Fakultät für Medizin

Deutsches Herzzentrum München - Klinik an der Technischen Universität München

Understanding the genetics of coronary artery disease through novel statistical approaches

Lingyao Zeng

Vollständiger Abdruck der von der Fakultät für Medizin der Technischen Universität München zur Erlangung des akademischen Grades eines

Doctor of Philosophy (Ph.D.)

genehmigten Dissertation.

Vorsitzende/r: Prof. Dr. Dr. Stefan Engelhardt

Betreuer/in: Prof. Dr. Adnan Kastrati

Prüfer der Dissertation:

1. Prof. Dr. Heribert Schunkert

2. Prof. Dr. Dr. Fabian Theis

Die Dissertation wurde am 29.07.2016 bei der Fakultät für Medizin der Technischen Universität München eingereicht und durch die Fakultät für Medizin am 14.09.2016 angenommen.

Acknowledgement

First of all, I would like to express my sincere gratitude and thanks to my advisor Prof. Heribert Schunkert. It is my honor to join his research group. He has been actively interested in my work and offered tremendous support and guidance at all levels of my PhD research over the last few years. I greatly appreciate all his knowledge, kindness, enthusiasm, and encouragement.

I would like to thank the other members of my committee, Prof. Adnan Kastrati and Prof. Fabian Theis. I thank Prof. Adnan Kastrati, who as my supervisor has given me the opportunity to pursue my PhD research at the Deutsches Herzzentrum München and Technische Universität München. I also thank him for the generosity and patience offering the clinical data and descriptions. I thank Prof. Fabian Theis for all his valuable comments and encouragement at both the committee meetings and e:AtheroSysMed project conferences.

I am grateful to Prof. Bertram Müller-Myhsok from the Department of Statistical Genetics, Max Planck Institute of Psychiatry, and his team member Dr. Nazanin Mirza-Schreiber, for their collaborative efforts in the epistasis project. We have spent much time together in co-running the analysis, trouble shooting, and stimulating discussions. Also many thanks to Dr. Till Andlauer who has offered us his scripts for a scrutinized genotype QC and imputation procedure.

My sincere thanks also goes to Prof. Jeanette Erdmann, who has warmly hosted me at Institut für Integrative und Experimentelle Genomik, Universität zu Lübeck, in my

early days of PhD research, and to Dr. Christina Willenborg in her team for her help with all data issues. I wish to thank Prof. Inke König for well hosting me at Institut für Medizinische Biometrie und Statistik, Universität zu Lübeck, in my early days of PhD research as well, and Dr. Christina Loley in her team for the patient instructions on basic data processing.

I specially thank my colleague Dr. Thorsten Keßler for his careful reviewing this thesis and offering comments and advices. Special thanks also goes to my colleague Dr. Barbara Stiller, who has largely helped in proof-reading and offered great suggestions.

I also thank all my other colleagues in AG Schunkert for contributing to the nice working atmosphere, stimulative discussion and supportive collaboration.

Grateful thanks to all my friends for always supporting me and for all the good times we have had.

Finally, I would like to sincerely thank my parents who have given me inspiration, encourage, and enlightenment over all years of my life.

Abstract

Coronary artery disease (CAD) is a complex disease with a strong genetic component. A better understanding of the molecular-genetic basis of CAD susceptibility is of vital importance for better prevention and treatment of this multifactorial and often lethal disease. The study of human genomes on a genome-wide scale may help to uncover novel genes involved in the pathophysiology and therefore broaden the knowledge from a systematic perspective. With the development of genotyping techniques and the progress of the Human 1000 Genome Project (1000G), quantification of variations on a genome-wide scale became feasible, which not only provides a powerful molecular repository for identifying complex genetic architectures of multifactorial diseases but also presents unprecedented data analysis challenges.

The aim of this thesis was to assess the possible genetic impact on the susceptibility of CAD using quantitative and statistical genetics research approaches for three aspects: 1. to conduct genome-wide association studies (GWAS) for CAD in the 1000G era; 2. to contribute to the understanding of the genetic complexity of GWAS signals of CAD; and 3. to detect whether CAD risk is modified by the interaction of different variants.

The first part of my research included GWAS analyses with the aim to uncover additional risk loci in the 1000G era. Specifically, I performed GWAS analyses after 1000G imputation on four German cohort studies based on the whole autosomal genome (1000G CAD GWAS), and also on one German cohort focused on the X-chromosome (1000G CAD X-Chr). The resultant summary statistics of these analyses served as

valuable components to the CARDIoGRAMplusC4D consortium, which has organized international collaborative efforts for the meta-analysis of genome-wide associations studies. So far one such meta-analysis of 1000G CAD GWAS, which was finally based on more than 185,000 CAD cases and controls and identified 10 novel CAD-associated loci, has been published. Further publications are under review.

Typically, only the best SNP at each locus from GWAS is reported. However, as typical for a complex disease, the genetic architecture of CAD is not monogenic. On one hand, multiple genes/loci are contributing to the CAD susceptibility. Within each of them there may also exist allelic heterogeneity. On the other hand, the genetic susceptibility loci to other complex traits may also contribute to CAD risk, due to a potentially shared genetic and functional background. The second part of my thesis was therefore to investigate the genetic complexity of GWAS signals of CAD, with the aim to examine and confirm - based on the individual-level genotype data - both the intra-locus allelic heterogeneity within known CAD loci and the multi-locus polygenic pleiotropic effect from susceptible loci of other traits. At known CAD loci, I collected individual-level genotype data from eight cohorts including over 10,000 cases and 10,000 controls, and examined the additive effect of multiple alleles, which were found to be independently associated with CAD risk at the respective locus based on a multi-locus polygenic score (PGS) approach. Indeed, at some loci multiple independent signals could be recovered with a combined effect that conferred incremental risk of CAD with the increase of the number of independent risk alleles. The results improved our understanding of the allelic structure at known CAD-associated loci, and also highlighted the importance and complexity of genetic heterogeneity. To investigate the potential pleiotropic effect of the genetic susceptible loci of other traits on CAD, I further constructed multi-locus polygenic scores for height and rheumatoid arthritis (RA) based on five and seven individual-level genotype datasets, respectively, and examined their effect on CAD onset. The results helped to support the notion that height directly and indirectly affects CAD

risk, as well as that genetic factors underlying RA carry a low likelihood to affect CAD risk. My analysis contributed to the respective publications.

GWAS are usually analyzed with the assumption that the genetic variants involved in a complex disease act independently and that their combined single effects are responsible for the observed phenotypes. It has now been accepted that epistatic effects, i.e., gene-gene interactions, may also play a significant role in determining complex traits. Limited by methodological issues, GWAS signals usually do not cover the higher order genetic architecture underlying CAD risk. The third part of my thesis was thus dedicated to detect gene-gene interactions. So far no large-scale systematic investigation of epistasis had been made in the context of CAD, mainly due to the challenge in both computation power and sample size. To enable such analysis, I collected individual-level genotype data of nine cohorts including 27,360 individuals, and, in collaboration with Max Planck Institute of Psychiatry, implemented the computation with a powerful GPU-based parallel computing tool. By strategy, I started with a searching space of broad sense CAD susceptibility regions, and investigated two-variant statistical epistasis assuming all possible genetic models. For each statistically significant epistasis pair, I subsequently searched for potential biological epistasis. Finally, I postulated a novel hypothesis on how genetic loci could convey their epistatic effect, firstly through perturbation of nuclear protein interactions, and secondly through perturbation of downstream pathways. These epistasis results make a great extension to our current knowledge of CAD genetics. The applied scheme and the GPU-based parallel computing tool may also enable researchers to further explore CAD epistasis at a genome-wide scale in the future.

In summary, the studies in my thesis, including statistical genetics approaches such as GWAS analyses, polygenic score calculation, and epistasis investigation, made efforts and contributions to the improvement of our understanding of the genetic etiology of CAD from several perspectives.

Contents

List of Abbreviations

1000G	The 1000 Genomes Project
ANOVA	Analysis of Variance
AUC	Area Under the Curve
CAD	Coronary Artery Disease
CI	Confidence Interval
EAF	Effect Allele Frequency
eQTL	Expression Quantitative Trait Loci
FDR	False Discovery Rate
GPU	Graphic Processing Unit
GWAS	Genome-wide Association Study
HapMap	The HapMap Project
HWE	Hardy-Weinberg Equilibrium
IBD	Identity by Desent
LD	Linkage Disequilibrium

MAF	Minor-Allele Frequency
MDS	Multidimensional Scaling
MI	Myocardial Infarction
NL	Nuclear Lamina
OR	Odds Ratio
PCA	Principal Component Analysis
PGS	Polygenic Score
QC	Quality Control
RA	Rheumatoid Arthritis
ROC	Receiver Operator Characteristic
SD	Standard Deviation
SNP	Single Nucleotide Polymorphism
TF	Transcriptional Factor
TSS	Transcription Start Site

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background of statistical genetics research in CAD

1.1.1 CAD is a complex disease with a strong genetic component

Coronary artery disease (CAD) is the most common cause of death globally [12]. The clinical presentation ranges from asymptomatic lesions in coronary vessels to ischemic symptoms such as angina pectoris or myocardial infarction, which causes myocardial damage and/or premature death. The anatomical underpinnings of CAD are atherosclerotic plaques developing throughout life which may finally lead to narrowing of coronary arteries or atherothrombosis.

CAD is considered as a semi-inherited disease. According to a large epidemiological survey 35% of the CAD patients have a positive family history [67], which suggests a strong genetic component in the etiology of the disease. Besides genetics, risk factors including hypertension, hyperlipidemia, smoking, obesity, diabetes, and sedentary life style modify the risk of CAD [69].

CAD is rather a complex than a monogenic disease. The high prevalence of patients with a positive family history suggests the sharing of numerous susceptibility genes.

Indeed, several genetic variants have been identified to increase the susceptibility of the disease suggesting a polygenic inheritance [99]. To the current knowledge, some CAD susceptibility loci show allelic heterogeneity or copy number variation. Some of the susceptibility loci also imply risk to other traits related to the disease pathogenesis [80, 102].

In order to improve prevention and treatment of CAD, it is of high importance to understand the multifactorial genetic basis of disease susceptibility.

1.1.2 Human genetic variations and 1000G project

Human genetic variations refer to genetic differences both within and among populations. There may be multiple variants of any given gene in a human population. Such polymorphisms may occur in the coding region – potentially affecting protein structure and function – or in regulatory regions. However, the vast majority of genetic variants may be functionally neutral. The study of human genetic variation is of significance for both, evolutionary and medical research, e.g., some disease-causing alleles occur more often in people from specific geographic regions [106].

A single nucleotide polymorphism (SNP) is a variation in a single nucleotide between members of one species. With the development of high-throughput genotyping techniques, great interest arose to perform case-control studies in order to detect the genetic variation underlying risk of diseases. These studies focused on the investigation of SNPs, as they can be easily detected in millions of people scattered throughout the whole human genome. Recent research on population genetics focused on SNPs that occur in at least 1% of the population, given that the statistical power to detect effects on disease risk is dependent on the distribution of alleles. Currently 10 to 30 million human SNPs have been explored in large genome-wide association studies (GWAS) [13].

The HapMap project, officially started in 2002, was organized as a global collaboration among researchers that aimed to develop a haplotype map (HapMap) to describe

the common patterns of human genetic variation [76]. This common human sequence variation database was a major contribution to facilitate large-scale association studies between genetic variation and human disease [107]. Out of this effort, genotypes at several hundred thousand chromosomal sites, combined with the knowledge of LD structure, allow the investigation of the vast majority of common variants with at least 0.5% minor allele frequency (MAF).

The 1000 Genomes Project (1000G) is an international research effort in order to establish the most detailed catalogue of human genetic variation. It was launched in 2008 with the goal to find genetic variants with frequencies of at least 1% in human populations [33]. The aim of the 1000 Genomes Project is to discover, genotype and provide accurate haplotype information on all forms of human DNA polymorphism in multiple human populations. Specifically, the goal is to characterize over 95% of variants that are in genomic regions accessible to current high-throughput sequencing technologies in each of five major population groups (populations in or with ancestry from Europe, East Asia, South Asia, West Africa, and the Americas). Lower MAF variants (down towards 0.1%) are also catalogued [33]. The 1000G data provides us with the most complete catalogue of human DNA variation. By the time I started my thesis, the latest version - "1000 Genomes Phase I integrated variant (v3)" was released in NCBI build 37 (hg19) coordinates with reference data for 1,092 individuals [34].

The availability of these large-scale data opened up an exciting era of new opportunities in the field of human genetics. The high volume data produced by these new technologies give geneticists the opportunity to study the genome from a broader perspectives. To enable the usage of the wealth of genetic information in order to analyze and understand the nature of complex diseases, new and highly interdisciplinary fields such as bioinformatics, functional genomics, and systems biology have emerged with the aim to tackle different aspects of the exploration and analysis of these data.

1.1.3 Genotyping arrays

Due to the importance of common human genetic variations in modulating disease risk, researchers joined in a global team in order to identify a steadily increasing number of SNPs and to genotype these in individuals of various ancestries. The technical tools for this endeavor were built by *Affymetrix* and *Illumina* which developed high-throughput genotyping arrays based on the biochemical principle that nucleotide bases in DNA molecules bind to their complementary partners. The two companies designed different platforms in order to house hundred thousands of defined polymorphisms for genotyping. Specialized equipment can produce a measure of the signal intensity associated with each probe and its target after hybridization. The underlying principle is that the signal intensity depends upon the amount of target DNA in the sample, as well as the affinity between target and probe. The raw signal intensity is then converted to genotypes via computational algorithms provided by the manufacturers. The advent of such high-throughput technologies largely facilitated genomic research studies [55].

1.2 Genome-Wide Association Studies

1.2.1 What is a GWAS analysis?

Advances in genotyping technologies and the multitude of genetic data available provide a vast amount of data that enables a better understanding of human genetic diseases by studying genetic variation. This has particularly led to the development of "hypothesis-free" experimental approaches such as genome-wide association studies (GWAS). In fact, these studies test the hypothesis that common variation at any given chromosomal locus may affect the regulation or function of a gene with subsequent implications for disease risk. Mainly, the whole genome obtained from groups of affected (cases) and unaffected (control) individuals is analyzed for differences in genetic variation to detect

associations with the disease of interest in form of functionally relevant SNPs. Indeed, this technology has become the most effective approach for identifying genetic variants that are associated with complex disease risk.

An extension of this approach is to first conduct a GWAS on population-matched cases and controls using whole genome genotyping arrays and, once significant variants have been found, to proceed with targeted sequencing or fine-mapping of these regions to refine the location of the causal variant(s). This process allows researchers to gain the maximum benefit from both methods and reduce costs since sequencing technology is still not affordable for large-scale studies.

1.2.2 Previous achievements and open challenges

In the recent years important progress has been made in unraveling the genetics of CAD through the advancement in both technology and global collaboration. In 2007, three independent GWASs identified SNPs on chromosome 9p21 as associated with myocardial infarction (MI) and CAD [39, 68, 89]. Additional successful studies of individual teams were carried out in 2009 [28, 50, 96]. Since then, further studies were conducted by the CARDIoGRAM [90] and CARDIoGRAMplusC4D [16] consortia on a large population scale for meta-analyses of GWAS on CAD patients and controls. The CARDIoGRAM GWAS was a meta-analysis of 22 GWAS studies of European descent imputed to HapMap 2 involving 22,233 cases and 64,762 controls. The CARDIoGRAMplusC4D GWAS was a two-stage meta-analysis of Metabochip and GWAS studies of European and South Asian descent involving 63,746 cases and 130,681 controls. These studies have generated 46 genetic risk variants with confirmed and replicated association with CAD.

Regarding the allelic architecture of complex common disease, a ‘common disease–common variant’ hypothesis was originated in 1996 [56], stating that if a heritable disease or trait is common in the population (greater than 1-5 percent life-time prevalence), then the causative genetic factors are also likely to be common. The total genetic

risk must be spread across multiple common genetic factors with small effect. Indeed, so far the genetic risk variants identified through GWAS studies are usually common. However, ongoing analyses on the new 1000G reference genomes may help us to further differentiate this concept.

Another generic feature of the risk loci is that they confer moderate odds ratios, and thus only a small fraction of risk is conferred by a single locus as compared to the total disease heritability, which may be substantially larger. Since even the totality of currently identified SNPs does not fully explain disease risk a so-called "missing heritability issue" is being considered [64]. With the refined reference genome panels or even larger case-control samples, hopefully more causal genetic variations will be identified and thus the knowledge on heritability of CAD can be increased.

There are also open questions independent of the coverage of genotypes. Most of the risk loci identified so far are located at non-coding regions of the genome, which makes it difficult to understand their molecular or functional mechanisms.

By method recent meta-analyses of GWAS were based on the assumption that the genetic variants involved in most complex diseases are acting independently and that their additive single effects are responsible for the observed phenotypes. Importantly, this approach precluded detecting the higher order genetic architectures for complex diseases such as CAD.

1.3 Genetic complexity of GWAS signals of CAD

1.3.1 What is intra-locus allelic heterogeneity?

Allelic heterogeneity is the phenomenon that different variants at the same locus cause a same or similar phenotype. In the context of GWAS, this is referred to as the presence of multiple association signals at a single locus.

The primary analysis of GWAS is carried out under the assumption that there is only

a single common SNP or allele underlying the signal. However, it is highly possible that multiple independent effects could be present within a gene or gene locus that is associated with a trait. Indeed, several recent studies have reported the presence of multiple independent SNPs and genetic heterogeneity in trait-associated loci. For example, in early years, multiple independent effects were observed in 19 loci associated with human height [57] and in six loci associated with Crohn disease [32]. In 2012, Yang et al developed an approximate conditional and joint association approach based on GWAS summary statistics, which was able to detect multiple additional variants influencing complex traits such as height [110]. These additional variants at single loci could explain additional phenotypic variation of common traits, and therefore contribute as a part of narrow-sense heritability.

1.3.2 What is multi-locus polygenic pleiotropy?

Pleiotropy generally means that a gene or genetic variant affects more than one phenotypic trait. GWAS studies in the recent years have identified many variants that might affect multiple distinct traits [91]

Polygenic scores have often been used to summarize and combine genetic effects among an ensemble of markers into a single score, which could be taken as a surrogate estimate for the genetic susceptibility of the trait of interest. By associating the polygenic score of one trait to the susceptibility of another trait, the shared genetic etiology between traits could be deduced.

1.3.3 Previous achievements and open challenges

In the process of my thesis work, the latest large-scale GWAS meta-analysis of CAD [77] based on 1000G imputation got published. Ten novel chromosomal regions were identified as conferring CAD risk, increasing the to-date list of known CAD loci to 56.

Intra-locus allelic heterogeneity is observed in some 1000G GWAS signals [77]. The authors of this study performed an approximate conditional and joint association analysis using GCTA software [110] based on the surroundings of variants that showed suggestive additive association ($p < 5 \times 10^{-5}$), which brought us to 202 variants with a low false discovery rate (FDR) ($q < 0.05$) at 129 loci. Ninety-five variants (explaining $13.3 \pm 0.4\%$ of CAD heritability) mapped to 44 significant CAD loci from GWAS. Thus, GWAS and fine-mapping of GWAS signals have made large progress in explaining the heritability of various diseases or traits including CAD. However several areas of uncertainty remain.

For example, recent evidence suggests the existence of intra-locus allelic heterogeneity at known CAD loci. However, the exact additive effects at each of these loci harboring multiple signals on the individual-level data is still unclear, which caught our interest for further exploration. Furthermore, it is also interesting to examine the extent by which these multiple independently acting variants increase the accuracy of CAD risk prediction. Moreover, it remains unclear how pleiotropy affects the polygenic inheritance patterns of CAD and other traits that are correlated with CAD.

1.4 Epistasis

1.4.1 What is epistasis?

Epistasis, also known as gene-gene interactions, has long been recognized as fundamentally important to understanding the structure and function of genetic pathways and the evolutionary dynamics of complex genetic systems [81]. Generally there two different definitions of epistasis: biological epistasis and statistical epistasis.

Biological epistasis

Early in 1909, William Bateson described the ”masking effect” where a variant or allele prevents another one from manifesting its effect in a similar manner. It is now more commonly defined as interactions between genes at different loci, where the effects of an allele at one locus are masked by an allele at another locus [17].

Biological epistasis has often been presumed to be ubiquitous given that many genes interact in complex ways with other genes [88]. It also seems to have an evolutionary role like maintaining deleterious variants under selection [40]. Biological epistasis can occur in several ways: through the interaction between transcription factors or promoter sequences, or through enzymes in a metabolic pathway [71]. The effects of epistasis can be beneficial or harmful in varying degrees. Epistasis is very common in nature and a key in characterizing the genetic basis of complex diseases. In model organisms, or agricultural crop, epistasis effects can be studied directly by genetic crosses, transgenes or DNA editing, or by other experimental approaches [88].

Statistical epistasis

Another definition of epistasis was proposed by Fisher and has been called "statistical interaction" [17]. Fisher describes the statistical interaction basically as a deviation from additivity in a linear model in the effect of alleles at different loci with respect to their contribution to a quantitative trait of a population. Specifically, he partitioned the genetic variance of a trait into several different components by fitting an additive model of genotypic values. The variance due to the deviations of each multilocus genotype from the additive model is the epistatic or interaction variance, which has nowadays been called statistical epistasis [88].

This definition is not equivalent to Bateson’s 1909 definition, however, *epistasis* used in the Fisher sense is closer to the usual concept of statistical interaction: departure from a specific linear model describing the relationship between predictive factors [17]. In

human genetics, direct studies on biological epistasis with transgenes or DNA editing, or by other experimental approaches, are rarely possible [88]. Detecting statistical epistasis in human populations is helpful to facilitate our understanding in genetic variations and inferences about biological epistasis.

1.4.2 Previous achievements and open challenges

Identifying epistatic interactions statistically in human genetics is a challenging issue. The technical difficulties involve – not exclusively – whether the statistical model is proper, whether the sample size is large enough, whether the significance level is properly settled, and whether the computation capacity is sufficient. To test the two SNPs epistasis of a set of N SNPs, the number of SNP combinations would be $M = N(N-1)/2$ [63].

Several approaches have been used to study epistatic interactions, including regression models, exhaustive search for two-locus versus high-order interactions, data-mining/machine learning and related approaches, Bayesian model selection approaches, etc [18]. Nevertheless, the overall success of empirical studies of epistasis in humans is unclear, as the reproducibility of gene–gene interactions is low [73]. Besides, statistical interaction does not necessarily imply interaction at a biological level [18].

Moreover, the investigation of interactions in recent genome-wide association studies has been restricted due to computation issues. So far, prior biological knowledge has been applied (e.g., known genes of interest, protein-protein interactions, pathways) by researchers to narrow the search space for potential epistasis effect in order to identify epistasis for different diseases or traits such as Behcet's disease, lipids, body-mass index, breast cancer. However, some of these investigations were either based on limited sample size ($N < 10000$), or were not able to identify epistatic SNP-pairs at a proper significance level [62, 70, 103].

Few studies have examined epistasis in CAD. When I started the work on my thesis, there was only one reported effort regarding exploration of epistasis at a systematic level.

In 2012, Lucas et al. performed a hypothesis-based analysis of gene-gene interactions and risk of myocardial infarction. They have focused their analysis on a set of 242 SNPs, which were selected based on prior knowledge to have suggestive implications to cardiovascular disease. A sample of 2,967 cases and 3,075 controls was taken into discovery phase and a sample of 1,766 cases of coronary heart disease and 2,938 controls was used as replication. However, no statistically significant interaction pairs were identified [61].

