncar. Projections of global mortality and burden of disease from 2002 to 2030. *PLoS medicine*, 3(11):e442, November 2006.

[18] Myocardial Infarction Genetics Consortium. Genome-wide association of early-onset myocardial infarction with single nucleotide polymorphisms and copy number variants. *Nature genetics*, 41(3):334–41, March 2009.

[19] Svante Pääbo. The mosaic that is our genome. *Nature*, 421(6921):409–12, January 2003.

[20] Michael Preuss, Inke R König, John R Thompson, Jeanette Erdmann, Devin Absher, Themistocles L Assimes, Stefan Blankenberg, Eric Boerwinkle, Li Chen, L Adrienne Cupples, Alistair S Hall, Eran Halperin, Christian Hengstenberg, Hilma Holm, Reijo Laaksonen, Mingyao Li, Winfried März, Ruth McPherson, Kiran Musunuru, Christopher P Nelson, Mary Susan Burnett, Stephen E Epstein, Christopher J O'Donnell, Thomas Quertermous, Daniel J Rader, Robert Roberts, Arne Schillert, Kari Stefansson, Alexandre F R Stewart, Gudmar Thorleifsson, Benjamin F Voight, George A Wells, Andreas Ziegler, Sekar Kathiresan, Muredach P Reilly, Nilesh J Samani, and Heribert Schunkert. Design of the Coronary ARtery DIsease Genome-Wide Replication And Meta-Analysis (CARDIoGRAM) Study: A Genome-wide association meta-analysis involving more than 22 000 cases and 60 000 controls. *Circulation. Cardiovascular genetics*, 3(5):475–83, October 2010.

[21] Shaun Purcell, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel A R Ferreira, David Bender, Julian Maller, Pamela Sklar, Paul I W de Bakker, Mark J Daly, and Pak C Sham. PLINK: a tool set for whole-genome association and population-based linkage analyses. *American journal of human genetics*, 81(3):559–75, September 2007.

[22] Jiannis Ragoussis. Genotyping technologies for genetic research. *Annual review of genomics and human genetics*, 10:117–33, January 2009.

[23] Stephen A Ramsey, Elizabeth S Gold, and Alan Aderem. A systems biology approach to understanding atherosclerosis. *EMBO molecular medicine*, 2(3):79–89, March 2010.

[24] D E Reich, M Cargill, S Bolk, J Ireland, P C Sabeti, D J Richter, T Lavery, R Kouyoumjian, S F Farhadian, R Ward, and E S Lander. Linkage disequilibrium in the human genome. *Nature*, 411(6834):199–204, May 2001.

[25] David E Reich, Stacey B Gabriel, and David Altshuler. Quality and completeness of SNP databases. *Nature genetics*, 33(4):457–8, April 2003.

[26] Eric E Schadt and Johan L M Björkegren. NEW: Network-Enabled Wisdom in Biology, Medicine, and Health Care. *Science translational medicine*, 4(115):115rv1, January 2012.

[27] Eric E Schadt, John Lamb, Xia Yang, Jun Zhu, Steve Edwards, Debraj Guhathakurta, Solveig K Sieberts, Stephanie Monks, Marc Reitman, Chunsheng Zhang, Pek Yee Lum, Amy Leonardson, Rolf Thieringer, Joseph M Metzger, Liming Yang, John Castle, Haoyuan Zhu, Shera F Kash, Thomas A Drake, Alan Sachs, and Aldons J Lusis. An integrative genomics approach to infer causal associations between gene expression and disease. *Nature genetics*, 37(7):710–7, July 2005.

[28] Eric E Schadt, Cliona Molony, Eugene Chudin, Ke Hao, Xia Yang, Pek Y Lum, Andrew Kasarskis, Bin Zhang, Susanna Wang, Christine Suver, Jun Zhu, Joshua Millstein, Solveig Sieberts, John Lamb, Debraj GuhaThakurta, Jonathan Derry, John D Storey, Iliana Avila-Campillo, Mark J Kruger, Jason M Johnson, Carol A Rohl, Atila Van Nas, Margarete Mehrabian, Thomas A Drake, Aldons J Lusis, Ryan C Smith, F Peter Guengerich, Stephen C Strom, Erin Schuetz, Thomas H Rushmore, and Roger Ulrich. Mapping the genetic architecture of gene expression in human liver. *PLoS Biology*, 6(5):13, May 2008.

[29] Nicholas J Schork, Sarah S Murray, Kelly A Frazer, and Eric J Topol. Common vs. rare allele hypotheses for complex diseases. *Current opinion in genetics & development*, 19(3):212–9, June 2009.

[30] Jesper Tegnér and Johan Björkegren. Perturbations to uncover gene networks. *Trends in genetics : TIG*, 23(1):34–41, January 2007.

[31] Jesper Tegnér, Josefin Skogsberg, and Johan Björkegren. Thematic review series: systems biology approaches to metabolic and cardiovascular disorders. Multi-organ whole-genome measurements and reverse engineering to uncover gene networks underlying complex traits. *Journal of lipid research*, 48(2):267–77, February 2007.

[32] The International HapMap 3 Consortium. Integrating common and rare genetic variation in diverse human populations. *Nature*, 467(7311):52–8, September 2010.

[33] The International HapMap Consortium. The International HapMap Project. *Nature*, 426(6968):789–796, December 2003.

[34] The International HapMap Consortium. A haplotype map of the human genome. *Nature*, 437(7063):1299–320, October 2005.

[35] The International HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. *Nature*, 449(7164):851–61, October 2007.

[36] The Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145):661–78, 2007.

[37] D G Wang, J B Fan, C J Siao, A Berno, P Young, R Sapolsky, G Ghandour, N Perkins, E Winchester, J Spencer, L Kruglyak, L Stein, L Hsie, T Topaloglou, E Hubbell, E Robinson, M Mittmann, M S Morris, N Shen, D Kilburn, J Rioux, C Nusbaum, S Rozen, T J Hudson, R Lipshutz, M Chee, and E S Lander. Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science (New York, N.Y.)*, 280(5366):1077–82, May 1998.

[38] Hua Zhong, John Beaulaurier, Pek Yee Lum, Cliona Molony, Xia Yang, Douglas J Macneil, Drew T Weingarth, Bin Zhang, Danielle Greenawalt, Radu Dobrin, Ke Hao, Sangsoon Woo, Christine Fabre-Suver, Su Qian, Michael R Tota, Mark P Keller, Christina M Kendziorski, Brian S Yandell, Victor Castro, Alan D Attie, Lee M Kaplan, and Eric E Schadt. Liver and adipose expression associated SNPs are enriched for association to type 2 diabetes. *PLoS genetics*, 6(5):e1000932, May 2010.

[39] Hua Zhong, Xia Yang, Lee M Kaplan, Cliona Molony, and Eric E Schadt. Integrating pathway analysis and genetics of gene expression for genome-wide association studies. *American journal of human genetics*, 86(4):581–91, April 2010.

www.kth.se