In the process of my thesis work, there has been recent progress in exploration of gene-gene interactions in relation to CAD. Musameh et al [74] studied a set of 913 common variants in candidate cardiovascular genes. Following the discover by a sample of 2,101 patients with CAD and 2,426 controls, they studied 2,967 patients and 3,075 controls as replication cohort. However, none of the interactions was statistically significant after correction for multiple testing, nor was a secondary exploratory interaction analysis on 11,332 independent common SNPs surviving quality control.

Turner et al [97] run a statistical epistatic analysis between variants in two genes, *SMAD3* and *COL4A2*, for which they have previously noticed that the upregulation of *COL4A1/2* is dependent on *SMAD3* in a TGF-β signaling pathway. Based on a meta-analysis of 5 cohorts with 4,956 cases and 2,774 controls, the most significant interaction was identified at a Bonferroni-corrected p-value 6.9×10^{-3}.

1.5 Specific aims

1.5.1 Aims of research

Following the considerations discussed above, the main purpose of my PhD project was to gain insight into the genetic etiology of CAD by innovative approaches of statistical genetics. The rationale is to make the most out of existing GWAS data and samples using different strategies. Specifically, my project is designed to address three issues:

1. Genome-wide association studies for CAD in the 1000G era
    - to improve the understanding of autosomal susceptible CAD loci based on traditional genotyping arrays
    - to investigate possible susceptible CAD loci on the X-chromosome
2. Understanding the genetic complexity of GWAS signals of CAD
    - to gain insights about allelic heterogeneity of the biological mechanisms underlying known CAD loci
    - to test and confirm the additive effects of multiple independent genetic variants within known CAD loci based on the individual-level genotype data
    - to investigate the relationship between genetic susceptible loci of other trait and the risk of CAD
3. Detecting epistasis that underlies CAD
    - to establish the feasible data-driven workflow for statistical epistasis identification for CAD
    - to explore the statistical epistasis effect in CAD
    - to explore the biological interpretation for the statistical epistasis effect in CAD

1.5.2 Structure of contents

In the following chapters of this dissertation I will address the above three issues in successive chapters by a subset of issues relating to it.

Chapter 2 describes the materials and methods that I have used for research, which starts with a section for data description, a section for general methods, and follows with specific methods regarding three parts of research respectively. Chapter 3 describes

the results that I have generated in respect to the three parts of issues sequentially. In Chapter 4 I will discuss the results from Chapter 3 in the same order. In Chapter 5 I will summarize the entire project, including all the findings of my research, and provide the outlook on future research.

Chapter 2

Methods

2.1 Cohort description

To achieve the objectives of my thesis, data from multiple CAD case-control cohorts have been collected and used for analysis.

Briefly, individual level data were obtained from the German Myocardial Infarction Family Studies (GerMIFS) I, II, III (KORA), IV, and V [16, 28, 51, 89, 90, 95], the Ludwigshafen Risk and Cardiovascular Heath Study (LURIC)/AtheroRemo [28, 90, 108], Myocardial Infarction Genetics Consortium (MIGen) [28, 90], the Cardiogenics Consortium [28, 89], and the Wellcome Trust Case Control Consortium (WTCCC) [89, 90, 104]. All subjects in all studies were Caucasians of European origin and gave written informed consent before participating. All studies were approved by their local Ethical Committees.

Genome-wide genotype data and associated phenotype data for GerMIFS I-V were generated by our group. Data for MIGen were obtained via the database of Genotypes And Phenotypes [20] (project number 6271). Data for LURIC were obtained via the eAtheroSysMed consortium [24]. Data for WTCCC and Cardiogenics were obtained via the Leducq network for CAD genomcis [78].

GerMIFS-I. *Cases:* All patients were Caucasians of German descent and selected from those who had suffered MI prior to the age of 60 years. The majority (>70%) was recruited in the vicinity of Augsburg and the southern part of Germany (Clinics in Starnberg/Höhenried, Prien) in the years 1997-2002. If at least one additional first-degree family member (preferentially a sibling) had suffered from MI or had severe coronary artery disease (percutaneous transluminal coronary angioplasty [PTCA] or bypass surgery [CABG]), the family (index patient, available parents and all siblings) was contacted and invited to participate in the study.*Controls:* Population-based subjects from the same area.

GerMIFS-II. *Cases:* Cases were identified following their admission for acute treatment of MI or in cardiac rehabilitation clinics, with a validated history of MI before the age of 60 years for both men and women. A positive family history for CAD was documented in 59.4% of individuals. *Controls:* Population-based subjects.

GerMIFS-III. *Cases:* Non fatal MI in the KORA registry with DNA available. Hospitalized survivors of MI who are 26 to 74 years of age are routinely entered into this registry. *Controls:* Population-based subjects from Augsburg KORA S4/F4 study and PopGen.

GerMIFS-IV. *Cases:* Consecutive patients referred for coronary angiography, classified as CAD or MI cases based on the coronary angiogram (at least a 50% stenosis in one major coronary vessel) and age of onset (<65 years in males, and <70 years in females). *Controls:* Population-based subjects as part of the Berlin Aging Study II (BASE-II).

GerMIFS-V. *Cases:* Participants of the Munich MI sample included in this study were consecutively recruited from 1993 to 2002 and examined with coronary angiography at Deutsches Herzzentrum Muenchen and 1. Medizinische Klinik rechts der Isar der Technischen Universitaet Muenchen. The diagnosis of MI was established in the presence of chest pain lasting >20 minutes combined with ST-segment elevation or pathological Q waves on a surface electrocardiogram. Patients with MI had to show

either an angiographically occluded infarct-related artery or regional wall motion abnormalities corresponding to the electrocardiographic infarct localization, or both. *Controls:* Population-based subjects.

LURIC. *Cases:* Cases were included as white patients hospitalized for coronary angiography between June 1997 and May 2001, with angiographically confirmed CAD (at least one coronary vessel with a stenosis > 50%) *Controls:* controls were from the German Blood Service (GerBS) control series that consists of healthy, unrelated blood donors. They were recruited from the southwestern area of Germany which corresponds to the geographical origin of the LURIC patients.

Cardiogenics. *Cases:* Cases from Germany and England were under the age of 65 with a confirmed primary MI within the preceding 3-36 months. Exclusion criteria were (i) a history of diabetes mellitus based on plasma glucose >7.0 mmol/l or HbA1C > 7.0, (ii) renal insufficiency, (iii) statin therapy, (iv) CRP level >10mg/dl, (v) fasting at the time of blood sampling or (vi) smokers. The Paris cohort comprised patients aged 33 to 87, recruited within the BAAAC study with symptoms of acute coronary syndrome who had one stenosis >50% diagnosed in at least one major coronary artery. *Controls:* Population-based subjects who were blood donors (aged 32 to 65 years) recruited as part of the Cambridge Bioresource in Cambridge.

WTCCC. *Cases:* Cases were recruited as part of the British Heart Foundation Family Heart Study and comprised subjects of European ancestry with a validated history of myocardial infarction or coronary revascularisation (PTCA or CABG) before their 66th birthday as well as a strong familial basis of CAD. Verification of the history of CAD was required either from hospital or primary care records. *Controls:* Controls comprised an equal number of subjects from the 1958 Birth Cohort and from blood donors recruited through the UK National Blood Service as part of the WTCCC study.

MIGen. *Cases:* A collection of early-onset myocardial infarction (in men less than or equal to 50 years old or women less than or equal to 60 years old) from six international

sites - Boston and Seattle in the United States as well as Sweden, Finland, Spain and Italy. MI was diagnosed on the basis of autopsy evidence of fatal MI or a combination of chest pain, electrocardiographic evidence of MI, or elevation of one or more cardiac biomarkers (creatine kinase or cardiac troponin). *Controls:* Population-based subjects.

A brief summary of individual statistics is shown in Table 2.1, and more detailed cohort descriptions could be sourced from the corresponding references.

2.2 General methods for genotyping array analyses

2.2.1 Quality control

Quality control (QC), that is, to check and clean the raw genotyping data, is always important prior to all analyses based on genotyping array, in order to reduce false positives and identify the true association in further analyses. The following criteria [3, 109, 116] were applied generally to the genotype level QC for all cohort studies.

Call rate

The missing call rate of a variant (genotype-level) is the proportion of individuals whose genotypes are not called for a given variant. The missing call rate of an individual is the proportion of variants whose genotypes are not called for a given individual. Low call rate may be an implication of poor DNA sample quality. Individuals with genotype-level call rate less than 98% in either cases or controls are filtered out. Likewise, variants with individual-level call rate less than 95% in either cases or controls are filtered out.

Minor-Allele Frequency

The current genotyping technology is still error-prone to detect loci with minor-allele frequency (MAF) less than 1%. Variants with MAF<0.01 were thus filtered out.

Cohort Name (Abbr.)	Platform	N Cases (CAD)	N Controls	Female N Cases (%)	Female N Controls (%)	Reference
GerMIFS-I (G1)	Affymetrix Mapping 500K Array Set	634	1608	211 (33.3)	817 (50.8)	[16, 28, 51, 89, 90]
GerMIFS-II (G2)	Affymetrix Genome-Wide Human SNP Array 6.0	1207	1288	246 (20.4)	618 (48.0)	
GerMIFS-III (G3)	Affymetrix Genome-Wide Human SNP Array 5.0/ Affymetrix Genome-Wide Human SNP Array 6.0	1060	1467	214 (20.2)	710 (48.4)	
GerMIFS-IV (G4)	Affymetrix Genome-Wide Human SNP Array 6.0	998	1147	349 (35.0)	704 (61.4)	
GerMIFS-V (G5)	Illumina HumanOmniExpress Omniuni2.5 OmniExpress	2532	1639	640 (25.3)	864 (52.7)	[95]
LURIC (LU)	Affymatrix Genome-Wide Human SNP Array 6.0	2364	697	596 (25.2)	293 (42.0)	[28, 90, 108]
Cardiogenics (CG)	Illumina Human660W-Quad	392	410	51 (13.0)	242 (59.0)	[28, 89]
WTCCC (WT)	Affymetrix Mapping 500K Array Set	1988	3004	406 (20.4)	1532 (51.0)	[89, 90, 104]
MIGen (MG)	Affymatrix Genome-Wide Human SNP Array 6.0	2967	3075	663 (22.3)	751 (24.4)	[28, 90]

Table 2.1: An overview of study cohorts utilized in this thesis

Departure from Hardy-Weinberg Equilibrium

The assumption of the Hardy-Weinberg equilibrium (HWE) is that the parental alleles in a heterozygous offspring SNP genotype can be estimated. When the ratios of homozygous and heterozygous genotypes significantly differ from the prediction under HWE assumptions, it can indicate genotyping errors, batch effects or population stratification. Checking HWE only in controls is usually recommended, as deviation from HWE in cases could represent a signal of true association. Variants with a p-value for the HWE less than 1×10^{-6} in controls were filtered out.

Sex check

Discrepancy of the sex between reported from raw genotyping data or sample records and the estimated (based on actual X-chromosome genotypes) simply indicates a processing error. Individuals with discrepancies in sex check were filtered out.

Population stratification and Outliers

Population structure can cause spurious findings in further analysis. Principal component analysis (PCA) or multidimensional scaling (MDS) are the most popular methods to capture stratification. The basic idea is to capture the hidden ancestry genetic background by inferring continuous axes of variation from genotype data. These axes ("PC"s for PCA, or "dimensions" for MDS) are independent from each other and suggest the variation from different aspects. The top several continuous axes of variation are then used as covariates to correct for stratification in the association analysis.

Individuals with the top two PCs deviate by more than 5 SD from the mean were filtered out during genotype QC.

Identity by descent

Identity by descent (IBD) means the degree of recent shared ancestry for a pair of individuals. The expectation is that IBD = 1 for duplicates or monozygotic twins, IBD = 0.5 for first-degree relatives, IBD = 0.25 for second-degree relatives and IBD = 0.125 for third-degree relatives. Individuals from each pair with an IBD value ≥ 0.25 were filtered out.

Heterozygosity

Heterozygosity rate for a given individual is the proportion of heterozygous genotypes for a given individual. Large deviation in heterozygosity is an indication of low chip quality. An additional heterozygosity F statistic can be calculated with the form $|F| = (1 - O/E)$, where O is the observed proportion of heterozygous genotypes for a given sample and E is the expected proportion of heterozygous genotypes for a given sample based on the minor allele frequency across all non-missing SNPs for a given sample. Individuals with a heterozygosity rate that deviates by more than 3 SD from the mean were filtered out.

Summary of individuals

During my thesis research several different analyses were conducted for different purposes, for which the control for the genotype quality were also slightly different. At a minimum, all classic QC criteria (meeting the requirement for the 1000G GWAS analysis) were taken in the collected studies. The final number of individuals included in the actual analyses was dependent on the criteria of QC. A summary is shown in Table 2.2.

		G1		G2		G3		G4	
		cases	controls	cases	controls	cases	controls	cases	controls
unQCed		634	1608	1207	1288	1060	1467	998	1147
	Call rate > 0.95	-1							
	Sex check		-2			-2		-11	
during QC (1000G criteria)	Outlier			-3	-2			-24	-1
	IBD check (pihat<0.25)		-6		-1				
	Heterozygosity	-1	-7	-2	-13		-10	0	-8
classic QCed		632	1593	1202	1272	1058	1457	963	1138
	Call rate > 0.98	-3	-14		-4				
during QC (epistasis critera)	Heterozygosity*	-7	-7	-10	-5	-2	-4	-5	-2
	IBD check (pihat<0.125)		-21		-7	-1	-12	-4	
refined QCed		622	1551	1192	1256	1055	1441	954	1136
	IBD check (pihat<0.0625)		-34	-4	-24	-6	-18	-2	-4
final QCed		622	1517	1188	1232	1049	1423	952	1132

Continued on next page

Table 2.2 – *Continued from last page*

		G5		LU		MG		CG	
		cases	controls	cases	controls	cases	controls	cases	controls
unQCed		2532	1639	2364	697	2967	3075	392	410
	Call rate > 0.95	-1		-7					
	Sex check	-4	-2			-15	-9	-4	
during QC (1000G criteria)	Outlier	-57				-17	-19	-5	-2
	IBD check (pihat<0.25)	-5	-26	-36	-10				
	Heterozygosity	-6		-6	-6	-1	-4		-2
classic QCed		2459	1611	2315	681	2934	3043	383	406
	Call rate > 0.98		-6	-8	-12	-4	-2		
during QC (epistasis critera)	Heterozygosity*	-14	-2	-28	-11	-29	-22	-1	-1
	IBD check (pihat<0.125)	-8	-29	-29	-14		-1		-1
refined QCed		2437	1574	2250	644	2901	3018	382	404
	IBD check (pihat<0.0625)	-14	-30	-38	-25	-11	-29	-1	
final QCed		2423	1544	2212	619	2890	2989	381	404

Continued on next page

Table 2.2 – *Continued from last page*

		WT	
		cases	controls
unQCed		1988	3004
	Call rate > 0.95	-21	-10
	Sex check		
during QC (1000G criteria)	Outlier	-16	-21
	IBD check (pihat<0.25)		
	Heterozygosity	-5	-2
classic QCed		1946	2971
sample list update		-27	-42
basic Qced final		1919	2929
	Call rate > 0.98	-5	-1
during QC (epistasis critera)	Heterozygosity*	-5	-11
	IBD check (pihat<0.125)	-9	-6
refined QCed		1900	2911
	IBD check (pihat<0.0625)	-5	-13
final QCed		1895	2898

Table 2.2: Summary of number of individuals at different levels of QC

2.2.2 Imputation

The purpose of imputation procedure is to infer the genotypes that are not measured by genotyping arrays by referencing HapMap haplotypes, so that marker density in the final analysis is increased. That is, the unobserved genotypes in a set of study individuals are predicted using a set of reference haplotypes and genotypes from a genotyping array. Imputation also facilitates meta-analyses for results derived from different groups which had originally been genotyped on different array platforms.

1000G reference panel

The 1000 Genomes Phase I integrated variant (v3) set released in NCBI build 37 (hg19) coordinates with reference data from March 2012 (updated August 2012) was utilized as the reference panel for imputation in all analyses in my thesis work.

Pre-imputation

Before imputation but after QC, all variants were updated to the same genome build, from human genome build 36 to build 37 as the reference genome, with the help of UCSC liftOver [41]. This was followed by allele flipping, which aligns all variants to the same positive strand as indicated in the reference genome. This procedure was performed using PLINK [85].

Pre-phasing

Haplotypes were then pre-phased from genotypes. This was performed with SHAPEIT2 haplotype estimation tool [22], which generates the best guess haplotypes based on the given genotypes.

Imputation

Then the best guess haplotypes were forwarded to IMPUTE2 [42]. The whole genome was splitted into chunks of 5Mb first, to perform imputation. The splitted chunks were curated and optimized, when the resultant "concordance" criteria, which shows the concordance between imputed genotypes and original genotypes for one variant, were low. All imputed chunks were afterwards concatenated together.

For each bi-allelic variant [A/B] for each individual, the main output of IMPUTE2 reported the three genotypes AA, AB and BB in the form of their probabilities accounting for the genotype imputation uncertainty, instead of giving an fixed designation.

Post-imputation QC

The qualities of imputation were assessed based on quality data given by IMPUTE2 "INFO" metric, which usually ranges between 0 to 1 and represents the imputation certainty. Variants with INFO < 0.8 in either cases or controls were filtered out in the downstream analysis.

2.2.3 Association analysis

Genotype coding

For a bi-allelic variation of interest, the common way to treat it in statistical genetics is to give a numeric value to the genotype effect with regard to its effect allele. The codings vary depending on the genetic model. Table 2.3 shows an example of typical genetic effect codings, supposing [G] is the effect allele of interest at a single [A/G] SNP.

Regression analysis

Quantitative traits are generally analyzed using a linear regression model, where the phenotype of interest is the outcome or the dependent variable (y) and the genotype for a

	AA	AG	GG
dosage/additive	0	1	2
dominant	0	1	1
heterozygous	0	1	0
recessive	0	0	1

Table 2.3: An Example of typical genetic effect codings for a single [A/G] SNP, assuming [G] is the effect allele of our interest.

variant (coded) is the predictor or the independent variable (x):

$$y = \beta_0 + \beta_1 x$$

In the situation of case-control studies, where the outcome of the disease is binary (i.e., case or control), logistic regression is used. The probability (p) of having the disease is modeled on a log odds scale:

$$log(p/(1-p)) = log(Prob(y=1)/Prob(y=0)) = \beta_0 + \beta_1 x$$

One can also include two or more predictive variables (e.g. one or two variants plus other covariates) in the model, as is

$$log(Prob(y=1)/Prob(y=0)) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$$

and an interaction term can be included as well:

$$log(Prob(y=1)/Prob(y=0)) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$$

2.3 Genome-wide association analysis for CAD in the 1000G era

2.3.1 Autosomal GWAS analysis

Study purpose. The following description refers to methods that were further used in the autosomal GWAS analysis based on traditional genotyping array, as part of the efforts contributed to CARDIoGRAMplusC4D for CAD in the 1000G era, which has been published as [77].

Participants

The individual level genotypes for autosomal GWAS analysis were collected from GerMIFS-I-IV studies(see Methods 2.1 for cohort descriptions).

Genotype processing

The genotype-level data were provided directly by our collaborator Dr. Christina Willenborg, from Institut für Integrative und Experimentelle Genomik (IIEG), Universität zu Lübeck, after QC processing and included 637 cases and 1,644 control individuals from GerMIFS-I, 1,222 cases and 1,298 control individuals from GerMIFS-II, 1,096 cases and 1,509 control individuals from GerMIFS-III, 1,016 cases and 1,147 control individuals from GerMIFS-IV. MDS analysis were performed based on these genotype data to generate the first two dimensions as a capture of genetic population stratification. The imputation procedures were performed as described previously (see Methods 2.2.2). The imputation INFO metric was recorded but without filtering, for the purpose of meta-analysis.

Association analysis

Logistic regression was performed using SNPTEST2 assuming additive, dominant and recessive models, with the top two dimensions resultant from autosomal MDS added to adjust for the population stratification.

As indicated by Pirinen et al [83], and also supported by Sarah Lewington and Robert Clarke from the CARDIoGRAMplusC4D consortium [15], the inclusion of non-confounding covariates into logistic regression models for genetic association analysis of case-control data has a variable impact on power. For the analysis of CAD, conventional risk factor covariates include age-of-onset of disease (and age-at-sampling for controls) and gender. These risk factors are usually assumed to be non-confounding with respect to gene associations. So their omission would not be expected to cause additional over-dispersion of the test statistics (i.e., inflation of the genomic control ratio), but improve the power to detect true signals. Therefore, age and gender were not considered in the adjustment of the association model.

2.3.2 X chromosome GWAS analysis

Study purpose. The following description refers to methods that were further used in the X chromosome GWAS analysis, as part of the efforts contributed to CARDIoGRAMplusC4D for CAD in the 1000G era, which has been summarized in the manuscript [59].

Participants

The individual level genotypes for X chromosome GWAS analysis were collected from GerMIFS-V cohort (see Methods 2.1 for cohort descriptions).

Genotype processing

The following pre-imputation QC criteria (classic QCed, Table 2.2) were taken:

Individual call rate $>= .95$

SNP call rate $> .98$

MAF $>= 0.01$

Sex check

Population outliers excluded (deviate beyond $mean \pm 5 \times sd$ for top two dimensions)

IBD < 0.25 (individuals distant away than second-degree relatives)

Heterozygosity within $mean \pm 3 \times sd$

Hardy-Weinberg $p > 1 \times 10^{-6}$

Both autosomal and X chromosomal genotypes underwent the above QC. After all quality control criteria, 2,459 cases and 1,611 controls from GerMIFS-V were taken into the analysis. The top two dimensions resulting from autosomal MDS analysis were used for adjusting the population stratification in the downstream association models. The imputation procedures were performed as formerly described (see Methods 2.2.2). Additional options (-chrX) in IMPUTE2, which is specific for imputation the non-pseudo-autosomal regions of chromosome X, was utilized. The imputation INFO metric was recorded but without filtering, for the purpose of meta-analysis. With the 1000G imputation, 957,983 X-chromosome variants were imputed and 315,957 SNPs remained after filtering out non-SNP variants and INFO ≤ 0.5. Finally, 220,458 SNPs were available before the central post-imputation QC.

Association analysis

The specific point to be noted for loci on the X chromosome is the inactivation issue (also called lyonization), a process by which one of the two X chromosomes in female mammals is inactivated. Since it is not known which loci are inactivated in females and to what degree, a combination of four different models was applied [60]. The association

analyses were performed in R in the four following models (model Ia, Ib, IIa, IIb):

$$glm(pheno \sim dose_a + sex + C1 + C2, family = binomial)$$
$$glm(pheno \sim dose_b + sex + C1 + C2, family = binomial)$$
$$glm(pheno \sim dose_a * sex + C1 + C2, family = binomial)$$
$$glm(pheno \sim dose_b * sex + C1 + C2, family = binomial)$$

where

$dose_a$: Assuming no inactivation of the locus in females [115]. With the standard coding in females with 0, 1, and 2 risk alleles; males are coded as 0 and 1 if carrying no or one risk allele, respectively.

$dose_b$: Assuming inactivation of the locus in females [11]. With the standard coding in females with 0, 1, and 2 risk alleles; males are coded as 0 and 2 if carrying no or one risk allele, respectively.

The adjustment for sex was added to the model to assure that type one error levels are not increased under unbalanced sample designs and sex-specific alelles frequencies [60]. The top two dimensions resulting from autosomal MDS were also added to adjust for the population stratification.

Effect estimates and standard errors from the above models were then submitted to CARDIoGRAMplusC4D [15] for further meta-analyses.

2.4 Understanding the genetic complexity of GWAS signals of CAD

2.4.1 Intra-locus allelic heterogeneity

Study purpose. The following description refers to methods that were further used in the intra-locus allelic heterogeneity analysis, as part of the efforts contributed to the manuscript [112].

Participants

Specifically for examination of intra-locus allelic heterogeneity, the individual level genotypes were collected from eight cohorts: GerMIFSI-V, LURIC, Cardiogenics and WTCCC-CAD (see Methods 2.1 for cohort descriptions).

Genotype processing

Refined QC (Table 2.2) and imputation procedures were performed as described previously (see Methods 2.2.2), resulting in 21,709 individuals in total.

SNP selection

A list of suggestive CAD risk variants was obtained from Supplementary Table 6 of the 1000G CAD GWAS publication [77], where the authors have generated a list of 202 variants and their corresponding effect sizes and p-values from conditional and joint analyses. Each of these variants was first annotated to their physically closest gene using Annovar [101], and then mapped to the known CAD loci, which were defined as the names that had been reported in either [16] or [77]. Independent variants within 2 centiMorgen (cM) distance (the flanking search space as recorded in the discovery procedure in [77]) were assigned to the known loci. Subsequently a list of known CAD

loci with more than one independent signal was compiled. The genotype data for the variants prepared above were then extracted from the genome-wide imputed data from the above described studies.

Polygenic score (PGS) calculation

I applied an intra-locus polygenic score to evaluate the combined effect of loci harboring multiple independent signals. At each locus, a weighted polygenic score was calculated by summarizing the number of risk alleles for each SNP (the posterior probabilities from imputation, a value from 0 to 2) and weighting the SNPs by their effect size ($beta$) reported in the conditional and joint analyses. The method is similar to the traditional polygenic score calculation, except that all the variants are located within a single locus.

For comparison, a second weighted genetic score was calculated simply based on the number of risk allele for the lead SNP and weighted by its corresponding effect size. Lead SNPs at each locus were defined as the SNP with the most significant p-value among all available SNPs in the dataset.

Evaluation of combined effects

Logistic regression was performed to evaluate the effect size of the polygenic scores (PGS) in each study. To consider all these PGSs on the same scale, they were modeled as a continuous variable and standardized into Z-scores (centered and scaled to have a mean of 0 and SD of 1). To adjust for the possible presence of population stratification, all analyses were adjusted with the top two dimensions resulting from autosomal MDS analysis, which were calculated with PLINK. A Nagelkerke's pseudo-R^2 was calculated to infer variance explained by the model.

The regression was performed for each study separately and afterwards a fixed-effect meta-analysis was performed to combine the effects across all studies. A combined p-value and R^2 were calculated via a weighted Z-score based on sample size of each

study.

I assessed the improvement in risk discrimination by comparing the area under the receiver operator characteristic (ROC) curves (AUC) in multi- versus lead- variant PGS.

Evaluation of relative effects

For loci harboring multiple independent signals, I grouped all individuals by the number of intra-locus risk alleles. The genotypes were assigned according to the largest posterior probabilities, and the dosage for the risk allele was rounded to a discrete value of 0 / 1 / 2 and the presence of CAD. Logistic regression was performed for each study separately based on the number of risk alleles as a categorical variable relative to the reference group. In order to achieve a better effect estimation, the group for the number of risk alleles which the majority of individuals carry was taken as the reference. The relative effect size of each group of individuals was evaluated in comparison with the majority group and was only computed when there were at least 5 cases and 5 controls in the group. The Cochran–Armitage test was performed to test the additive trend of CAD odds ratio with the incremental increase of risk alleles.

Afterwards, a fixed-effect meta-analysis was performed to combine the effects. A combined p-value and R^2 were calculated via a weighted Z-score based on sample size.

2.4.2 Multi-locus pleiotropy

Genetic association between height and CAD

Study purpose. The following description refers to methods that were further used in the multi-locus genetic risk score analysis between height and CAD, as part of the efforts contributed to the publication [75].

Participants. To investigate the association between height and CAD, the individual level genotypes were collected from five cohorts, GerMIFS-I-V (see Methods 2.1 for

cohort descriptions).

Genotype processing. Classic QC (Table 2.2) and imputation procedures were performed as described previously (see Methods 2.2.2), resulting in 13,385 individuals in total.

Polygenic Score Calculation. Based on 180 height-associated genetic variants [57], a weighted genetic risk score was calculated. For every SNP for each individual, the score was calculated on the basis of the sum of the dosage (posterior probabilities generated from imputation, a continuous value between 0 to 2) of the height-increasing allele and multiplied by the effect size observed for height. I then totaled these values across all SNPs for each individual, and the individuals were then grouped into quartiles.

Genetic association between rheumatoid arthritis (RA) and CAD

Study purpose. The following description refers to methods that were further used in the multi-locus genetic risk score analysis between RA and CAD, as part of the efforts contributed to the manuscript [46].

Participants. To investigate the association between RA and CAD, the individual level genotypes were collected from 7 cohorts, GerMIFS I-V, Cardiogenics and WTCCC (see Methods 2.1 for cohort descriptions).

Genotype processing. Refined QC (Table 2.2) and imputation procedures were performed as described previously (see Methods 2.2.2), resulting in 18,815 individuals in total.

Polygenic Score Calculation. Based on 61 reported RA-associated SNPs [29], a weighted genetic risk score was calculated. For each individual, we summed the dosage of RA risk alleles for each SNP (posterior probabilities generated from imputation, a continuous value between 0 to 2) and weighted the SNPs by their reported estimated effect size on RA.

The average genetic RA-risk score between CAD cases and controls was compared

using a two-sided T-test between CAD cases and controls in each study. Then a combined p-value was calculated via a weighted Z-score based on sample size.

Afterwards I analyzed association between CAD and tertiles of PGS using logistic regression adjusted for population stratification. Each cohort was analyzed separately, and the estimates weighted on the inverse of their square of standard errors were combined across cohorts with fixed effects meta-analysis. A combined p-value was calculated via a weighted Z-score based on sample size.

2.5 Detecting epistasis that underlies CAD

Study purpose. The following description refers to methods that were further used in the epistasis analysis of CAD, as part of the efforts contributed to the manuscript [113].

2.5.1 Participants

Individual level genotypes specifically for examination of intra-locus allelic heterogeneity were collected from nine cohorts: GerMIFS I-V, LURIC, Cardiogenics, WTCCC-CAD and MIGen (see Methods 2.1 for cohort descriptions).

2.5.2 Genotype processing

The following pre-imputation QC criteria (final QCed, Table 2.2) were taken:

Individual call rate $>= .98$

SNP call rate $> .98$

MAF > 0.01

Sex check

Population outliers excluded (deviate beyond $mean \pm 5 \times sd$ for top two dimensions)

IBD < 0.0625 (individuals distant away than fourth-degree relatives)

Heterozygosity rate within $mean \pm 3 \times sd$; F-statistic within $mean \pm 4 \times sd$

Hardy-Weinberg $p > 1 \times 10^{-6}$

which resulting in 27,370 individuals in total. The imputation procedures were performed as formerly described (see Methods 2.2.2).

The following post-imputation QC criteria were taken:

SNP call rate $> .98$

MAF > 0.05

Hardy-Weinberg $p > 1 \times 10^{-5}$

2.5.3 Heritability estimation

The lead-variants at the 46 loci reported from [16] and the 10 novel loci reported from [77] were collected from the main table of the original publications. The imputed individual-level data were merged from nine studies and only variants available in all studies were included. All variants in the flanking region around the lead-variants at these 56 known loci were extracted from the merged imputed genotype data. For variants that were not available in all studies, a proxy variant was resorted for each of them as the one having highest LD r^2 within a $\pm 200kb$ distance to it. The genetic kinship matrix was then calculated based on using LDAK with LD-adjustment [92]. The narrow-sense heritability of CAD was then estimated in the measure of the total variance in liability (assuming the prevalence of CAD as 5%) explained by all variants together, which was calculated via REML algorithm incorporated in the software GCTA [111].

2.5.4 GLIDE implementation

GLIDE is a high-performance GPU-based tool for detecting epistasis systematically based on regression analysis [48]. For instance, GLIDE enables to conduct a systematic epistasis search on the GWAS data published by the Wellcome Trust Consortium in

about 6 h per data set using a relatively inexpensive setup of 12 GPUs. This is a huge speed upgrade in comparison to a single-core CPU-based setup, where a similar approach would take roughly 1 year to be completed [48]. This speed upgrade is achieved via GPU threads cooperation and parallel computation. For an epistatic interactions matrix of size $n \times n$ to be computed, it is divided into chunks of size $nGPU \times nGPU$. The chunks are computed sequentially. Each chunk is divided into blocks of the size $BS \times BS$. Each of those blocks is computed in parallel by $BS \times BS$ threads [48]. In collaboration with the Max Planck Institute of Psychiatry, where a high-performance GPU cluster has been set up, we were able to search for epistasis with GLIDE implementation.

2.5.5 Genotype coding

For each pair of variants to test, all 4 genotype codings in Table 2.3 were applied, resulting in 4×4 combinations of possible genotypic effect patterns.

2.5.6 Statistical epistasis test for CAD

We explored the statistical epistasis in two steps (Figure 2.1): step I served as a primary filtering of potential candidates and step II served as main screening and final confirmation.

At step I, we searched for pairwise epistasis for all variants located physically at a broad-sense of a CAD susceptibility region and with LD-pruned $r^2 < 0.5$ from each other (n =8,068 SNPs in total). Logistic regression (Eq. 2.1) was first performed with GLIDE for all 16 possible genotypic effect patterns for each study, and then a fixed-effect meta-analysis was performed to estimate the effect size and standard error. A meta-analysis p-value was also calculated via a weighted Z-score based on the sample size. For all resulting pairs from all 16 genotypic models from GLIDE, those that passed a loose and arbitrary meta-analysis $p < 1 \times 10^{-8}$ were taken as potential candidates and chosen for

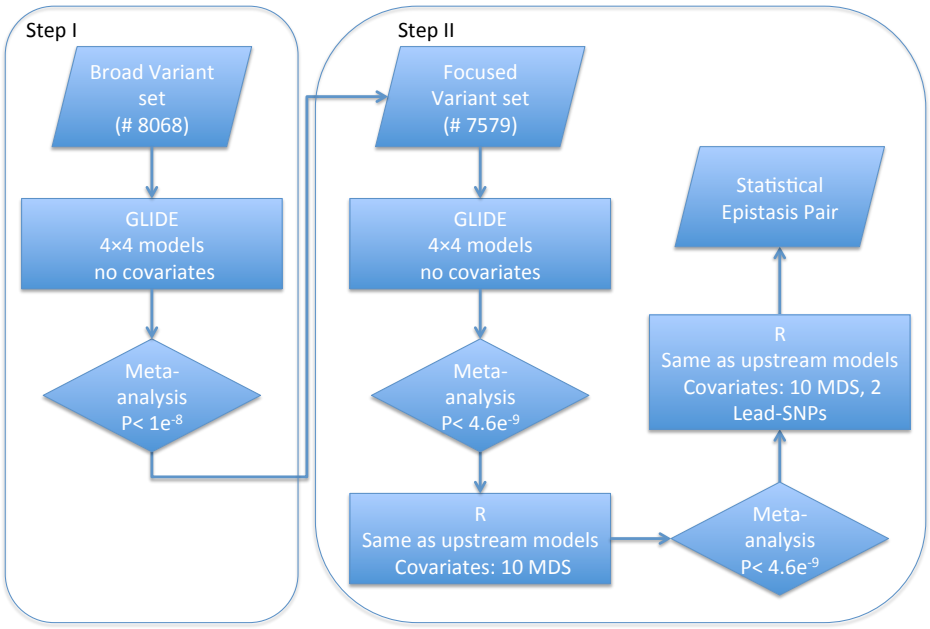


Figure 2.1: Scheme of statistical epistasis test for CAD

the subsequent step II analysis.

At step II, we searched for pairwise epistasis for all variants within the full LD-expanded region of the potential candidate pairs from step I analysis (n =7,579 SNPs in total). In order to calculate a proper significance level, LD-based clumping was performed via PLINK [85] to determine the total number of independent SNPs, which gave us a number of n =4,654. In this way a final significance level was calculated with Bonferroni adjustment $0.05/(n \times (n-1)/2) = 4.618 \times 10^{-9}$.

Logistic regression was performed with GLIDE in the same way as in step I (Eq. 2.1). For those pairs that passed the significance level, further regression analyses including population covariates (Eq. 2.2) were performed with R [86] based on the same genetic model. For those pairs still passing the significance level, further regression analyses including the nearest CAD lead-SNPs were performed with R again to confirm the source of effect (Eq. 2.3).

$$log(Prob(y=1)/Prob(y=0)) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 \tag{2.1}$$

$$\begin{aligned} log(Prob(y=1)/Prob(y=0)) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \\ \beta_4 C_1 + \beta_5 C_2 + \beta_6 C_3 ... + \beta_{13} C_{10} \end{aligned} \tag{2.2}$$

$$\begin{aligned} log(Prob(y=1)/Prob(y=0)) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \\ \beta_4 C_1 + \beta_5 C_2 + \beta_6 C_3 ... + \beta_{13} C_{10} + \beta_{14} z_1 + \beta_{15} z_2 \end{aligned} \tag{2.3}$$

where, y is the binary status of CAD ($y = 1$ if case; $y = 0$ if control), x_1 and x_2 are genotype codings for the two variants in the test, and the covariates C_1 to C_{10} are the top 10 dimensions resulting from autosomal MDS were to adjust for the population stratification. z_1 and z_2 are genotype codings for the two CAD lead-SNPs closest to the two variants in the test, β_1 reports the main effects of the coded variables of variant 1.

β_2 reports the main effects of the coded variables of variant 2. β_3 reports the interaction effects of the coded variables of variant 1 and 2.

2.5.7 Statistical epistasis test for gene expression

Part of the individuals in Cardiogenics (see Methods 2.1 for cohort descriptions) were also investigated with transcript abundance via gene expression microarray for monocytes and macrophages from blood samples.

The individuals with both genotypes and gene expression data available were extracted and categorized into six sample groups: (i) CAD cases with gene expression data available in monocytes, (ii) CAD-free controls with gene expression data available in monocytes, (iii) all individuals (regardless of CAD onset or not) with gene expression data available in monocytes, (iv) CAD cases with gene expression data available in macrophages, (v) CAD-free controls with gene expression data available in macrophages, and (vi) all individuals (regardless of CAD onset or not) with gene expression data available in macrophages.

Both linear regression (Eq. 2.4) and ANOVA analyses (Eq. 2.5) were performed for each gene on the microarray and for each of the 6 sample groups, to test whether the CAD epistatic pair also shows an epistatic effect on any gene expression level.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \beta_4 C_1 + \beta_5 C_2 + \beta_6 C_3 ... + \beta_{13} C_{10} \tag{2.4}$$

$$y = \beta_0 + \beta_1 x_{1l} + \beta_2 x_{2l} + \beta_3 x_{il} + \beta_4 C_1 + \beta_5 C_2 + \beta_6 C_3 ... + \beta_{13} C_{10} \tag{2.5}$$

where, y is the normalized gene expression for each gene. x_1 and x_2 are genotype codings for the two variants in regression. x_{1l} and x_{2l} are genotype levels for the two variants in ANOVA, and x_{il} is the genotype level for any of the 3×3 genotype

combinations of the two variants. The covariates C_1 to C_{10} are the top 10 dimensions resulting from autosomal MDS were to adjust for the population stratification.

The p-values for the effect of the interaction term β_3 for each test were recorded as the observed p-values. To adjust for multiple testing, 1000x Permutation were performed at first by scrambling 1000x the indexes of the samples for the gene expression matrix to destroy the true genotype-expression relationship, and generating 1000x the pseudo effects and p-values. Finally, an exact p-value was calculated according to [82].

2.5.8 Correlation between gene expression and CAD

For each of the $3 \times 3 = 9$ genotype combinations according to the pair of epistatic SNPs, a putative CAD odds ratio (putative OR-CAD) was calculated based on the pooled (all 9 studies) number of individuals of CAD cases and controls. In this way a putative OR-CAD could be assigned to each individual according to its genotype combination of the epistatic SNP pair. Pearson correlation test was performed for each gene on the microarray to test the linear correlation between gene expression and the log scale putative OR-CAD across the individuals.

2.5.9 Statistical epistasis test for clinical traits

Several basic traits and CAD-related traits are also available from the GerMIFS-I, GerMIFS-II, GerMIFS-V and LURIC studies, including, body-mass index, total cholesterol, LDL cholesterol, HDL cholesterol, and triglycerides. Triglyceride levels are typically skewed, which was also the case in our data, thus they were converted into the log scale. For the epistasis SNP-pair of interest, a similar statistical epistasis test was performed for each of these available traits. Linear regression analysis (Eq. 2.4) was performed (here, y is the trait value) to test whether the CAD epistatis pair also shows an epistatic effect on any of the above five CAD-related traits (outliers (out of range

$mean \pm 5 \times sd$) were excluded). Subsequently a combined p-value was calculated via a weighted Z-score based on sample size of each study.

2.5.10 Motif enrichment

For all transcripts having expression data available, the DNA sequences were extracted from UCSC Genome Browser [49], at the flanking 2kb of transcriptional starting site (TSS), except for genes with multiple non-unique TSS. These TSS$\pm$2kb DNA sequences were taken as putative gene promoter sequences and divided into two sets: *set-a* for genes associated with the epistatis pair; *set-b* for genes not associated with the epistatis pair.

Motif enrichment analyses were performed with MEME–Suite 4.11 [5]. MEME was first applied to search for ungapped motifs enriched in the sequence *set-a*, relative to (i) random sequences simulated with base-pair frequencies same as *set-a*, and (ii) *set-b* sequences. The resultant enriched motifs were followed with CentriMo, which tests whether the motif has a particular location preference in the input sequences. The enriched motifs were also forwarded to TOMTOM, by which an alignment to the known transcription factor binding motifs was produced. The JASPAR 2016 core vertebrates database [66] was taken as the known motif database.

Chapter 3

Results

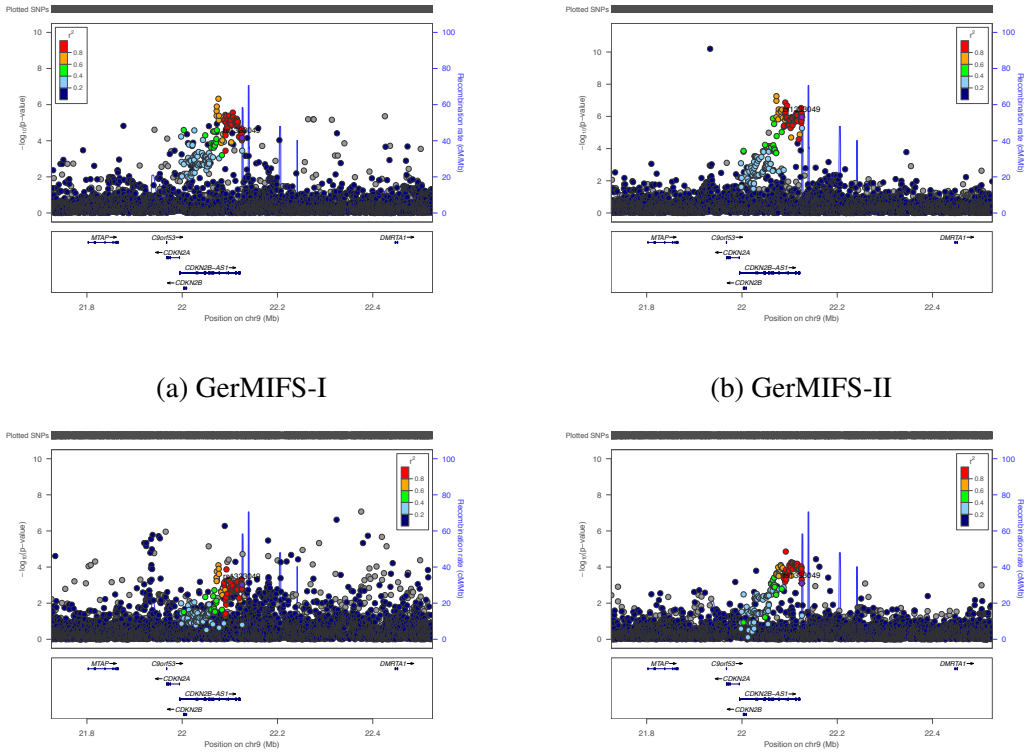
3.1 Genome-wide association analysis for CAD in the 1000G era

3.1.1 Autosomal GWAS analysis

Study purpose. The following description refers to results that were generated in the autosomal GWAS analysis based on traditional genotyping array, as part of the efforts contributed to CARDIoGRAMplusC4D for CAD in the 1000G era, which has been published in 2015 [77].

Single-study analysis

We investigated the genome-wide association for the autosomal genomes in German MI Family studies I-IV. The association analyses were performed assuming additive, dominant, or recessive models. Based on summary statistics for the additive model in all studies analyzed, the most significant signals were reached at 9p21 locus (Figure 3.1), which is the best replicated genetic susceptibility locus for CAD. The p-values for GWAS studies ranged from 1.39×10^{-5} in GerMIFS-I to 6.33×10^{-11} in GerMIFS-II.



(a) GerMIFS-I

(b) GerMIFS-II

(c) GerMIFS-III

(d) GerMIFS-IV

Figure 3.1: The y axis shows the log10 P values of all variants within the region ±400kb of SNP rs1333049 under the additive genetic model, and the x axis shows their chromosomal positions. Panels (a-d) present the results for GerMIFS I-IV, respectively.

The summary statistics for the association studies assuming additive, dominant, or recessive models were submitted to the statistical team of the CARDIoGRAMplusC4D consortium [15] for further post-imputation QC and meta-analyses.

Meta-analysis

The meta-analysis for 1000G CAD autosomal GWAS was performed by the statistical team of the CARDIoGRAMplusC4D consortium [77], with 60,801 cases and 123,504 controls from 48 studies. The meta-analysis results have been published by Nikpay et al [77].

Genome-wide associations were scanned for both additive and non-additive models. Based on the summary statistics of meta-analysis results for the additive genetic model, 47 out of the 48 loci previously reported as CAD susceptibility loci were recovered with nominally significance. The only exception was the lead SNP of a region, which had been previously detected as being specific for Han Chinese. In total 2,213 variants showed significant associations with CAD ($p<5 \times 10^{-8}$) with a low false discovery rate (FDR q value $< 2.1 \times 10^{-4}$) assuming the additive model [77], which represented eight novel regions at genome-wide levels of significance. In addition, based on the summary statistics for the recessive model, two novel recessive susceptibility loci were identified. For the dominant model, multiple strong associations were also identified, all of which overlapped with loci that had already been identified in the additive model [77]. All these ten newly identified CAD associations are represented by risk alleles with a frequency of >5%. Figure 3.2 displays a summary of the 1000G CAD GWAS results in a circular Manhattan plot.

FDR and heritability analysis

Based on variants around each lead-variant showing suggestive additive association in the meta-analysis, the statistical team of the CARDIoGRAMplusC4D consortium [77]

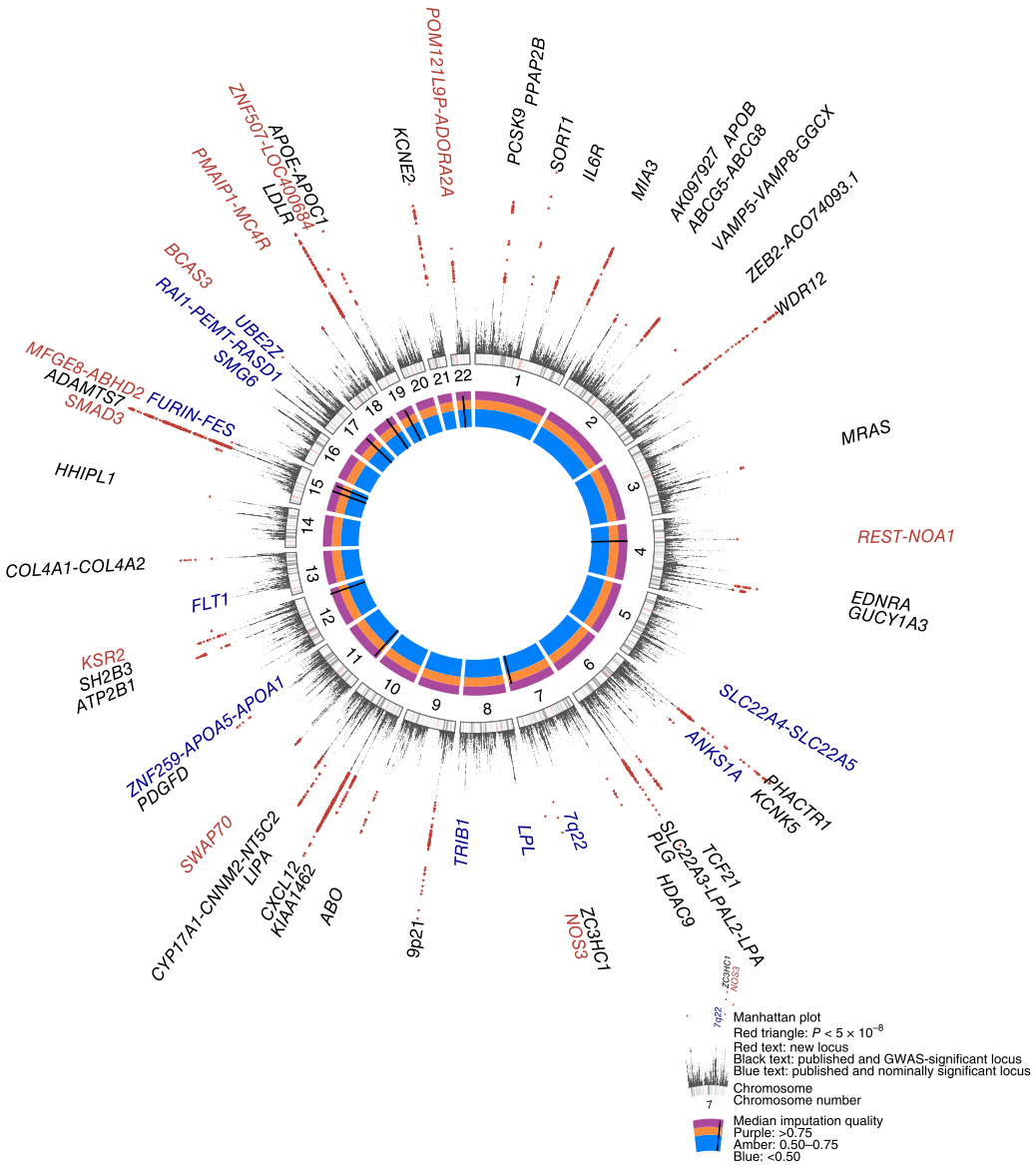


Figure 3.2: (taken directly from the publication Nikpay et al [77]) A circular Manhattan plot summarizing the 1000G Genomes Project CAD association results. Red text: ten new CAD-associated loci. Black text: previously reported loci showing genome-wide significant association. Blue text: previously reported loci showing nominal significance ($p < 0.05$). Inner track: the imputation quality of the lead variants in the new loci. Middle track: numbered chromosome ideograms with centromeres represented by pink bars.

further performed conditional and joint analysis, by which 202 FDR variants (median MAF = 0.22) were identified (FDR q value < 0.05). Fifteen low-frequency (MAF < 0.05) variants explained only $2.1 \pm 0.2\%$ of CAD heritability, and all were either a lead variant or were jointly associated (q value < 0.05) with a common variant. Ninety-five variants (explaining $13.3 \pm 0.4\%$ of CAD heritability) mapped to 44 significant loci from GWAS, suggesting the presence of multiple independent signal at a single locus.

3.1.2 X chromosome GWAS analysis

Study purpose. The following description refers to results that were generated in the X chromosome GWAS analysis, as part of the efforts contributed to CARDIoGRAMplusC4D for CAD in the 1000G era, which has been summarized in the manuscript Loley et al [59].

Single-study analysis

We investigated the genome-wide association for the non-pseudo-autosomal region of chromosome X in the GerMIFS-V study. The association analysis was performed assuming no inactivation and no sex interaction (model Ia), with inactivation and no sex interaction (model Ib), no inactivation and with sex interaction (model IIa), and with both inactivation and sex interaction (model IIb). Our single study resulted in 77 variants reaching genome-wide significance ($p < 5 \times 10^{-8}$) in any of the models out of all 195,102 variants with good imputation quality (INFO > 0.8). The effect estimates and standard errors from four different models assuming X-inactivation status were submitted to the statistical team of the CARDIoGRAMplusC4D consortium [15] for further post-imputation QC and meta-analyses.

Meta-analysis

The meta-analysis for 1000G CAD X-chromosome GWAS was performed by the statistical team of the CARDIoGRAMplusC4D consortium [59], with about 200,000 X chromosomal SNPs after QC and 43,120 cases with CAD and 58,291 controls from 35 studies in total. The results have been submitted and are under review [59].

The meta-analysis with random effect models were calculated for each of the four models (i.e., Ia, Ib, IIa, IIb). The meta-analyses on the effect estimates for the SNP-sex interaction were also performed in the same way. As a result, none of the statistical models used for meta-analysis revealed genome-wide significant association with CAD for any SNP, even with stricter quality control or excluding non-European individuals [59].

3.2 Understanding the genetic complexity of GWAS signals of CAD

3.2.1 Multiple independent signals at known CAD susceptibility loci

Study purpose. The following description refers to results that were generated in the intra-locus allelic heterogeneity analysis, as part of the efforts contributed to the manuscript [112].

Known CAD loci harboring multiple independent signals

From all the suggestive CAD susceptibility I compiled variants reported in the conditional and joint analysis results of the 1000G CAD GWAS [77]. For each variant the risk allele was defined as the one with positive beta from the joint analysis. For each locus, the lead variant was defined as the one with the most significant p-value from the joint analysis

that was also available in our datasets. In total we obtained a list of 80 variants at 25 loci (Table 3.1).

Known Locus	SNP	Chr	Risk/ Other Allele (C/J)	RAF	beta (C/J)	p (C/J)	Avail-ability	Lead-SNP
PPAP2B	rs9970807	1	C/T	0.92	0.13	2.12E-15	1	1
PPAP2B	rs61772626	1	G/A	0.12	0.08	3.76E-07	1	0
SORT1	rs7528419	1	A/G	0.79	0.08	4.88E-08	1	1
SORT1	rs1277930	1	A/G	0.74	0.07	4.08E-06	1	0
SORT1	chr1:110299165:I	1	D/I	0.63	0.07	1.21E-10	0	0
IL6R	rs6689306	1	A/G	0.45	0.06	2.61E-09	1	0
IL6R	rs72702224	1	A/G	0.4	0.08	3.11E-12	1	1
MIA3	rs67180937	1	G/T	0.66	0.09	3.37E-16	1	1
MIA3	rs75082168	1	A/T	0.04	0.15	4.75E-06	1	0
ABCG5/ABCG8	rs13420649	2	C/T	0.21	0.08	1.30E-08	1	1
ABCG5/ABCG8	chr2:44074126:D	2	I/D	0.74	0.14	7.41E-19	0	0
VAMP5/VAMP8/GGCX	rs11126366	2	C/G	0.81	0.09	1.50E-08	1	1
VAMP5/VAMP8/GGCX	rs11126387	2	C/T	0.45	0.07	1.00E-07	1	0
ZEB2/ACO74093.1	rs7564469	2	C/T	0.2	0.06	2.71E-06	1	0
ZEB2/ACO74093.1	rs17678683	2	G/T	0.09	0.09	6.20E-08	1	1
ZEB2/ACO74093.1	rs2252654	2	A/G	0.31	0.05	2.01E-06	1	0
WDR12	rs7559543	2	T/C	0.18	0.06	2.16E-06	1	1
WDR12	chr2:203828796:I	2	I/D	0.11	0.15	7.91E-21	0	0
EDNRA	rs4593108	4	C/G	0.8	0.1	1.83E-15	1	1
EDNRA	rs6842241	4	A/C	0.17	0.09	5.63E-12	1	0
GUCY1A3	chr4:156366138:I	4	D/I	0.96	0.21	1.12E-12	0	0
GUCY1A3	rs13140296	4	G/A	0.53	0.06	2.22E-09	1	0
GUCY1A3	rs1001037	4	T/C	0.85	0.07	7.72E-08	1	0

Continued on next page

Table 3.1 – *Continued from last page*

Known Locus	SNP	Chr	Risk/ Other Allele (C/J)	RAF	beta (C/J)	p (C/J)	Avail-ability	Lead-SNP
GUCY1A3	rs72685791	4	G/A	0.8	0.08	9.81E-11	1	1
PHACTR1	chr6:12619932:D	6	I/D	0.37	0.05	2.84E-06	0	0
PHACTR1	rs9349379	6	G/A	0.43	0.18	0.00E+00	1	1
TCF21	rs12202017	6	A/G	0.7	0.06	1.80E-10	1	1
TCF21	rs2327433	6	G/A	0.14	0.07	6.08E-07	1	0
SLC22A3/LPAL2/LPA	rs9364537	6	G/A	0.32	0.09	2.63E-15	1	0
SLC22A3/LPAL2/LPA	chr6:160265331:D	6	I/D	0.98	0.25	8.62E-09	0	0
SLC22A3/LPAL2/LPA	rs6932293	6	C/T	0.04	0.22	3.59E-06	1	0
SLC22A3/LPAL2/LPA	rs624249	6	C/A	0.63	0.06	8.49E-09	1	0
SLC22A3/LPAL2/LPA	chr6:160776695:I	6	I/D	0.01	0.43	3.13E-09	0	0
SLC22A3/LPAL2/LPA	rs9457927	6	G/A	0.02	0.43	4.77E-24	1	1
SLC22A3/LPAL2/LPA	rs55730499	6	T/C	0.06	0.19	7.67E-09	1	0
SLC22A3/LPAL2/LPA	rs12201989	6	A/T	0.83	0.08	1.14E-06	1	0
SLC22A3/LPAL2/LPA	rs56393506	6	T/C	0.16	0.12	6.77E-14	1	0
SLC22A3/LPAL2/LPA	rs1998043	6	G/A	0.16	0.09	3.19E-09	1	0
SLC22A3/LPAL2/LPA	rs6935921	6	T/C	0.65	0.07	2.48E-10	1	0
SLC22A3/LPAL2/LPA	rs186696265	6	T/C	0.01	0.35	5.47E-12	1	0
SLC22A3/LPAL2/LPA	rs75176946	6	T/C	0.01	0.24	3.21E-06	0	0
SLC22A3/LPAL2/LPA	rs112215831	6	G/A	0.9	0.1	6.00E-07	1	0
9p21	rs13301437	9	C/T	0.14	0.08	1.37E-06	1	0
9p21	rs7855162	9	C/T	0.04	0.16	4.17E-09	1	0
9p21	rs1970112	9	C/T	0.49	0.12	4.50E-22	1	0
9p21	rs62555370	9	G/A	0.88	0.09	1.99E-07	1	0
9p21	rs1333046	9	A/T	0.5	0.12	1.41E-22	1	1

Continued on next page

Table 3.1 – *Continued from last page*

Known Locus	SNP	Chr	Risk/ Other Allele (C/J)	RAF	beta (C/J)	p (C/J)	Avail-ability	Lead-SNP
KIAA1462	rs7917431	10	C/T	0.66	0.05	1.24E-06	1	0
KIAA1462	rs2487928	10	A/G	0.42	0.06	7.83E-11	1	1
CXCL12	rs58030109	10	A/G	0.02	0.17	3.14E-06	1	0
CXCL12	rs11238720	10	G/C	0.13	0.1	1.71E-09	1	0
CXCL12	rs1870634	10	G/T	0.64	0.07	4.68E-13	1	0
CXCL12	rs1746050	10	C/A	0.85	0.12	6.38E-16	1	1
CYP17A1/CNNM2/NT5C2	rs11191416	10	T/G	0.87	0.09	5.97E-11	1	1
CYP17A1/CNNM2/NT5C2	rs11813268	10	C/T	0.83	0.07	1.38E-07	1	0
SWAP70	rs4627080	11	G/T	0.1	0.07	2.40E-06	1	0
SWAP70	rs10840293	11	A/G	0.55	0.06	4.74E-09	1	1
SH2B3	rs7967514	12	A/G	0.06	0.13	1.03E-06	1	0
SH2B3	rs4766578	12	T/A	0.43	0.09	7.39E-15	1	1
COL4A1/A2	rs11617955	13	T/A	0.89	0.1	1.39E-09	1	0
COL4A1/A2	rs4773141	13	G/C	0.36	0.08	1.78E-10	1	0
COL4A1/A2	rs11838776	13	A/G	0.26	0.07	8.86E-11	1	1
COL4A1/A2	rs9515203	13	T/C	0.76	0.07	1.08E-08	1	0
COL4A1/A2	rs34905765	13	T/C	0.1	0.08	6.41E-07	1	0
COL4A1/A2	rs56003851	13	C/A	0.8	0.08	1.97E-10	1	0
COL4A1/A2	rs61969072	13	G/T	0.17	0.06	2.87E-06	1	0
ADAMTS7	rs11635330	15	C/T	0.6	0.05	3.30E-07	1	0
ADAMTS7	rs4887109	15	C/T	0.68	0.08	4.46E-11	1	1
ADAMTS7	rs4468572	15	C/T	0.59	0.06	1.67E-08	1	0
BCAS3	rs7212798	17	C/T	0.15	0.1	9.51E-12	1	1
BCAS3	rs2270114	17	C/G	0.65	0.06	2.01E-07	1	0

Continued on next page

Table 3.1 – *Continued from last page*

Known Locus	SNP	Chr	Risk/ Other Allele (C/J)	RAF	beta (C/J)	p (C/J)	Avail-ability	Lead-SNP
LDLR	rs12979495	19	G/A	0.75	0.05	6.97E-07	1	0
LDLR	rs56289821	19	G/A	0.9	0.11	2.45E-10	1	1
LDLR	rs6511721	19	G/A	0.48	0.06	4.59E-07	1	0
APOE/APOC1	rs118147862	19	G/A	0.97	0.17	9.84E-08	1	0
APOE/APOC1	rs405509	19	T/G	0.5	0.06	1.45E-09	1	1
APOE/APOC1	rs4420638	19	G/A	0.17	0.08	2.08E-09	1	0
APOE/APOC1	chr19:45801579:D	19	D/I	0.38	0.05	2.02E-07	0	0
KCNE2	rs28451064	21	A/G	0.12	0.12	3.08E-12	1	1
KCNE2	rs7280276	21	A/G	0.24	0.06	9.98E-07	1	0

Table 3.1: Known CAD loci harboring multiple independent signals

The number of multiple independent signals was highest at the *LPA* locus (14 variants), followed by *COL4A1/A2* (7 variants) and the 9p21 locus (5 variants) (Table 3.2). Three loci (*ABCG5/ABCG8*, *WDR12*, *PHACTR1*) were reported to harbor two independent variants, but only one of the variants at each of these loci was available in the individual-level genotypes after QC. The *MIA3* locus was also reported to harbor two independent variants, but due to the low risk allele frequencies of one of them (0.03) the data were inconclusive in most samples. Therefore, these four loci were filtered out in further examination.

Combined effect of loci harboring multiple independent signals

First, I aimed to examine whether the combined analysis of variants at loci with multiple independent signals would give stronger effects on CAD risk than effects that had been

Known Locus	No. Independent Signals
SLC22A3/LPAL2/LPA	14
COL4A1/A2	7
9p21	5
APOE-APOC1	4
CXCL12	4
GUCY1A3	4
ADAMTS7	3
LDLR	3
SORT1	3
ZEB2/ACO74093.1	3
ABCG5/ABCG8	2
BCAS3	2
CYP17A1/CNNM2/NT5C2	2
EDNRA	2
IL6R	2
KCNE2 (gene desert)	2
KIAA1462	2
MIA3	2
PHACTR1	2
PPAP2B	2
SH2B3	2
SWAP70	2
TCF21	2
VAMP5/VAMP8/GGCX	2
WDR12	2

Table 3.2: Number of independent signals at known CAD loci. The list were compiled based on the conditional and joint analysis results of 1000G CAD GWAS [77].

reported for the lead SNPs in our individual-level genotype analysis. For each of the 21 loci, two measures were calculated: one score based only on a lead variant (lead-variant PGS), which served as a baseline effect of the known locus in our individual genotype data; another score based on multiple independent variants at a locus (multi-variant PGS) and the risk of CAD for comparison. A basic logistic regression model was applied for lead- and multi- variant PGS.

The reported risk loci for CAD were robust in our individual level datasets for all 21 loci considered to harbor multiple signals, as the positive effect could be observed for almost all the lead-variant PGS, despite the possible sampling difference (Table 3.3). The only exception is at the *VAMP5/VAMP8/GGCX* locus, which obtained a negative effect for the multi-variant PGS. For 14 out of the 21 loci, the lead-variant PGS showed a marginally associated trend (meta-analysis p-value<0.05), and for 17 out of 21 loci, the multi-variant PGS was associated (meta-analysis p-value<0.05) with CAD.

A combined effect was observed for most of the loci, which was measured as the odds ratio for each SD of the polygenic scores. When comparing the effect of scores based on multiple variants versus that of the single lead variant for each of the 21 loci, two-third of them showed a larger effect and explained more variance (Figure 3.3). Nevertheless, the improvements of fit for additional variance explained at a single locus were all very slight, ranging from 0.02% (*LDLR*) to 0.24% (*COL4A1/A2*). The only exception was at the *APOE* locus, which had additional 4.367% of variance explained. However, the p-value indicated non-significance for both PGS (p>0.2) (Table 3.3). The remaining one-third of the 21 loci did not show a larger effect or more variance explained, only to a slight degree as well (Table 3.3, Figure 3.3).

Among all these loci, four achieved genome-wide significance ($p < 5 \times 10^{-8}$) with multi-variant PGS, including three loci that did not achieve the same significance level for the lead-variant PGS (the *SLC22A3/LPAL2/LPA* locus, the *COL4A1/A2* locus, and the *KCNE2* (gene-desert) locus), and the 9p21 locus, which already shows a high effect

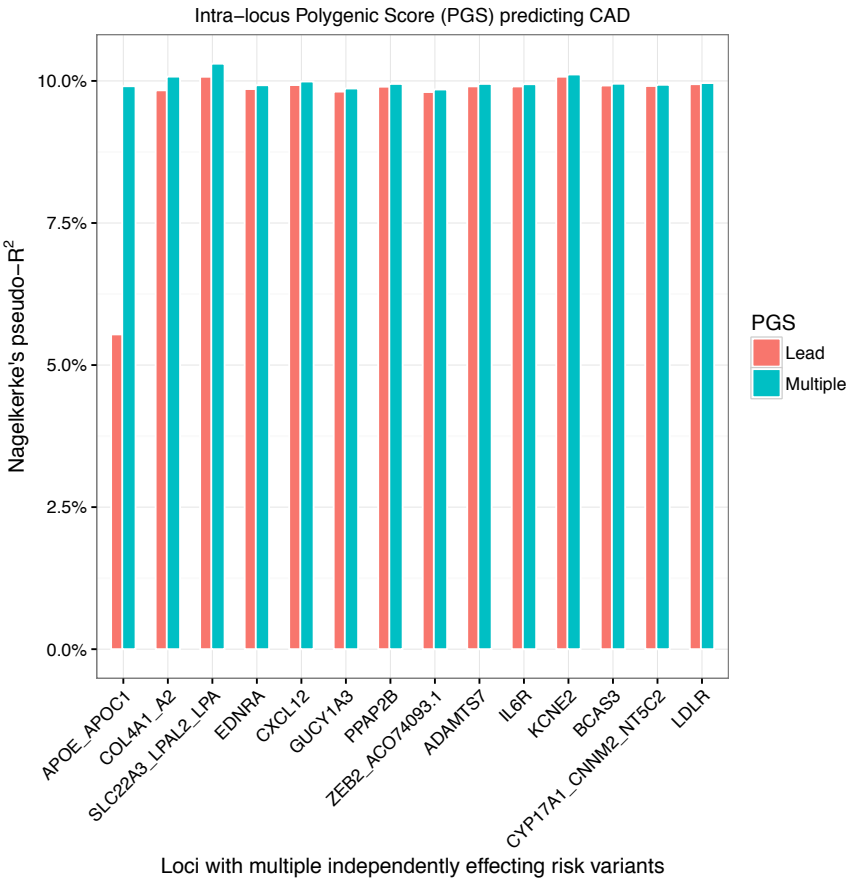


Figure 3.3: More variance explained (estimated in Nagelkerke's pseudo-R^2) for logistic regression based on multi-variant PGS than lead-variant PGS. The increase trend of variance explained (Nagelkerke's pseudo-R^2) by PGS for 14 loci are displayed.

by only the lead-variant itself. All these four loci presented a combined effect of OR > 1.09 for one standard deviation increase of multi-variant PGS, as well as a goodness of fit measure $R^2 > 10\%$ (Table 3.3).

Prediction evaluation

To further explore the utility of a PGS approach in predicting CAD risk, we selected two loci with the potentially largest predictive value for this investigation, i.e., *SLC22A3/LPAL2/LPA* and *COL4A1/A2*. The two loci were the only ones that not only achieved genome-wide significance, but also showed a much larger additional R^2 compared to others. Also, they were reported to harbor the largest number of independent signals (14 at *LPA*; 7 at *COL4A1/A2*) . We therefore performed ROC analysis to compare the discriminatory ability of the regression model based on multi-variant PGS versus lead-variant PGS.

However, neither of the two loci showed satisfactory AUC (above 0.8) in any study. The AUC of multi-variant PGS range from 0.52 (LURIC) to 0.77 (Cardiogenics) for *LPA*; and from 0.54 (LURIC) to 0.77 (Cardiogenics) for *COL4A1/A2*. the measure of AUC, based on paired T-test for the AUC values for all eight studies, the PGS was significantly but only slightly better in discriminating CAD patients from controls at the *LPA* locus, and below the marginal significance level (0.05) at the *COL4A1/A2* locus (Figure 3.4).

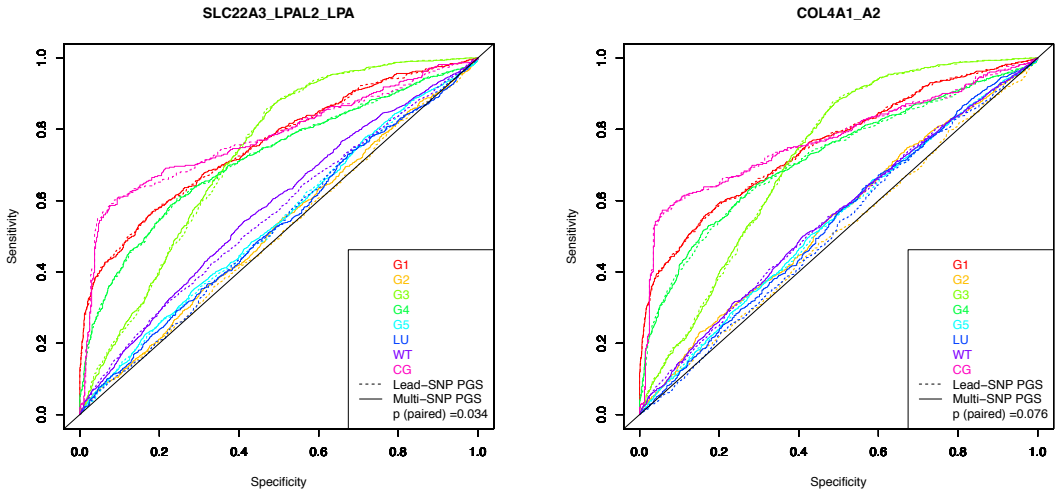
Intra-locus allelic heterogeneity

For the four loci showing the largest multi-variant effects (the *SLC22A3/LPAL2/LPA* locus, the *COL4A1/A2* locus, and the *KCNE2* (gene-desert) locus , and the 9p21 locus), I further interrogated the distribution of multiple variants among the individuals.

I grouped all individuals by the number of risk alleles and calculated an relative odds ratio based on individual counts in CAD cases and controls. For all of these four loci an incremental increase in the odds ratios was observed with increasing numbers of risk alleles. Comparing individuals with the highest number of risk alleles with those

Known Locus	Lead-variant PGS				Multi-variant PGS				Additional $R^2(\%)$
	OR	.95CI	P	$R^2(\%)$	OR	.95CI	P	$R^2(\%)$	
SLC22A3/LPAL2/LPA	1.09	[1.06,1.13]	2.86E-07	10.07	1.14	[1.11,1.17]	4.44E-16	10.3	0.23
COL4A1/A2	1.04	[1.01,1.07]	9.19E-02	9.83	1.099	[1.07,1.13]	1.50E-08	10.08	0.241
KCNE2	1.09	[1.06,1.12]	6.43E-07	10.07	1.096	[1.06,1.13]	3.93E-08	10.11	0.039
CXCL12	1.07	[1.04,1.1]	8.64E-05	9.93	1.087	[1.06,1.12]	8.76E-07	9.99	0.062
BCAS3	1.05	[1.02,1.08]	2.37E-04	9.92	1.055	[1.02,1.09]	2.68E-05	9.95	0.031
PPAP2B	1.06	[1.03,1.09]	5.20E-03	9.9	1.076	[1.04,1.11]	6.10E-05	9.95	0.049
EDNRA	1.05	[1.02,1.08]	4.88E-02	9.86	1.058	[1.03,1.09]	2.00E-04	9.92	0.067
LDLR	1.06	[1.03,1.1]	1.94E-03	9.94	1.067	[1.04,1.1]	5.33E-04	9.96	0.02
CYP17A1/CNNM2/NT5C2	1.06	[1.03,1.09]	3.53E-03	9.91	1.065	[1.03,1.1]	1.12E-03	9.93	0.021
ADAMTS7	1.05	[1.02,1.08]	1.01E-01	9.9	1.063	[1.03,1.09]	6.21E-03	9.95	0.043
IL6R	1.04	[1.01,1.07]	7.14E-02	9.9	1.054	[1.02,1.08]	1.10E-02	9.94	0.04
GUCY1A3	1.03	[1,1.06]	6.38E-01	9.81	1.051	[1.02,1.08]	5.38E-02	9.87	0.054
ZEB2/ACO74093.1	1.02	[0.99,1.05]	5.57E-01	9.8	1.046	[1.02,1.08]	5.78E-02	9.85	0.045
APOE/APOC1	1.02	[0.99,1.06]	2.79E-01	5.54	1.044	[1.01,1.07]	2.83E-01	9.91	4.367
9p21	1.2	[1.16,1.23]	2.05E-31	10.67	1.193	[1.16,1.23]	6.98E-31	10.66	-0.006
SORT1	1.09	[1.06,1.13]	3.78E-07	10.07	1.094	[1.06,1.13]	7.57E-07	10.06	-0.005
KIAA1462	1.06	[1.03,1.09]	1.18E-04	10.02	1.043	[1.01,1.07]	4.98E-05	10.01	-0.01
SH2B3	1.06	[1.03,1.09]	1.56E-04	9.99	1.062	[1.03,1.09]	2.95E-04	9.98	-0.005
TCF21	1.06	[1.03,1.09]	1.76E-05	9.95	1.059	[1.03,1.09]	4.30E-03	9.9	-0.056
SWAP70	1.03	[1,1.06]	8.48E-03	9.9	1.034	[1,1.06]	2.64E-02	9.87	-0.027
VAMP5/VAMP8/GGCX	1.01	[0.99,1.04]	3.20E-01	9.82	0.997	[0.97,1.03]	5.86E-01	9.81	-0.01

Table 3.3: Combined effect of loci harboring multiple independent signals. For each locus both, lead-variant PGS and multi-variant PGS were calculated. Logistic regression was performed to estimate the effect of two PGS. R^2 represents the Nagelkerke's R^2.



(a) *SLC22A3/LPAL2/LPA* (b) *COL4A1/COL4A2*

Figure 3.4: ROC plot for multi-variant PGS. (a) the *LPA* locus, (b) the *COL4A1/A2* locus.

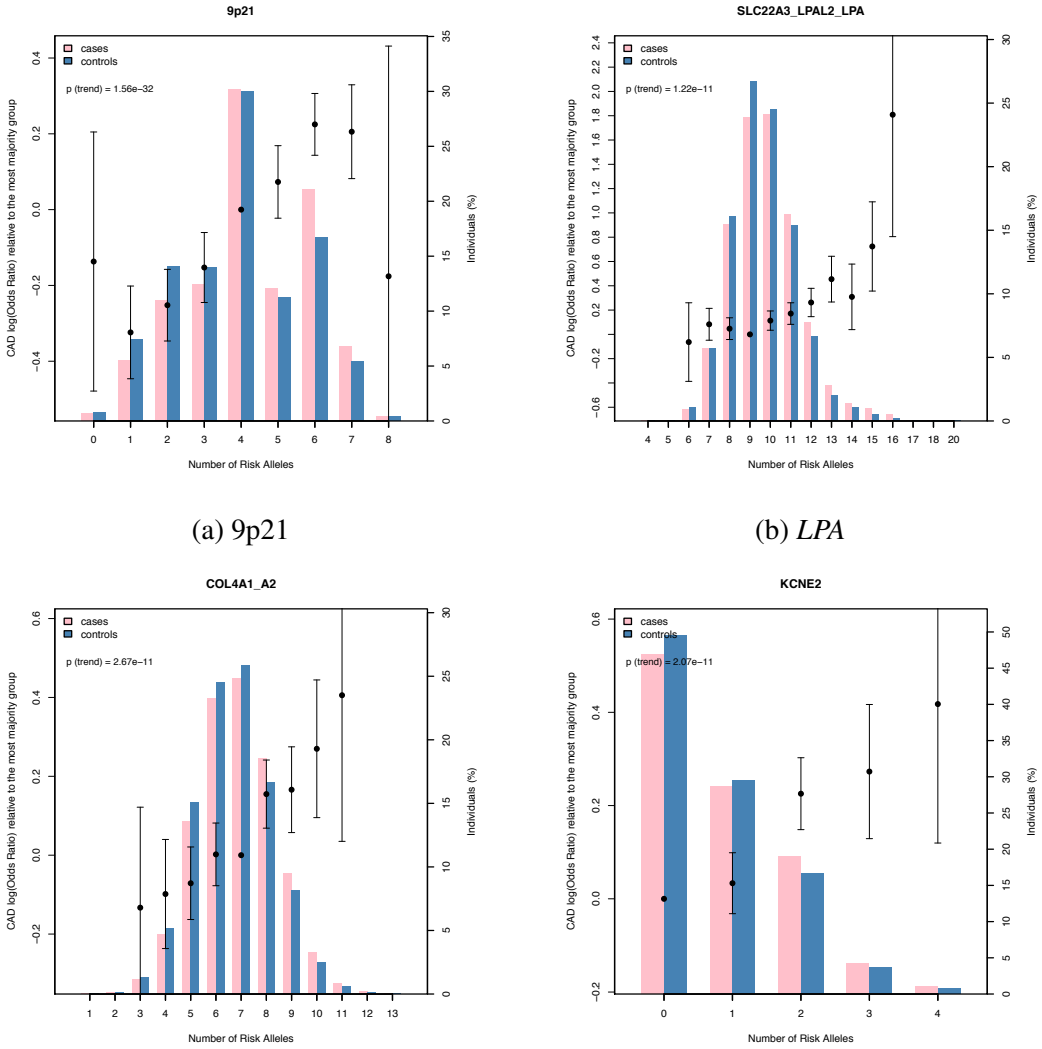
carrying the lowest number of risk alleles, the relative odds ratio of CAD ranged from 0.93 (nRA=6) to 2.06 (nRA=15) at the *LPA* locus, from 0.87 (nRA=3) to 1.50 (nRA=11) at the *COL4A1/A2* locus, from 1 (nRA=0) to 1.52 (nRA=4) at the *KCNE2* (gene desert) locus and from 0.87 (nRA=0) to 1.23 (nRA=7) at the 9p21 locus, as shown in Figure 3.5.

3.2.2 Polygenic score of other traits in relation to CAD

Inverse genetic association between height and CAD

Study purpose. The following description refers to results that were generated in the multi-locus genetic risk score analysis between height and CAD, as part of the efforts contributed to the publication Nelson et al [75].

I applied a weighted multi-locus polygenic score to represent the genetic risk for height based on the reported height increasing variants from GWAS studies [57]. Subsequently all individuals were categorized into quartiles corresponding to four grades of the genetic increase in height. Table 3.4 shows a graded trend ($p = 1.14 \times 10^{-11}$) between



(a) 9p21

(b) *LPA*

(c) *COL4A1/A2*

(d) *KCNE2* (gene desert)

Figure 3.5: Incremental increase of risk with additive load of risk alleles. Barplot: Relative frequency distributions of number of risk alleles in CAD patients and controls, respectively. CI-lines: Relative CAD log(odds ratios) and 95% confidence intervals for individuals in each group with the incremental increases of number of risk alleles in comparison with the most majority group as the reference group. P-value: Cochrane-Armitage test for increase trend of the odds ratio. (a) the 9p21 locus, (b) the *LPA* locus, (c) the *COL4A1/A2* locus, (d) the *KCNE2* (gene desert) locus.

the presence of an increased height PGS and a reduced ratio of CAD cases in the five GerMIF studies.

Study Abbr.	Quartiles of Height PGS	Q1	Q2	Q3	Q4
G1	N cases (CAD)	111	151	162	208
	N controls (CAD)	445	405	394	349
G2	N cases (CAD)	324	307	304	267
	N controls (CAD)	295	311	314	352
G3	N cases (CAD)	356	272	253	177
	N controls (CAD)	273	356	376	452
G4	N cases (CAD)	200	246	248	269
	N controls (CAD)	325	279	277	257
G5	N cases (CAD)	708	641	610	500
	N controls (CAD)	310	376	407	518

Table 3.4: Number of individuals in each category of PGS quartiles for height

These results were sent to our collaborator Christopher Nelson at University of Leicester, where the genetic score based on WTCCC cohort was integrated. Afterwards, regression analysis was performed to model the multi-locus polygenic scores for height to estimate the combined odds ratios for CAD, which have been published in the process of my thesis work [75]. An inverse association between height and CAD was confirmed, with odds ratio 0.74; 95% CI, 0.68 to 0.84 for height quartile 4 versus quartile 1 ($P < 0.001$) [75]. Figure 3.6 is taken from the original publication, which presents the results of association analysis performed by Nelson et al based on the integrated individuals of both our five GerMIFS cohorts and the WTCCC cohorts (Figure 3.6).

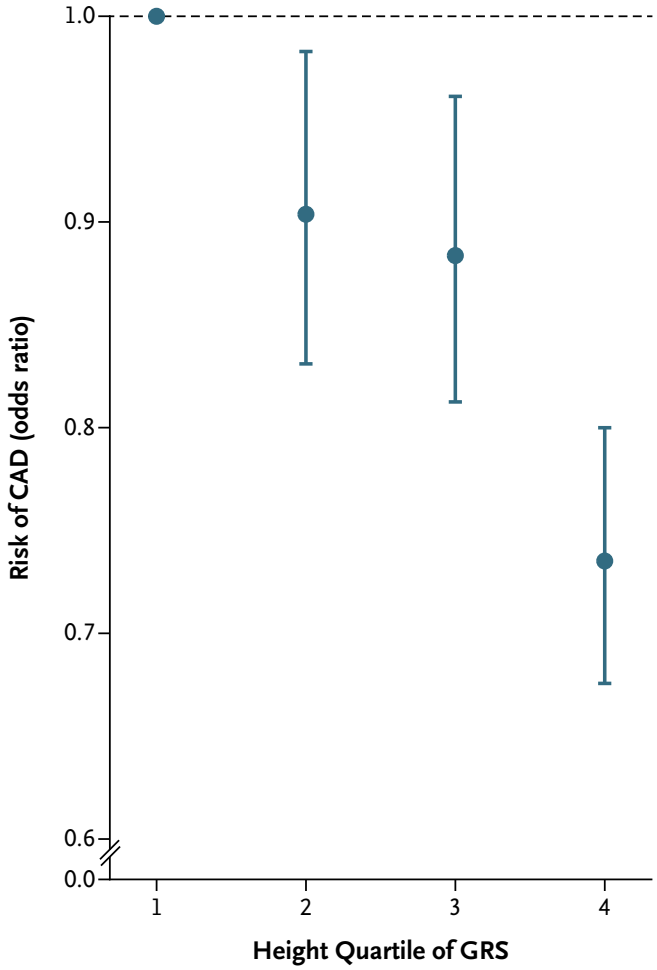


Figure 3.6: (taken directly from the original publication Nelson et al [75]). Inverse genetic association between height and CAD. Analysis of the association between the presence of an increasing number of height-related alleles and the risk of CAD, according to quartile of genetic risk score. Points and Lines: odds ratios and 95% confidence intervals. Quartile 1 (reference) were participants carrying the fewest number of height-increasing alleles.

No association between genetic risk variants of rheumatoid arthritis and CAD

Study purpose. The following description refers to results that were further used in the multi-locus genetic risk score analysis between rheumatoid arthritis (RA) and CAD, as part of the efforts contributed to the manuscript [46].

I applied a weighted multi-locus polygenic score to represent the genetic risk for RA, based on all genetic variants known to affect RA risk [29]. No difference was noticed between the RA-PGS in CAD cases and controls ($p = 0.26$). Then, all individuals were categorized into tertiles corresponding to three grades of genetic risk for RA, and logistic regression was performed to estimate the association between CAD and tertiles of multi-locus RA PGS. A non-significant result was obtained ($p = 0.21$) (Table 3.5 , Figure 3.7). Table 3.5 shows the number of individuals in each category of PGS tertiles for RA, the median number of RA risk alleles in each category as well as the standard error. An odds ratio relative to the lowest RA risk groups was also calculated. No specific trend was noticed comparing CAD cases and controls (Figure 3.7).

Study		**Tertiles of RA PGS**		
		Q1	**Q2**	**Q3**
G1	N cases (CAD)	183	221	218
	N controls(CAD)	541	503	507
	N Risk Alleles (RA) Median	50	54	57
	N Risk Alleles (RA) Se	0.098	0.075	0.1
	CAD log(OR) vs Q1	0	0.262	0.24
G2	N cases (CAD)	394	404	394
	N controls(CAD)	422	412	422
	N Risk Alleles (RA) Median	50	54	57
	N Risk Alleles (RA) Se	0.126	0.11	0.119
	CAD log(OR) vs Q1	0	0.049	0

Continued on next page

Table 3.5 – *Continued from last page*

Study		Tertiles of RA PGS		
		Q1	**Q2**	**Q3**
G3	N cases (CAD)	367	352	336
	N controls(CAD)	465	480	496
	N Risk Alleles (RA) Median	51	54	58
	N Risk Alleles (RA) Se	0.119	0.107	0.11
	CAD log(OR) vs Q1	0	-0.073	-0.153
G4	N cases (CAD)	311	322	321
	N controls(CAD)	386	374	376
	N Risk Alleles (RA) Median	51	54	58
	N Risk Alleles (RA) Se	0.134	0.113	0.127
	CAD log(OR) vs Q1	0	0.066	0.0578
G5	N cases (CAD)	796	804	837
	N controls(CAD)	541	533	500
	N Risk Alleles (RA) Median	50	54	58
	N Risk Alleles (RA) Se	0.096	0.089	0.094
	CAD log(OR) vs Q1	0	0.025	0.129
WT	N cases (CAD)	610	619	671
	N controls(CAD)	994	984	933
	N Risk Alleles (RA) Median	50	54	57.5
	N Risk Alleles (RA) Se	0.086	0.078	0.086
	CAD log(OR) vs Q1	0	0.025	0.16

Continued on next page

Table 3.5 – *Continued from last page*

Study		Tertiles of RA PGS		
		Q1	**Q2**	**Q3**
CG	N cases (CAD)	131	130	121
	N controls(CAD)	131	132	141
	N Risk Alleles (RA) Median	49	54	58
	N Risk Alleles (RA) Se	0.167	0.139	0.187
	CAD log(OR) vs Q1	0	-0.015	-0.153

Table 3.5: Number of individuals in each category of PGS tertile for RA

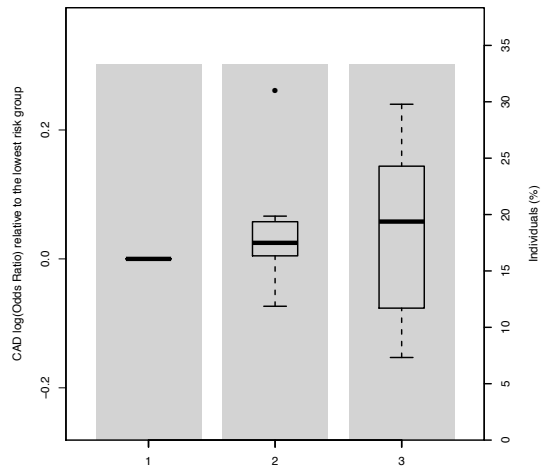


Figure 3.7: Odds ratio of CAD for each category of PGS tertiles for RA

3.3 Detecting epistasis that underlies CAD

Study purpose. The following description refers to methods that were generated in the epistasis analysis of CAD, as part of the efforts contributed to the manuscript [113].

3.3.1 Broad-sense of CAD susceptibility region

Knowing that multiple independent signals at the surrounding region of known CAD lead-SNPs could explain additional heritability to CAD, we aimed to define a broad-sense of CAD susceptibility region, on which basis we perform the epistasis analysis. For this purpose, we physically expanded progressively the region of the known CAD loci at a certain flanking range, so that growing numbers of variants within the expanded region could provide additional heritability. To decide a proper threshold for the flanking range, we calculated the narrow-sense heritability for all variants at the expanded step from $\pm$100kb to $\pm$1mb surrounding each CAD lead-SNP at the known risk loci. An incremental increase in the heritability could be observed with enlargement. However, a plateau was achieved at approximately $\pm$500kb surrounding the lead-SNPs (Figure 3.8). The additional increase was then more of moderate nature. Therefore, we decided on $\pm$500kb as the surrounding region to expand the searching space. Accordingly, all independent (LD-pruned $r^2 < 0.5$) variants within this region were picked for downstream analyses.

3.3.2 Significant trans-epistatic SNP pair for CAD

We explored the statistical epistasis in two steps, as shown in Figure 2.1. The primary filtering step was to find potential epistasis candidates out of all independent (LD-pruned $r^2 < 0.5$) variants within the broad-sense CAD susceptibility regions (n = 8,068). And then in the main screening step we included all variants without LD-pruning for fine-mapping of the epistasis candidates.

As a result, there was one SNP-pair which passed filtering. The two SNPs, rs71524277 [T/C] on chromosome 7 (effect allele frequency (EAF) C = 0.07) and rs679958 [T/C] on chromosome 13 (EAF T = 0.27), showed a statistical significant trans-epistasis effect on CAD (meta-analysis p-value = 3.06×10^{-11}). Both SNPs were well-imputed in our

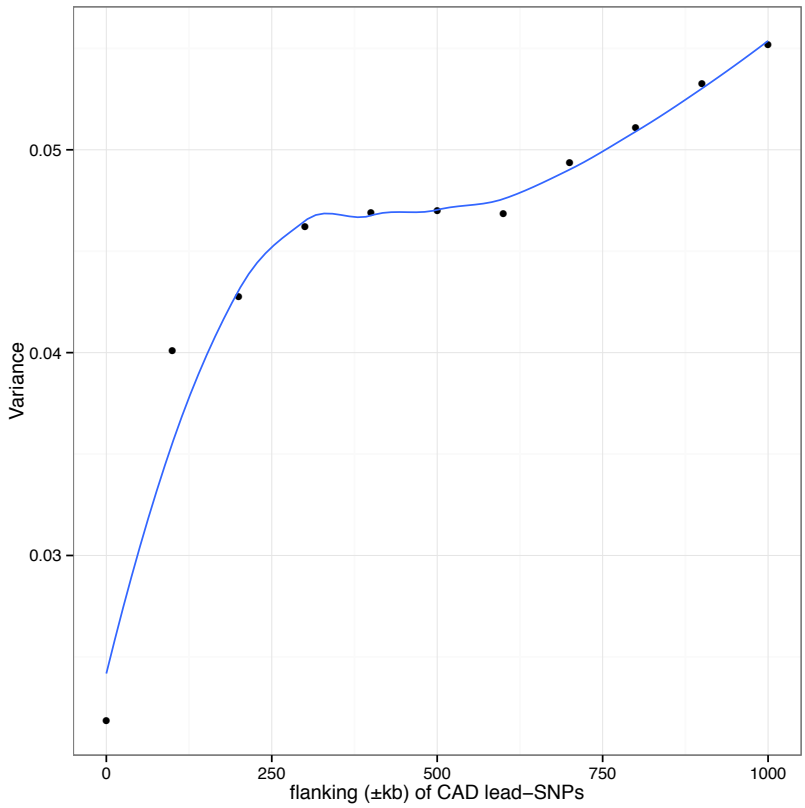


Figure 3.8: Increased variance explained with physically expansion for CAD lead-SNPs. All variants in the flanking region around the lead-variants at these 56 known loci were extracted from the merged imputed data for all studies. The LD-adjusted kinship matrix was calculated with LDAK [92] and forwarded to GCTA [111] to estimate the narrow-sense heritability of CAD in the measure of the total variance in liability (assuming the prevalence of CAD as 5%) explained by all variants together.

datasets, with mean INFO score of 0.95 (minimum of 0.866) for rs71524277 in all studies and mean INFO 0.96 (minumum 0.86) for rs679958.

The genetic epistasis model suggested both SNPs were heterozygous for the epistasis. Although the epistasis effect was also identified between rs71524277-C-dominant and rs679958-T-heterozygous (meta-analysis p-value = 1.35×10^{-10}), and between rs71524277-C-dosage and rs679958-T-heterozygous (meta-analysis p-value = 2.46×10^{-9}), but not as significant as the model for both SNPs being heterozygous. Figure 3.9 shows the study-wise effect for the epistasis SNP pair in a forest plot, where a very consistent effect across all studies was observed, with the meta-analysis effect for the interaction term *beta* as 0.58[0.41,0.75].

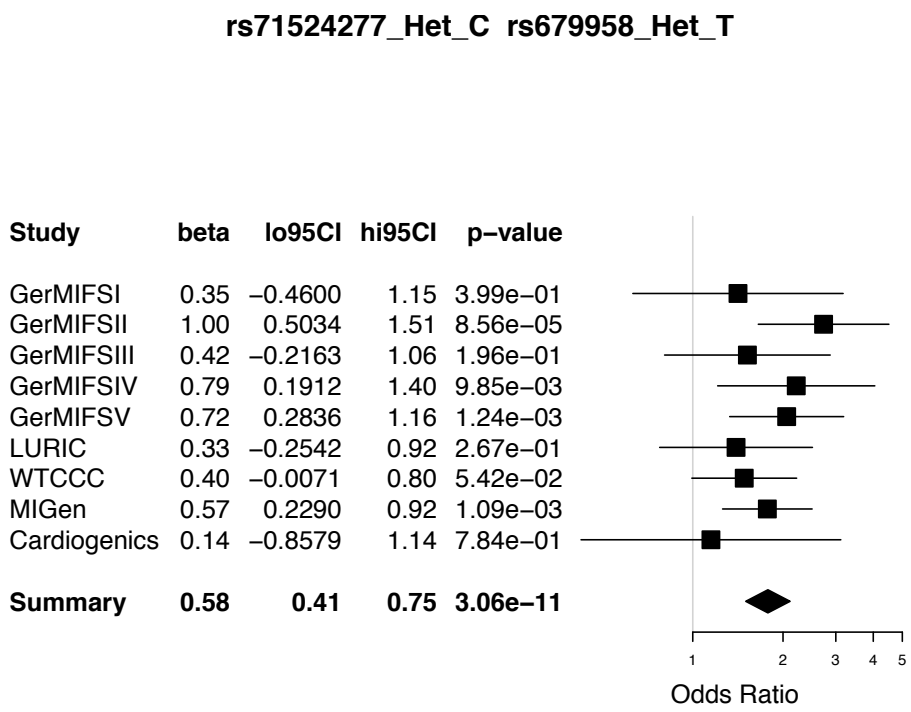
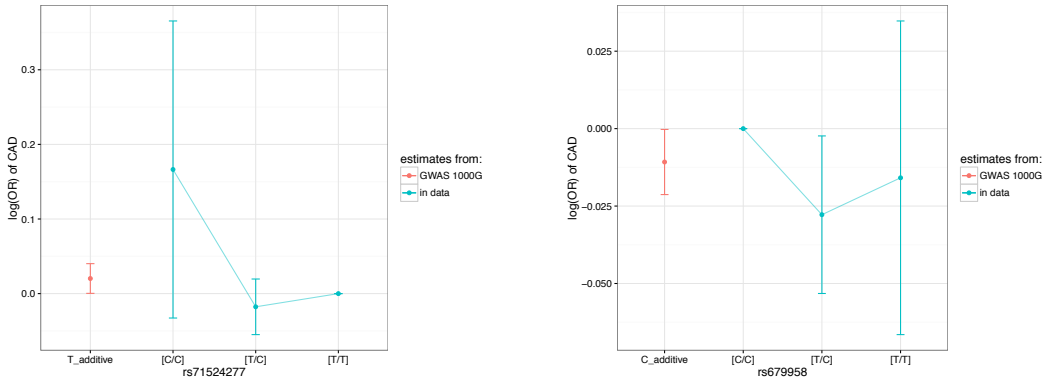


Figure 3.9: Effects of the epistatic SNP-pair across studies.

Rs71524277 is located in the intergenic region between gene *FERD3L* and *TWISTNB*. The nearest known CAD lead-SNP is rs2023938 in *HDAC9*, but the two SNPs are

independent regarding LD ($r^2 < 0.0125$). Rs679958 is located in the intronic region of *COL4A1*. The nearest but also LD-independent ($r^2 < 0.0035$) CAD lead-SNP is rs4773144 located in the intronic region of *COL4A2*. Conditional logistic regression (Eq. 2.3) adjusting the effect of these two lead-SNPs still gave a significant epistasis effect (conditional analysis p-value = 6.05×10^{-11}). According to the summary statistics from the 1000G CAD GWAS [77], the two SNPs themselves do not have a univariate association signal in an additive model, with a *beta* of -0.02 and p-value of 0.3 for rs71524277-C (Figure 3.10a), and *beta* of 0.01 and p-value of 0.3 for rs679958-T (Figure 3.10b).



(a) (b)

Figure 3.10: Relative log (odds ratio) for genotypes of each single SNP in the epistasis pair. Red: The log odds ratio (*beta*) with standard error (*se*) error-bars based on the reported summary statistics from the 1000G CAD GWAS assuming the additive model for rs71524277 T in (a) and rs679958 C in (b) [77]. Blue: The log odds ratio (*beta*) with standard error (*se*) error-bars based on all individuals used in the epistasis analysis for each of the 3 genotypes. Reference group: rs71524277 T/T in (a) and rs679958 C/C in (b).

Figure 3.11 displays the relative effect for the 9 combinations of genotypes for the epistasis pair, where the odds ratio of CAD for the majority group, rs71524277 T/T and

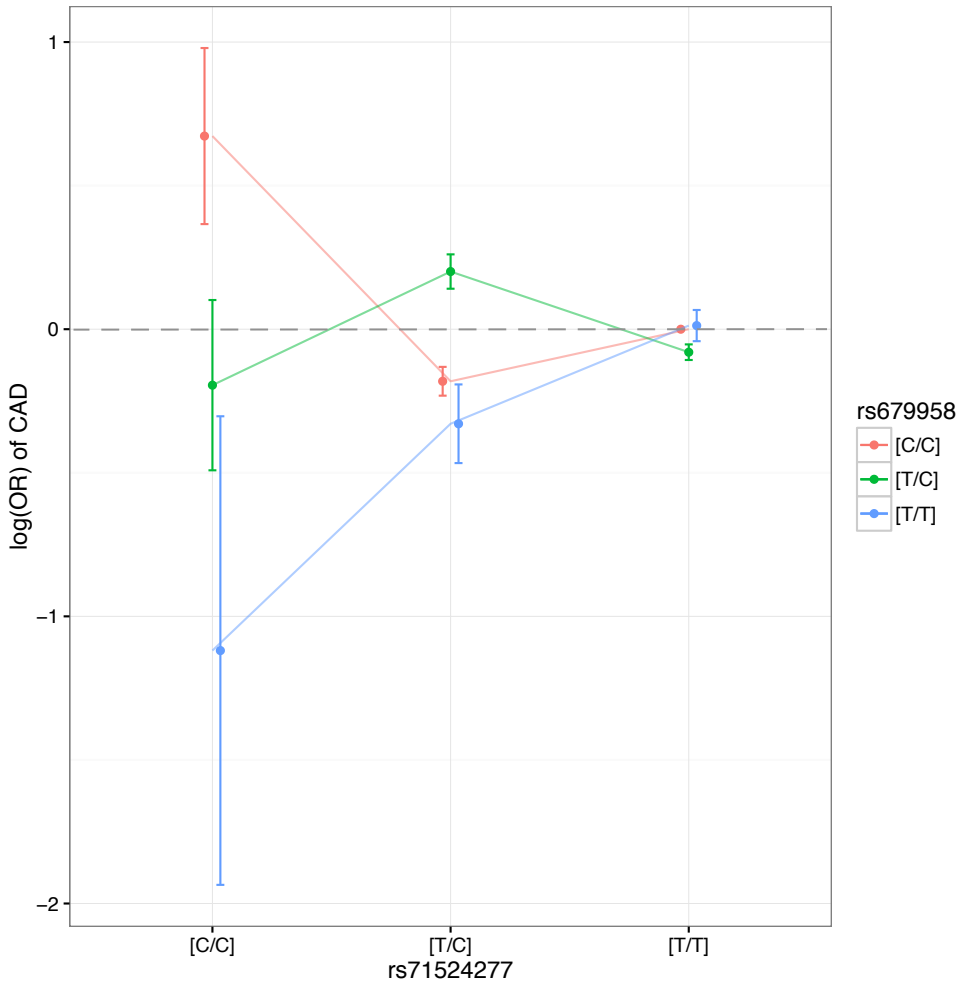


Figure 3.11: Relative log (odds ratio) for the 9 combinations of genotypes for the epistasis pair. The log odds ratio (*beta*) with standard error (*se*) error-bars based on all individuals available in the epistasis analysis for each of the 9 genotype combinations are displayed. Reference group: rs71524277 T/T and rs679958 C/C.

rs679958 C/C, is set as the baseline for comparison. A cross over of the effect trend could be observed when both SNPs displayed a heterozygous genotype, which is the deviation from expected and was the interaction item being tested.

3.3.3 Potential genes intermediate between epistasis pair and CAD

In order to interpret the identified statistical epistasis, we examined whether the same epistasis effects could be recovered on the expression level of some genes, which might play an intermediate role in modulating CAD risk.

For this purpose we utilized the parallel genotype and gene expression data from the Cardiogenics study (see Methods 2.1 for cohort description), where gene expression data from monocytes and macrophages was available for part of the genotyped individuals. The gene expression datasets were pre-processed and underwent QC by our collaborator, Veronica Codoni, at INSERM. After adjustment for batch effect and normalization, 15,539 probes in 684 macrophage samples and 849 monocyte samples were available for gene expression analysis. The numbers of individuals with both genotype data and gene expression data available are shown in Table 3.6.

	Monocytes	**Macrophages**
CAD cases	354	303
CAD-free controls	389	301
All	743	604

Table 3.6: Number of individuals with both genotype and gene expression data in Cardiogenics

First I looked at the univariate eQTL effect of single SNPs. A previous eQTL analysis has been performed by the INSERM institute for both cis and trans eQTLs

in these two tissues, from which we have requested the summary statistics through the Leducq Consortium CADgenomics. Rs71524277 did not show a single eQTL effect. For rs679958, one trans-eQTL effect with the gene SP3 in macrophages exists ($p = 1.76 \times 10^{-6}$), but with a FDR of 0.52.

Therefore, we performed a specific statistical epistasis test to identify genes affected by the epistasis pair. For gene expression in either monocytes or macrophages and sample groups with any status of CAD cases (Table 3.6), linear regression was performed (Eq. 2.4), followed by ANOVA test (Eq. 2.5) to analyze if any gene expression level shows difference at different genotype combinations according to the epistasis pair. Furthermore, Pearson correlation tests were performed to identify genes whose expression level were also correlated with the putative CAD odds ratio. As a result, 111 genes were identified as significantly associated with the epistasis pair in any sample group (1000x permutation $p < 5 \times 10^{-3}$ for both linear regression and ANOVA, and $p < 5 \times 10^{-3}$ for correlation) .

Gene	Tissue	CAD	p (lm)	p.perm (lm)	p (anova)	p.perm (anova)	p.cor
TMEM176B	mac	cases	6.90E-05	0.002496	3.52E-05	0.002496	1.63E-06
TMEM176A	mac	cases	1.32E-04	0.001501	7.99E-05	0.001501	5.55E-06
ATP5F1	mon	all	2.46E-03	0.002496	2.34E-03	0.002496	1.20E-05
TMEM176A	mon	cases	1.36E-03	0.000581	8.92E-04	0.000581	1.24E-05
THAP10	mon	controls	1.69E-03	0.002496	8.33E-04	0.000581	1.38E-05
TMEM176B	mon	cases	1.55E-03	0.001501	1.08E-03	0.000581	3.73E-05
GTF2A2	mon	controls	3.27E-03	0.003496	1.88E-03	0.003496	9.27E-05
C20orf24	mon	all	1.01E-03	0.001501	6.98E-04	0.001501	1.02E-04
LARGE	mon	cases	2.11E-03	0.001501	1.52E-03	0.001501	1.15E-04
SIGLEC1	mon	controls	4.59E-03	0.003496	3.32E-03	0.001501	1.32E-04
RHOQ	mon	controls	1.19E-04	0.000581	7.60E-05	0.000581	1.52E-04
SPTBN2	mon	all	1.70E-03	0.000581	1.20E-03	0.000581	1.87E-04
UTP23	mon	cases	1.45E-03	0.001501	1.30E-03	0.001501	2.10E-04
NLRP3	mon	all	2.19E-03	0.001501	8.98E-04	0.000581	2.11E-04
SPTAN1	mon	cases	1.38E-04	0.000581	3.17E-04	0.000581	2.17E-04
PPP2R5A	mac	all	1.85E-06	0.000581	4.31E-06	0.000581	2.26E-04
PPP2R5A	mac	controls	1.54E-04	0.000581	1.22E-04	0.000581	2.28E-04
PDE8A	mac	all	3.40E-05	0.000581	1.20E-05	0.000581	2.30E-04
PGD	mon	all	5.96E-04	0.000581	3.44E-04	0.000581	2.57E-04
RARB	mon	cases	2.50E-03	0.001501	4.31E-03	0.001501	2.91E-04

Continued on next page

Table 3.7 – *Continued from last page*

Gene	Tissue	CAD	p (lm)	p.perm (lm)	p (anova)	p.perm (anova)	p.cor
SIGLEC10	mac	all	2.10E-04	0.000581	4.10E-04	0.000581	2.96E-04
PIBF1	mon	controls	1.37E-03	0.002496	2.25E-03	0.003496	3.05E-04
SDC1	mon	all	5.79E-04	0.000581	3.59E-04	0.000581	3.14E-04
SEC22C	mon	all	3.31E-03	0.002496	2.45E-03	0.002496	3.51E-04
PSRC1	mon	controls	1.98E-03	0.002496	8.46E-04	0.001501	3.56E-04
ZCRB1	mon	controls	5.13E-04	0.001501	3.46E-04	0.001501	3.57E-04
PLXNA3	mon	cases	3.85E-03	0.003496	4.41E-03	0.004495	3.83E-04
SSSCA1	mon	all	1.48E-03	0.002496	9.31E-04	0.001501	4.63E-04

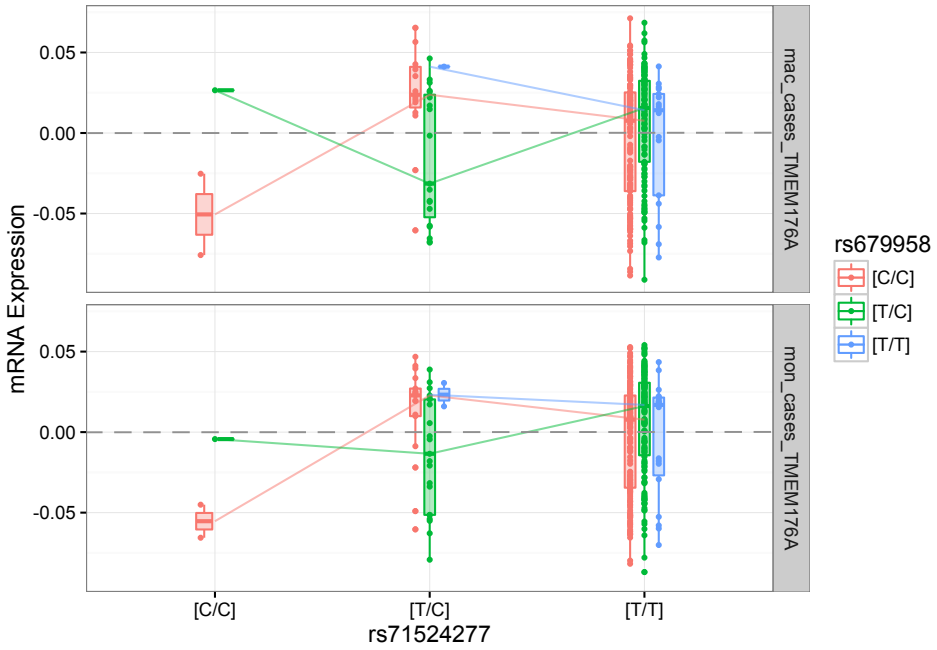
Table 3.7: Potential genes intermediate between epistasis pair and CAD. The tissues are either monocytes (mon) or macrophages (mac); the individual groups are CAD cases, controls, or both combined. P-values for linear regression (lm) and ANOVA test (anova) are listed both as original (p) and as permutation adjusted (p.perm); p-values for the correlation test (p.cor) $<5 \times 10^{-4}$ between the expression level and the putative CAD odds ratio are listed here.

Most of these 111 genes only showed association in specific tissues and sample groups. Only 14 of them displayed association (1000x permutation $p<0.05$) in at least three sample groups. Table 3.7 displays the list of associated genes with $p < 5 \times 10^{-4}$ for correlation between expression level and the estimated odds ratio of CAD deduced by the epistasis pair. The full list of 111 genes is available in Table A.1. However, no gene ontology, functional categories, or pathways could be enriched based on DAVID Bioinformatics Resources 6.8 [44] (Benjamini adjusted p-value > 0.05).

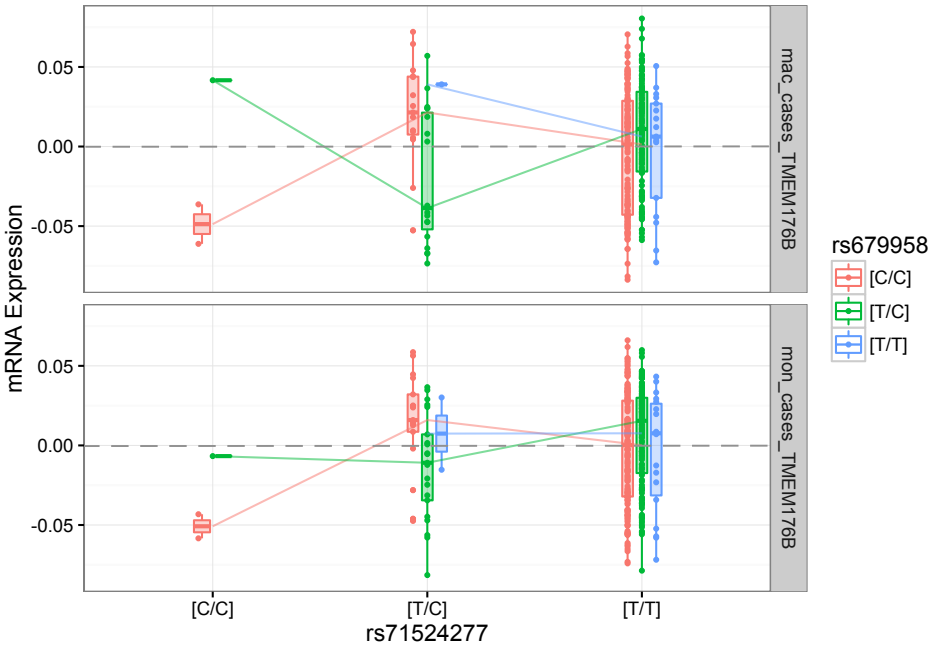
Nevertheless, two transmembrane protein coding genes, *TMEM176A* and *TMEM176B*, still caught our attention, as both of them showed repetitive significance (Table 3.8) based on expression epistasis results in CAD cases in both monocytes and macrophages. The expression levels were highly associated not only with the epistasis pair, but also with the putative CAD odds ratio (Figure 3.12). However, the association could not be recovered in CAD-free control samples.

Gene	Tissue	CAD	p (lm)	p.perm (lm)	p (anova)	p.perm (anova)	p.cor
TMEM176A	mon	cases	1.40E-03	0.00058	8.90E-04	0.00058	1.20E-05
		controls	8.00E-01	0.77373	7.60E-01	0.74475	5.80E-01
		all	5.40E-02	0.04346	3.50E-02	0.02248	7.30E-04
	mac	cases	1.30E-04	0.0015	8.00E-05	0.0015	5.60E-06
		controls	9.50E-01	0.94156	8.80E-01	0.87962	3.40E-01
		all	1.20E-02	0.00949	1.00E-02	0.00849	1.20E-04
TMEM176B	mon	cases	1.50E-03	0.0015	1.10E-03	0.00058	3.70E-05
		controls	9.20E-01	0.91858	8.90E-01	0.87363	4.70E-01
		all	4.70E-02	0.04446	3.00E-02	0.02947	8.70E-04
	mac	cases	6.90E-05	0.0025	3.50E-05	0.0025	1.60E-06
		controls	9.80E-01	0.97452	8.40E-01	0.83167	3.50E-01
		all	7.50E-03	0.00849	5.90E-03	0.00449	6.10E-05

Table 3.8: Potential genes *TMEM176A* and *TMEM176B* intermediate between epistasis pair and CAD. The tissues are either monocytes(mon) or macrophages(mac); the individual groups are CAD cases, controls, or both combined. P-values for linear regression and ANOVA test are listed both as original and as permutation adjusted; p-values for the correlation test between the expression level and the putative CAD odds ratio are also listed.



(a) *TMEM176A*



(b) *TMEM176B*

Figure 3.12: *TMEM176A/B* expressions at different genotype combinations based on the epistasis SNP-pair (rs71524277-C-heterozygous and rs679958-T-heterozygous). In either monocytes or macrophages from CAD case individuals, the expression of *TMEM176A/B* is correlated with the putative CAD odds ratio according to the groups of epistatic genotypes (which is displayed in Figure 3.11).

3.3.4 No epistasis effect on other CAD-related traits

In order to investigate the possibility that the epistasis pair conveys its association with CAD via other CAD-related traits, we performed a similar statistical epistasis tests on the expression of body-mass index, total cholesterol, LDL cholesterol, HDL cholesterol, and triglycerides. Information on these traits was available in 7,932 individuals, 579 from GerMIFS-I, 791 from GerMIFS-II, 3,252 from GerMIFS-V, and 3,310 from LURIC. However, none of them have reached a marginal significance level (Table 3.9).

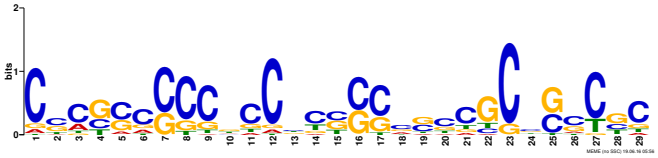
Study Abbr.	**BMI**	**Cholesterol**	**LDL**	**HDL**	**log(TG)**
G1	0.21	0.67	0.58	0.3	0.39
G2	0.5	0.44	0.44	0.92	0.78
G5	0.89	0.17	0.45	0.66	0.22
LU	0.31	0.24	0.19	0.47	0.26
meta-analysis	0.6	0.16	0.26	0.7	0.22

Table 3.9: P-value for the association between the epistasis SNP-pair and five CAD-related traits. Linear regression analysis in each study to test whether the CAD epistasis pair (rs71524277-C-heterozygous and rs679958-T-heterozygous) also shows an epistasis effect on any of the above five CAD-related traits. A combined p-value was calculated via a weighted Z-score based on sample size of each study.

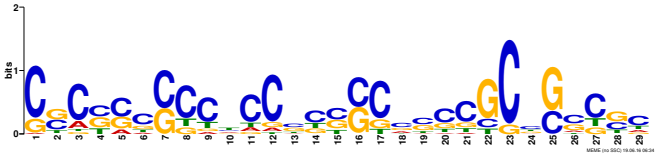
3.3.5 Motif enrichment with intermediate genes

The large number of genes likely to be affected in their expression by the epistasis pair led to the assumption that the genes with these expression patterns may be co-regulated by a common transcriptional factor. Therefore, we further investigated the putative promoter regions (TSS$\pm$2kb DNA sequences) for the 111 genes whose expression were

significantly associated with the epistasis pair. The DNA sequences for these 111 genes' promoters were compared to (i) normal mode: the random sequences simulated with same base-pair frequencies, and (ii) discriminative mode: the promoter sequences for all other genes available on the gene expression array (see Methods 2.5.10).



(a) Normal



(b) Discriminative

Figure 3.13: Enriched DNA motif sequence based on the potential genes intermediate between epistasis pair and CAD. (a) The top enriched motif resulting from normal mode of MEME enrichment analysis. (b) The top enriched motif resulting from discriminative mode of MEME enrichment analysis.

As a result, the normal mode of MEME enrichment analysis indicated one top significant GC-rich motif with MEME E-value at 1.5×10^{-106} (Figure 3.13a). The number of sequences contributing to the construction of this top motif was up to 73 (out of 105 putative promoter sequences in total). The best site for the enriched motif predicted by Centrimo was just around the center of our given sequences, which was actually the TSS site of our input sequences. This further supports the possibility of a common transcriptional factor binding for the epistasis pair (Figure 3.14). The most centrally enriched transcriptional factor binding matrix was predicted as SP2. With motif similarity searching for the top motif using TOMTOM, several motifs from C2H2 zinc finger transcription factors turned out to be in high similarity, headed with MA0516.1

(SP2), MA0162.2 (EGR1), MA0528.1 (ZNF263) and MA0079.3 (SP1) (Table 3.10).

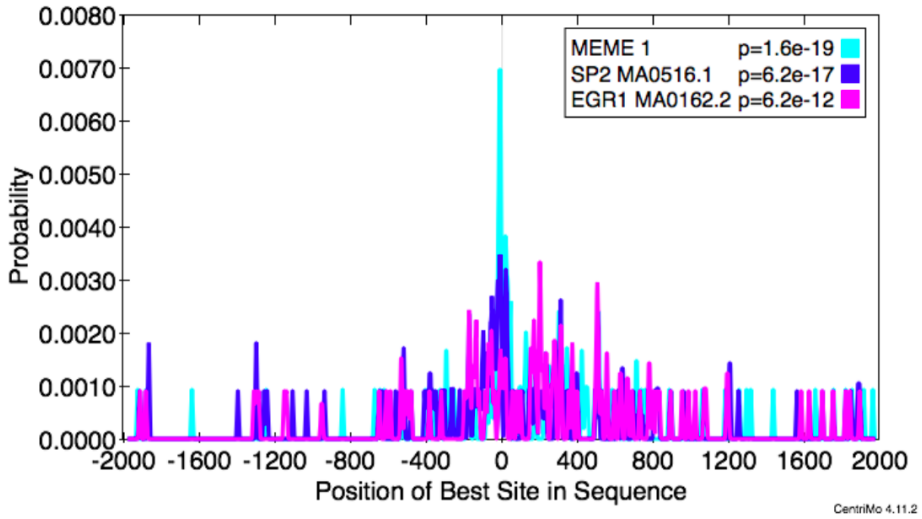


Figure 3.14: Position of the best sequence enriched site predicted by Centrimo. MEME 1 is the top enriched motif resultant from MEME (Figure 3.13a). The transcription factor binding motifs for SP2 (MA0516.1) and EGR1 (MA0162.2) are also displayed. The p-values represent the expected number of motifs that would have at least one region as enriched for best matches to the motif as the reported region.

The discriminative mode provided us with similar results. The top enriched motif deduced by MEME was also a GC-enrich sequence with an E-value=7.9×10^{-107} and the number of sequences contributing up to 72 (out of 105 sequences in total) (Figure 3.13b). TOMTOM predicted that the top enriched motif was highly similar to the binding motifs of several C2H2 zinc finger transcription factors such as MA0162.2 (EGR1), MA0516.1 (SP2) and MA0528.1 (ZNF263), followed by several AP2/ERF domain transcription factors, such as MA0992.1 (ERF4) and MA0975.1 (CRF2) (Table 3.10).

We took the top enriched motif from MEME and the aligned known motifs with E-value < 0.05 in both the normal and discriminative analyses as putatively confident co-regulators for the potential intermediate genes (Table 3.10) and searched for further evidence. Among these motifs MA0516.1 (SP2), MA0162.2 (EGR1) and MA0079.3 (SP1)

MEME mode	Top enriched motif for the 111 significantly associated genes	TOMTOM alignment to the known transcription factor binding motifs				
		Target ID	Putative TF (JASPAR)	p-value	E-value	Target consensus
Normal	CCCGCCCCCGCC CCCCCGCCGCCG CCGC	MA0516.1	SP2	3.72E-07	0.00019	GCCCCGCCCCCTCCC
		MA0162.2	EGR1	8.10E-07	0.00042	CCCCCGCCCCCGCC
		MA0528.1	ZNF263	1.30E-05	0.00674	TCCTCCTCCCCCTCCTCCTCC
	E-value = 1.5E-106	MA0079.3	SP1	2.97E-05	0.01543	GCCCCGCCCCC
Discriminative	CGCCCCCCCCTCCC CCCCCCCGCCG CCGC	MA0162.2	EGR1	9.94E-07	0.00108	CCCCCGCCCCCGCC
		MA0516.1	SP2	2.61E-06	0.00282	GCCCCGCCCCCTCCC
		MA0528.1	ZNF263	6.93E-06	0.0075	TCCTCCTCCCCCTCCTCCTCC
		MA0992.1	ERF4	1.76E-05	0.01903	CCGCCGCC
	E-value = 7.9E-107	MA0975.1	CRF2	2.62E-05	0.02839	CCGCCGCC

Table 3.10: Alignment of the top enriched DNA motif sequence with known transcriptional factor binding motifs

belong to the human transcription factor class *three-zinc finger Krüppel-related factors*, which is known for sequence-specific interactions with GC-rich promoter elements.

3.3.6 High nuclear lamina contacts at the rs71524277 genomic region

The nuclear lamina (NL) is a structure near the inner nuclear membrane and the peripheral chromatin [37]. The role of the nuclear lamina is not restricted to the maintenance of nuclear shape and structure. Through direct binding of the lamina to both, chromatin and a range of nuclear envelope proteins and transcription factors, it also play a role in regulating transcription, controlling differentiation and organizing chromatin [9].

In 2015 Kind et al. have developed a method based on DamID technology to map NL contacts genome-wide in single human cells. For each cell, they binned the whole-genome in 100-kb contiguous genomic segments and then calculated for each segment

an observed over expected (OE) score to indicate whether there is more Dam-LmnB1 methylation than may be expected randomly.

On a genome-wide scale, they reported that only about 15% of the segments have stable contact (CF >80%). We extracted the OE values for the loci where our epistatic SNP rs71524277 is located, i.e., chromosome 7 region [19200001,19300000](GRC37) and counted contact frequency (CF) based on the OE values as described by Kind et al. A cutoff of 1 was applied to the OE score to suggest that loci were either in a "contact" or "no-contact" state [52]. As a result, 88.98% of CF was obtained in this region for the 118 single cells in their experiments, suggesting the high possibility that this locus might be located at the NL.

Chapter 4

Discussion

4.1 Genome-wide association analysis for CAD in the 1000G era

Large-scale meta-analyses of GWAS for CAD had already successfully identified 46 CAD risk loci based on HapMap imputation training sets or tagging SNP arrays. With the update of the 1000 Genomes Project, which has considerably expanded the coverage of human genetic variation, a new avenue for GWAS analyses in CAD has been enabled. Aiming to uncover additional risk loci in the 1000G era, the CARDIoGRAMplusC4D consortium has led the effort to collect multiple studies and performed large-scale meta-analyses, including both the whole autosomal genome (1000G CAD GWAS) and the X-chromosome (1000G CAD X-Chr).

4.1.1 Autosomal GWAS

The meta-analysis of the 1000G CAD GWAS has been published in the process of my thesis work [77].

The 1000G CAD GWAS has taken advantage of the progress from the 1000 Genomes

Project, which provided a substantial upgrade of the genotype imputation panel in terms of coverage of lower-frequency variants and the integration of indel polymorphism. Ten novel CAD risk loci have been identified according to the final results of the meta-analysis [77]. The lead SNPs for four of the ten newly identified CAD loci were either absent or imperfectly tagged ($r^2 < 0.8$) in the HapMap 2 training set, which demonstrated that the power of GWAS to investigate the genetic architecture of complex traits is further enhanced by the 1000 Genomes Project.

The 1000G CAD GWAS meta-analysis included ∼185,000 CAD cases and controls, thus provided more statistical power to detect CAD susceptibility signals. For example, only based on the results from single GerMIF studies the signal at the 9p21 locus could also be recovered, but could not necessarily reach the genome-wide level of significance ($p < 5 \times 10^{-8}$), while in the meta-analysis it was shown to be highly significant and robust ($P < 3.8 \times 10^{-93}$). According to the post-hoc power calculation for the 1000G CAD GWAS meta-analysis, 93.3% of common variants (MAF>0.05) with OR>1.15, and 88.3% low frequency variants (0.005 < MAF < 0.05) with OR>1.5 could be detected with predicted power >90% at $p = 5 \times 10^{-8}$ [77], which strongly demonstrates the need for large-scale studies in genome-wide genetic association analyses.

The results of this comprehensive analysis strongly supported the common disease–common variant hypothesis. Despite the possibility and power to detect the associations for low-frequency variants owing to the 1000G reference genome and the large-scale meta-analysis, all ten newly identified CAD associations found in the present analysis, as well as all but one of the previously identified loci, were represented by risk alleles with a frequency of >5%. In addition, from the conditional and joint analysis where more suggestive susceptibility variants were identified, fifteen low-frequency (MAF < 0.05) variants explained only $2.1 \pm 0.2\%$ of CAD heritability, and all were either a lead variant or were jointly associated (q value < 0.05) with a common variant [77]. Our group conducted further analyses to see whether the common lead variants on the

array, or other, rare variants at these loci are responsible for the strongest signals of significance. Such explorations revealed that low-frequency variants do not explain a significant portion of the missing heritability at a population-based level [93].

4.1.2 X-chromosome

The meta-analysis of the 1000G CAD X-Chr has been completed and summarized in a manuscript by the CARDIoGRAMplusC4D consortium (Loley et al [59]).

Previously, little was known about the role of X-chromosomal variants in CAD, mainly due to the sex-specific data structure of chromosome X which requires special test statistics and statistical models other than the routine GWAS analyses. The meta-analysis included more than 43,000 CAD cases and 58,000 controls from 35 international study cohorts, which makes it, with the inclusion of more than 100,000 individuals, the to date largest study compared to all current GWAS reported on the X chromosome (<50,000 individuals, according to the NHGRI GWAS catalog [105] reports on more than 600 traits available). Regarding the statistical model, four possible models assuming any combination of X chromosome inactivation and sex interaction were investigated. Therefore, the 1000G CAD comprehensive X-chromosome meta-analysis by itself has provided a thorough investigation for the possible associations of genetic variants and CAD.

The meta-analysis revealed that no significant signals were identified at genome-wide significance level. Even stricter quality control and exclusion of non-European studies demonstrated that there is no univariate association between any genetic variants at X-chromosome and CAD susceptibility [59]. Although several variants reached the level of genome-wide significance in our single study (GerMIFS-V), the same signal was not captured anymore in the final meta-analysis [59]. This is probably due to limited sample size and sampling heterogeneity. The negative result could also be due to the suboptimal statistical model or biological assumptions. Nevertheless, considering the complex

of genetic architecture underlying complex diseases, it is also highly possible that X-chromosomal variants may affect CAD susceptibility in a non-univariate way, such as cis-epistasis between X-chromosomal variants, trans-epistasis between X-chromosomal and autosomal variants, or gene-environment interactions on the X chromosome, all of which have not been investigated yet in the context of CAD.

4.2 Understanding the genetic complexity of GWAS signals of CAD

4.2.1 Intra-locus allelic heterogeneity at known CAD loci

For complex diseases as CAD, it is important to refine known loci to comprehensively identify and understand the genetic architecture. The authors of the 1000G CAD GWAS meta-analysis [77] performed an approximate conditional and joint association analysis based on the suggestive additive association region surrounding (2cM) the variants with $p < 5 \times 10^{-5}$. As a result, 95 variants (explaining $13.3 \pm 0.4\%$ of CAD heritability) mapping to 44 significant CAD loci identified by GWAS were detected suggesting that multiple independent signals could be recovered at the surrounding regions of known CAD loci.

Here, based on individual-level genotype data from eight CAD case-control cohorts, we applied a weighted intra-locus polygenic scores (PGS) approach to estimate and compare the combined effect of multiple variants versus that of the lead variant only. The weight we used was the effect size reported from the conditional and joint analysis, with the assumption that the effect could be estimated more precisely due to large sample sizes. Weighted scores may increase statistical power compared to unweighted scores, provided that the weights are accurately determined [8]. The effects we reported for the PGS were per standard deviation, so that the scales of effects could be comparable

between the multi-variant and lead-variant PGS. While on the other hand, the odds ratio for PGS was then not easy to interpret as it was not comparable with the odds ratio for the risk allele (which is reported by most GWAS the odds ratio for the additive genetic model).

We observed that larger effect and more variance explained could be recovered on our individual-level genotypes for two-thirds of the loci identified as known loci harboring multiple independent signals for multi-variant PGS than that for lead-variant PGS. However, the improvements of fit for additional variance explained at single loci were all very slight. We also observed that four loci presented genome-wide significant combined effect for multi-variant PGS. These results supported the overall combined effects of multiple signals at single loci.

However, for one-third of the variants no improvement of fit for multi-variant PGS, compared to the lead-variant PGS, was detected, not even for the 9p21 locus, which had a genome-wide significant effect for multi-variant PGS. One reason could be the inaccurate estimate of the effect in the joint analysis, which was performed based on the summary statistics of the GWAS meta-analyses with the reference genome to deduce the LD structure of the whole genotype data. Thus, there could be sampling genetic heterogeneity between our genotype datasets and the datasets included in the GWAS meta-analysis. Additionally, there could be multicollinearity between multiple variants. Indeed, when we looked for the effect of each single variant in the additive univariate GWAS for 9p21, one of the SNPs (rs7855162) had a negative effect ($\beta = -0.04$) for the risk allele according to the joint analysis ($\beta = 0.16$), which supports the possibility of collinearity.

Regarding predictive power for CAD based on PGS, we demonstrated that even for the top two loci (i.e., *SLC22A3/LPAL2/LPA* and *COL4A1/A2*, which not only achieved genome-wide significance but also showed a much larger additional R^2 compared to other loci), no significant improvement in prediction could be achieved. This further

implies that although the conditional and joint analysis could provide additional variants independent in the sense of conditional analysis, predictions based on variants at single loci still require effort for better prioritization for both, the variants and the prediction model.

Despite the observation that the PGS for a single locus could not present much total predictive power for CAD, even with combined effects of multiple signals, the observation that multi-variant PGS presented per SD higher odds ratio as well as more variance explained than the lead-variant PGS still supports the existence of intra-locus allelic heterogeneity. For the four loci showing the largest multi-variant PGS effects we observed that incremental increases in the odds ratios were accompanied by increasing numbers of risk alleles, which further supports the existence of intra-locus allelic heterogeneity. However, closer investigation at each locus is needed to provide more knowledge about the functional mechanisms influencing the genetic etiology of CAD.

The current conditional and joint analyses by the GWAS meta-analysis were performed only based on summary statistics, with the reference panel taken into account to derive the LD structure [77]. For many of the known CAD loci, the density of variation probes of the traditional GWAS array is not as high as the Metabochip genotyping array, which may affect the power to identify all the independently associated signals at single loci. This raises the hope that more intra-locus allelic heterogeneity probably could be uncovered with the coming results of fine-mapping GWAS meta-analysis.

4.2.2 Multi-locus pleiotropy

It has been shown that there is a large degree of polygenic overlap between CAD and cardiovascular risk factors, which underlines the shared polygeneticity of these phenotypes and also promotes our understanding of genetic background underlying CAD [58]. Accumulating information about multiple SNPs with a small effect into a single genetic risk score, has become a useful tool to examine the cumulative predictive

ability of genetic variation at known loci on cardiovascular disease outcome and other related phenotypes.

Risk prediction based on common genetic variation has gained widespread attention in the last years. Multi-locus polygenic scores are usually constructed to estimate the combined effect in order to investigate the correlation between different diseases or traits and to explore novel shared or unshared genetic information.

Here, we applied the multi-locus polygenic score approach to test possible associations between the risk of CAD and adult height as well as rheumatoid arthritis (RA). Multi-locus polygenic scores were constructed combining the effects of height-increasing alleles and RA risk alleles, respectively. After modeling these multi-locus polygenic scores into logistic regression to estimate combined odds ratios for CAD, an interesting inverse genetic association was noticed between height and CAD. However, no association between RA and CAD was detected. These results provide better understanding of the shared and unshared genetic background between CAD and height, and the unshared genetic background between CAD and RA.

It is noteworthy that the results from the polygenic score analysis do not disclose the causal role of these risk factors. The variants used in the score calculation were picked from the literature based on results from corresponding GWAS analyses. However, the variants identified by GWAS are not necessarily causal but can nevertheless be in in LD with one or more causal variants. Regarding complex diseases, it is also likely that the so far identified genetic loci from GWAS do not cover the complete genetic spectrum. In addition, the clinical utility of polygenic scores need further validation. As the individual samples in the original GWAS analyses could be largely different compared to our test samples, the effect sizes estimated in the original analyses might lead to a bias for the polygenic score prediction in the test datasets.

4.3 Detecting epistasis that underlies CAD

It is now widely accepted that multiple genes influence many complex diseases such as CAD. The exploration of epistasis in CAD, however, is still at the beginning. The aim of this work was to make a first large-scale exploration not only of the statistical epistasis of CAD but also to narrow down the gap between statistical epistasis and biological interpretation. In line with this, we also aimed to establish a workflow for further epistasis research in the future.

The GPU-based GLIDE epistasis computation tool helped us to handle the computation burden of epistasis which is challenging using normal CPU-based computing clusters. With the aim to scan a rather large genomic region without losing the high-potential susceptibility region for a pilot study, we started our analysis searching for pair-wise epistatis effects in a broad-sense of CAD susceptibility regions. The searching space of our current study set was not restricted to the known lead-SNPs, as most of the included variants actually did not show significant associations in GWAS and were not in LD with the known CAD lead-SNPs. Nevertheless, the searching space was still restricted to certain regions of the genome. Thus, our results can only describe a part of CAD epistasis. Further expansion of a similar pipeline to a genome-wide search scale, as well as including the search for all three-way or four-way or any higher level interactions, may finally help us identifying a landscape of epistasis activity in CAD genetics.

Traditional GWAS, or so-far reported, epistasis analyses have mainly focused on additive, dosage, or dominant models for genetic variations. The heterozygous or recessive model has often been ignored. Here, we made a thorough search for the $4(dosage, dominant, heterozygou, recessive) \times 4$ genetic model combinations for a pair of variants. For statistical epistasis we applied a two-step approach with the aim to both, enhance the speed and the chance of positive finding: In the first step we tried to discover all possible candidate pairs in high speed by taking the advantage of the GLIDE software. Only LD-independent variants were tested and no covariates were included in the model,

we just applied a loose significant threshold for primary filtering. In the second step we performed a careful screening with the aim to assure the positive finding. Therefore, all variants in the full LD block were tested and up to ten genetic covariates were included in the model. A Bonferroni significance level was also carefully calculated for multiple testing correction.

The significant epistasis pair we identified is composed of one variant in an intergenic region (rs71524277) and the other variant in an intronic region (rs679958). This also highlights the important role of the non-coding parts of the genome, which are also the major source of the thus far by traditional GWA studies identified common genetic variants [26].

In order to interpret the identified epistasis pair, we went on to analyze gene expression and clinical trait data to elucidate the disease mechanism. However, we did not find significant association between the epistasis pair and other CAD-related traits, e.g., body-mass index, total cholesterol, LDL cholesterol, HDL cholesterol, and triglycerides. This indicated that the identified epistasis pair could be directly associated with CAD rather than intermediate with CAD-related traits.

On the other hand, we could identify 111 genes, significantly associated with both, the epistasis pair and the odds ratio for CAD. However, no enriched functional categories could be identified only with this group of genes, and these genes presented significant association almost exclusively in specific tissue-phenotype sample groups. The genes presented the largest cross-tissue-phenotype consistency for the epistasis pair were *TMEM176B* and *TMEM176A*, in their expressions in monocytes and macrophages of CAD cases. These two genes encode membrane proteins associated with the immature state of dendritic cells, which play a central role in the induction and maintenance of immune tolerance [14]. Furthermore they were shown to have a similar mRNA expression patterns among various murine tissues [14]. The inconsistent association across different tissue and disease status for the 111 genes implies the possibility that

they could be co-regulated via a common upstream regulator, which plays a cellular context-specific regulatory role in cells and CAD disease status.

The non-enrichment of functions of these genes also implies the possibility that the monocytes and macrophages that we investigated may not be the major tissue where the upstream regulator conducts its functional role in the pathological pathway. The gene expression dataset we used in our analysis had a small size (less than 800 samples) and was also tissue restricted (monocytes and macrophages), which by itself is a limitation. Larger sample sizes as well as gene expression data from multiple tissues might provide the possibility of replication and validation of the functional role of the genes that are associated with the epistasis pair and CAD.

Interestingly, these genes, which were significantly associated with both the epistasis pair and the odds ratio for CAD, were highly enriched in GC-rich motif patterns in their putative promoter regions based on two different motif enrichment computations. The corresponding TFs (SP2, EGR1) predicted to bind to these motifs repetitively occurred in both computations. Indeed, SP2, EGR1 and SP1 have been frequently reported to share overlapping promoter binding sites [2, 4, 100, 114]. Interestingly, the only trans-eQTL gene to SNP rs679958 in macrophages, *SP3*, although with high FDR, also belongs to the Sp-family. Regarding promoter binding activity, Sp2 has been reported to carry the least conserved DNA-binding domain among Sp-family members and binds poorly to a subset of target DNA sequences bound by other family members and has little or no capacity to stimulate transcription of promoters that are potently activated by Sp1 or Sp3 [72]. Sp1, Sp3, and Egr1 have been reported to compete for their DNA-binding sites [43, 72].

Testing of statistical epistasis rather than biology epistasis has the advantage that even if two molecules are not physically interacting with each other, it could also be captured in the statistical epistasis [25]. Indeed, in molecular biology it happens that two physically irrelevant proteins may be involved in the same pathway, cause different but

dependent cascading events or regulate different but context-dependent pathways, thus together affecting disease susceptibility. All these possibilities could be hypothesized from statistical epistasis but would probably be missed from biological epistasis analysis.

We hypothesize that the epistatic pair (rs71524277-rs679958) may perturb the function or regulation of the transcription factor Sp-family and EGR1, thus further causing the corresponding change in the gene expression for a group of genes.

The observation that the putative Sp-family transcription factors and EGR1 themselves are not among the 111 genes, thus showing statistically intermediate associations between the epistasis pair and CAD, led to the thought that dysregulation of Sp proteins due to the epistasis pair is not caused by their disruption at mRNA level but rather at the protein level, e.g., alteration of protein structure, post-translation modification, or altered domain-binding activity affected by cellular microenvironment.

Mammalian chromosomes are spatially organized inside interphase nuclei, the contact frequency with the NL is locus specific. Only about 15% of the human genome have stable contact (CF>80%) with NL, which are usually extremely gene poor and suggesting a structural role [52]. The rs71524277 flanking genomic region has high contact frequency (88.98% of CF, which was much higher than the 80% threshold for stable contact given by Kind et al [52]) with the nuclear lamina. Besides, this variant itself is located in the intergenic region, suggesting the high possibility that the variant and its flanking region could be NL-linked.

SP1, SP2, and SP3 are nuclear proteins, by their subcellular localization. Sp2 preferentially localizes to subnuclear foci associated with the nuclear matrix and it has been speculated that this subcellular localization plays an important role in the regulation of Sp2 function [72]. Furthermore, both, SP1 and SP3 are bound to the nuclear matrix with different nuclear matrix-associated sites [38]. We therefore hypothesize that the epistatic pair (rs71524277-rs679958) may convey its genetic effect through the perturbation of the interplay between nuclear matrix proteins and the nuclear lamina

(Figure 4.1).

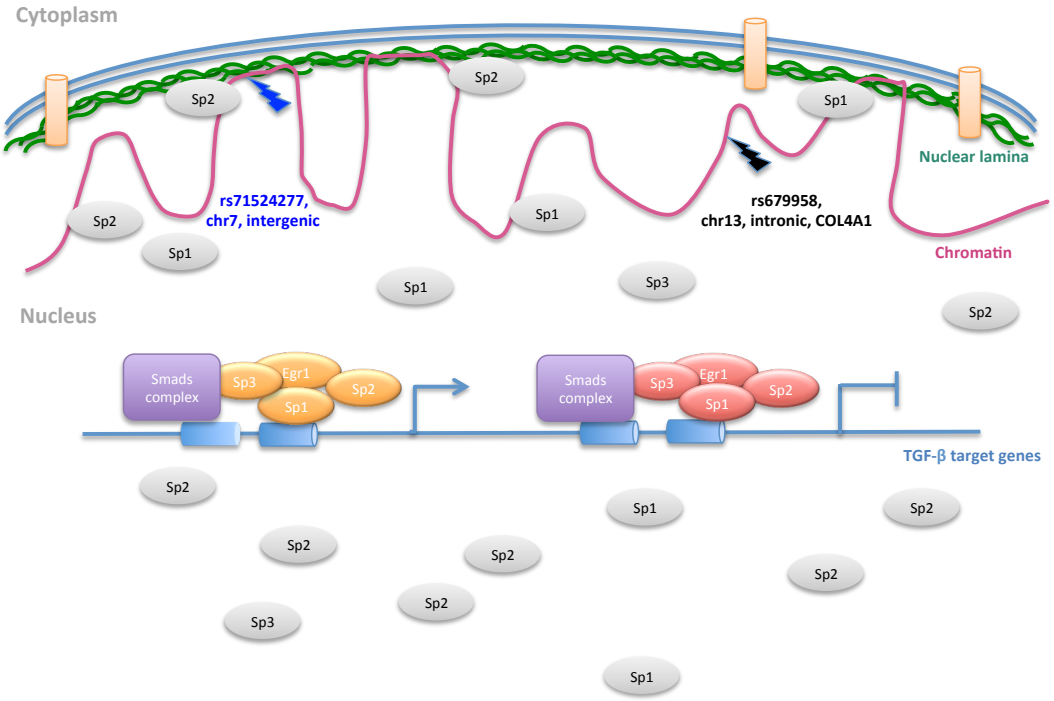


Figure 4.1: Hypothesized mechanism for the epistasis pair underlying CAD. The epistasis pair (rs71524277-rs679958) may convey its genetic effect first through the perturbation of the interplay between nuclear matrix proteins and nuclear lamina and then through the downstream perturbation on the TGF-β signaling pathway.

Interestingly, Sp-family transcription factors as well as EGR1 are all tightly linked to the transforming growth factor beta (TGF-β) - SMAD signaling pathway, which is associated with a series of cardiovascular diseases, including atherosclerosis, hypertension, restenosis and heart failure [6, 7, 36, 87]. SP1 has been reported to cooperate with SMAD complexes to mediate the TGF-β signaling in different cell types [30, 47, 84]. SP1 and SP3 together have been reported to collaborate with SMAD3 in the TGF-β signaling pathway [19]. Furthermore EGR1 has been reported to interact physically and

functionally with SMAD3 in promoter-specific fashion [31].

EGR1 and SP1 were also identified as the potential transcriptional regulators of *TGFB1* in human atherosclerotic arteries through computational methods. Interestingly in the same study, the authors also identified ZNF263, which was also a putative transcription factor (TF) with significant enriched motif in our study, as a regulator in the fine-tuning of *TGFB1* expression in atherosclerosis [23]. Further experimental validation by the authors in cultured human vascular smooth muscle cells showed that inhibition of the activity of SP1 and EGR1 induced a comparable decrease in the expression of some tightly co-expressed genes of the *TGFB1* cluster.

Besides these links from the TFs, putatively affected by the espistasis pair, to TGF-β signaling, rs679958, one of the SNPs of the epistasis pair, is by itself located at the intronic region of the gene coding for COL4A1, which directly interacts with bone morphogenic protein (BMP) to modulate TGF-β regulation [53].

Taken together, for the epistasis SNP-pair that we have identified, we hypothesize that the two variants may convey its genetic effect firstly through the perturbation of the interplay between nuclear matrix proteins and nuclear lamina, and, secondly, through the downstream perturbation of the TGF-β signaling pathway (Figure 4.1). However, further efforts need to be made in order to understand the exact biological mechanisms.

A recent study by Turner et al. suggested a statistical epistatic interaction effect on CAD risk between rs72655775 at the *COL4A1/COL4A2* locus and rs12441344 at the *SMAD3* locus based on a meta-analysis of 5 cohorts with 4,956 cases and 2,774 controls [97]. We made efforts replicate this pair in our 9 cohorts testing all the variants within the $\pm$500kb flanking region of the reported SNP with all 16 models. However, the epistatic effect between *COL4A2* and *SMAD3* could not be replicated with our samples. It is possible that this discrepancy is due to genomic heterogeneity between the study samples. It is also possible that the sample size in the analyses of Turner et al (N<10,100) was not powerful enough to get conclusive results.

On the other hand, as already indicated translating the statistical epistasis in biological epistasis can be challenging. The complexity of biological systems usually cause the need of organisms to interact with their environment to adjust, acclimatize, or maintain homeostasis, which may result in negligible levels of statistical epistasis even when the biological epistasis is pervasive [88]. Nevertheless, systematically characterizing statistical epistasis and proposing and validating hypotheses, would largely improve our chance of identify unbiased and unknown epistasis.

Chapter 5

Conclusion and Outlook

5.1 Summary

The aim of my dissertation was to improve the understanding of CAD through statistical genetics approaches. My work includes three research parts addressing the complex genetic architecture from different aspects: 1. Genome-wide association studies for CAD in the 1000G era; 2. Understanding the genetic complexity of regional GWAS signals of CAD; 3. Detecting epistasis underlying CAD.

5.1.1 Genome-wide association analysis for CAD in the 1000G era

I investigated the univariate effects of genetic variants in a genome-wide scale. The GWAS approach has the advantage of hypothesis-free, such that all variants on the genome are investigated and the results are thus not biased towards the prior knowledge. I performed analyses based on traditional genotyping array data on four cohort studies with the aim to add further knowledge to the genetic loci affecting CAD risk. Subsequently, I analyzed traditional genotyping array data on one cohort study to identify potential genetic susceptible loci on the X-chromosome.

The analyses conducted by me were limited to render the full spectrum of genetic causes of CAD, mainly due to the fact that the study samples were of limited size and thus did not provide enough power for detecting significant new loci on a genome-wide scale. Nevertheless, all of my findings served as valuable components for the CARDIoGRAMplusC4D consortium, where a meta-analysis was conducted.

So far one such meta-analysis was published [77] based on a whole collection of ∼185,000 CAD cases and controls, and 10 novel CAD-associated loci could be identified. Findings from the 1000G GWAS meta-analysis results support the common disease–common-variant hypothesis for CAD.

The meta-analysis of the 1000G CAD X-Chr has been completed based on more than 43,000 CAD cases and 58,000 controls, and summarized in a manuscript by Loley et al [59]. Findings from the 1000G X-Chromosome meta-analysis revealed that there was no genome-wide significant X-chromosomal variants associated with the risk of CAD.

5.1.2 Understanding the genetic complexity of GWAS Signals of CAD

Intra-locus allelic heterogeneity

With the aim to study additive effects of risk alleles as well as their impact on risk prediction, I have assessed the effect of multiple independent signals at 25 known CAD loci on 8 cohorts of individual-level genotype datasets for the suggestive SNPs derived from 1000G GWAS meta-analysis.

Typically for GWAS results only the best SNP at each locus is reported. However, a single SNP may not capture the overall amount of variation at a locus because there may be multiple causal variants. By conditional analysis the 1000G CAD GWAS meta-analysis [77] allowed the identification of multiple independently associated alleles at known loci, which increases the total contribution of the respective loci to genetic

variance of CAD.

My investigation went specifically into 8 cohorts with individual-level genotype data and examined the additive effects of multiple associated alleles at locus basis. Indeed, for some of the loci the combined effect of loci harboring multiple independent signals could be recovered, and the incremental effect on the association with CAD is also noticed with the increase of the number of independent alleles. The results improve our understanding of the allelic structure of these individual CAD-associated loci. It also highlights the importance and complexity of the issue of genetic heterogeneity. Moreover, these findings demonstrate that the functional contribution of the mechanisms affected by a locus (or a given gene) may go beyond the odds ratios reported for the lead SNP.

Multi-locus polygenic pleiotropy

The genetic architecture for complex diseases is usually complex as well. It happens often that both polygenic (one phenotype can also be caused by multiple genes/loci) and genetic pleiotropy (one gene/locus causes multiple phenotypes) exist. With the popularity of large-scale GWAS analyses, genetic variants identified to be associated with different complex traits could be utilized to explore the shared genetic background for traits of interest. When individual-level genotype data is available, a multi-locus polygenic score (PGS) technique is often further applied to assess the potential risk prediction value based on the shared genetic loci.

My investigations on the relationship between genetic susceptible loci of height and rheumatoid arthritis (RA) and the risk of CAD based on multiple individual-level genotype datasets have helped to support the notion that height affects directly and indirectly CAD risk, as well as the notion that genetic factors underlying RA carry a low likelihood to affect CAD risk.

5.1.3 Detecting Epistasis that Underlie CAD

The genetic architecture of complex traits are considered to be much more complex than an independent additive model. Epistasis has been suggested to be important for complex disease [64, 81, 117], however, no large-scale systematic investigation has been made in the context of CAD.

To our current knowledge, our investigation at meta-analysis level with 27,360 individuals and with the searching space of 8,068 SNPs is by far the largest scale for CAD epistasis. Besides, assuming the 4×4 epistasis model allowed us to detect the potential epistasis effect unbiased towards any genotype models.

One statistical trans-epistatic SNP-pair (rs71524277-rs679958) has been first identified at a p-value of 3.06×10^{-11}. Furthermore, the same epispastic effect for the lead pair has also been identified at the gene expression level. Finally, the biological interpretation underlying this trans-epistasis pair has been postulated, which proposes a novel hypothesis on how genetic loci could interact between each other to affect the risk of CAD. We hypothesize that the epistasis pair may convey its genetic effect first through the perturbation of the interplay between nuclear matrix proteins and nuclear lamina and then through the downstream perturbation on the TGF-β signaling pathway, which is a great extension to our current knowledge of the CAD genetics.

The success with our approach also supports the potential to further explore epistasis affecting CAD risk at a genome-wide scale.

5.2 Outlook

Despite the all the efforts and success from all kinds of statistical genetic studies such as GWAS analyses, polygenic score calculation, epistasis investigation, as well as the genetic etiology of complex diseases, CAD is still far from being completely understood.

5.2.1 Missing heritability

By referring to missing heritability for a trait in a population, researchers now focus on the missing gap between the narrow-sense heritability, which is the additive genetic portion of the phenotypic variance explained by a set of known genetic variants [117] and the total heritability, which is estimated via comparison the phenotypic concordance of monozygotic (MZ, identical) twins versus dizygotic (DZ, fraternal) twins.

The unexplained heritability has long been an unsolved question in statistical genetic research in complex diseases. Different reasons for the missing heritability have been proposed. These include large numbers of variants of smaller effect size that yet need to be found, structural variants poorly captured by the existing array technology, low power to detect gene-gene interactions, and inadequate accounting for shared environment among relatives [27, 64].

According the current result of GWAS meta-analysis, all the current known CAD loci can only explain a limited proportion of total heritability of CAD. Nevertheless, GWAS may probably still be an efficient way of investigating missing heritability, with the development of improved genotyping arrays, methods for imputation, next generation sequencing, and advanced statistical methods, the numbers as well as the regions (from single nucleotide polymorphism to long structural variants) of loci found to be associated with diseases will hopefully continue to expand. By the end of my thesis work, the 1000 Genomes Project has released the latest version of the Phase 3 reference genome. The final data set contains data from 2,504 individuals from 26 populations. Low coverage exome sequence data are present for all of these individuals, 24 individuals were also sequenced at high coverage for validation purposes [35, 94]. This new reference panel will be a valuable source for future exploration.

The multiple independent variants within known CAD loci could be another source of add-on explanation for the heritability. As suggested by the conditional and joint analysis from the 1000G GWAS meta-analysis, the ninety-five variants mapped to 44

known CAD loci from GWAS could explain $13.3 \pm 0.4\%$ of CAD heritability. Besides the above evidence, as a part of pre-analysis for epistasis project (see Results 3.3.1), additive variance was observed when we estimated the narrow-sense heritability around the 56 known CAD ris loci. All these data have suggested that for loci that have already been identified, a further mining into depth may still be a possible way to uncover part of the missing heritability.

Epistasis, although it is a capture of non-additive effects by definition, has also been shown to contribute to the narrow-sense heritability. This may be due to statistical illusion of additive variance, but it could be due to real additive variation as marginal effects from higher order genetic interaction [40]. By future exploration into the CAD epistasis, the extent to which epistasis could contribute to the missing heritability may be further elucidated.

5.2.2 Understanding the complex genetic architecture

As multiple independent variants exist within known CAD loci, it may be important to systematically identify instances of allelic heterogeneity and to examine the extent to which additional SNPs can help to shed light on the functional basis of genetic variations in CAD. With the rapid growth of next-generation-sequencing data, the allelic structure of disease-associated loci could be more refined.

Dissecting the allelic heterogeneity on a locus-by-locus basis to closely examine the patterns/existence of dependencies and additive or interactive effects may expand our knowledge of the general genetic mechanisms of complex diseases. For example, Kuo et al. have used a series of Col4a1 and Col4a2 mutant mouse lines to investigate the allelic heterogeneity of COL4A1/A2 in ocular dysgenesis, myopathy and brain malformations [54]. They observed that different Col4a1 and Col4a2 mutations had distinct effects on COL4A1 and COL4A2 biosynthesis and distinct molecular consequences that lead to ocular, cerebral and myopathic phenotypes of variable severity and penetrance, which

reflected the mechanistic heterogeneity.

In atherosclerotic plaques substantial clinical heterogeneity can be observed in morphological differences [98] which could be due to differences in the genetic susceptibilities. There is no doubt that in the context of CAD further investigation into the functional mechanisms underlying the allelic heterogeneity at each locus would not only enhance our understanding of the CAD genetic and molecular etiology, but also offer more knowledge for the potential possibility in the development of therapeutic interventions.

Nowadays it has been widely accepted that exploring epistasis is likely to be crucial to understand complex diseases [63]. By design, the GWAS approach ignores the interaction information, which involves multiple genetic variants and interactions. The development and improvement on the methods to handle the statistical and computational challenges will largely promote the identification of more statistical epistasis, and thus provide us with new opportunities to understand how naturally occurring genetic variants jointly act to modulate disease risk. Further efforts in narrowing the gap between statistical epistasis and biological epistasis and in functional validation and elucidation of the exact epistasis mechanism, will provide valuable insights into the complexities of the genetic architecture and molecular mechanisms of CAD.

Besides epistasis or gene-gene interaction, gene-environment interactions may also play a significant role in determining complex diseases [45]. The onset of CAD has been know to be affected by a large number of environmental factors, such as smoking, obesity, diabetes and sedentary life style. However, little is known how various types of environment factors interact with the genetic variations and co-affect the disease susceptibility.

5.2.3 Clinical implications

Recent advances in genome-wide association studies have stimulated interest in personalized medicine, where genetic information together with clinical information could be used to predict the individuals' disease risk. Thereby, preventive measures, focused diagnostic procedures or early interventions may be facilitated [10].

Current genetic risk prediction approaches are mostly based on the information combining the lead-SNPs reported by GWAS analysis. For example, the NHGRI GWAS Catalog [105] is a popular repository for researcher to collect disease-associated SNPs-However, usually only the top SNP reported in the original literature is listed. Previous efforts in the multi-locus polygenic risk score prediction of CAD and cardiovascular disease have mostly indicated that multi-locus PGSs are not particularly successful at predicting the incidence of CAD events [21, 79].

While the detection of multiple independent variants at single CAD risk locus is encouraging, it also highlights the limitations of current genetic risk score approaches. Although prediction based on variants at single loci still requires effort to better prioritize for both, the variants and the prediction model, the information from allelic heterogeneity or additive SNPs at known loci may have the potential to serve a better basic feature set than only the lead-SNPs and may finally lead to a better prediction with proper genetic risk score modeling. Further elucidation of the possible distinct biological consequences of multiple variants at single loci would be valuable as genetic counseling information to improve the accuracy of prognoses.

The accuracy of genetic predictive models may also be facilitated with the identification of disease-associated statistical epistasis. Despite the difficulty in translating the statistical epistasis into biological mechanisms, the effect estimates derived from population-level studies by themselves would be informative for risk prediction. Recently, methods about the integration of epistasis information into the genetic risk prediction models have been explored [1, 65]. This provides us with the plausible outlook, that

epistasis could bring potential benefit in personalized medicine for CAD.

Bibliography

[1] Deniz Akdemir and Jean-Luc Jannink. Locally Epistatic Models for Genome-wide Prediction and Association by Importance Sampling. *arXiv:1603.08813*, 2016.

[2] A Al-Sarraj, R M Day, and G Thiel. Specificity of transcriptional regulation by the zinc finger transcription factors Sp1, Sp3, and Egr-1. *J Cell Biochem*, 94(1):153–167, 2005.

[3] C A Anderson, F H Pettersson, G M Clarke, L R Cardon, A P Morris, and K T Zondervan. Data quality control in genetic case-control association studies. *Nat Protoc*, 5(9):1564–1573, 2010.

[4] S W Bahouth, M J Beauchamp, and K N Vu. Reciprocal regulation of beta(1)-adrenergic receptor gene transcription by Sp1 and early growth response gene 1: induction of EGR-1 inhibits the expression of the beta(1)-adrenergic receptor gene. *Mol Pharmacol*, 61(2):379–390, 2002.

[5] T L Bailey, M Boden, F A Buske, M Frith, C E Grant, L Clementi, J Ren, W W Li, and W S Noble. MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res*, 37(Web Server issue):W202–8, 2009.

[6] A Bobik. Transforming Growth Factor-s and Vascular Disorders. *Arteriosclerosis, Thrombosis, and Vascular Biology*, 26(8):1712–1720, 2006.

[7] M Bujak and N G Frangogiannis. The role of TGF-beta signaling in myocardial infarction and cardiac remodeling. *Cardiovasc Res*, 74(2):184–195, 2007.

[8] S Burgess and S G Thompson. Use of allele scores as instrumental variables for Mendelian randomization. *Int J Epidemiol*, 42(4):1134–1144, 2013.

[9] B C Capell and F S Collins. Human laminopathies: nuclei gone genetically awry. *Nat Rev Genet*, 7(12):940–952, 2006.

[10] N Chatterjee, J Shi, and M Garcia-Closas. Developing and evaluating polygenic risk prediction models for stratified disease prevention. *Nat Rev Genet*, 2016.

[11] D Clayton. Testing for association on the X chromosome. *Biostatistics*, 9(4):593–600, 2008.

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

[13] F S Collins, L D Brooks, and A Chakravarti. A DNA Polymorphism Discovery Resource for Research on Human Genetic Variation. *Genome Res*, 8:122901231, 1998.

[14] T Condamine, L Le Texier, D Howie, A Lavault, M Hill, F Halary, S Cobbold, H Waldmann, M C Cuturi, and E Chiffoleau. Tmem176B and Tmem176A are associated with the immature state of dendritic cells. *J Leukoc Biol*, 88(3):507–515, 2010.

[15] CARDIoGRAMplusC4D Consortium. Cardiogramplusc4d consortium. `http://www.cardiogramplusc4d.org`, 2016.

[16] CARDIoGRAMplusC4D Consortium, P Deloukas, S Kanoni, C Willenborg, M Farrall, T L Assimes, J R Thompson, E Ingelsson, D Saleheen, J Erdmann, B A Goldstein, K Stirrups, I R Konig, J B Cazier, A Johansson, A S Hall, J Y Lee, C J Willer, J C Chambers, T Esko, L Folkersen, A Goel, E Grundberg, A S Havulinna, W K Ho, J C Hopewell, N Eriksson, M E Kleber, K Kristiansson, P Lundmark, L P Lyytikainen, S Rafelt, D Shungin, R J Strawbridge, G Thorleifsson, E Tikkanen, N Van Zuydam, B F Voight, L L Waite, W Zhang, A Ziegler, D Absher, D Altshuler, A J Balmforth, I Barroso, P S Braund, C Burgdorf, S Claudi-Boehm, D Cox, M Dimitriou, R Do, Diagram Consortium, Cardiogenics Consortium, A S Doney, N El Mokhtari, P Eriksson, K Fischer, P Fontanillas, A Franco-Cereceda, B Gigante, L Groop, S Gustafsson, J Hager, G Hallmans, B G Han, S E Hunt, H M Kang, T Illig, T Kessler, J W Knowles, G Kolovou, J Kuusisto, C Langenberg, C Langford, K Leander, M L Lokki, A Lundmark, M I McCarthy, C Meisinger, O Melander, E Mihailov, S Maouche, A D Morris, M Muller-Nurasyid, Ther Consortium Mu, K Nikus, J F Peden, N W Rayner, A Rasheed, S Rosinger, D Rubin, M P Rumpf, A Schafer, M Sivananthan, C Song, A F Stewart, S T Tan, G Thorgeirsson, C E van der Schoot, P J Wagner, Consortium Wellcome Trust Case Control, G A Wells, P S Wild, T P Yang, P Amouyel, D Arveiler, H Basart, M Boehnke, E Boerwinkle, P Brambilla, F Cambien, A L Cupples, U de Faire, A Dehghan, P Diemert, S E Epstein, A Evans, M M Ferrario, J Ferrieres, D Gauguier, A S Go, A H Goodall, V Gudnason, S L Hazen, H Holm, C Iribarren, Y Jang, M Kahonen, F Kee, H S Kim, N Klopp, W Koenig, W Kratzer, K Kuulasmaa, M Laakso, R Laaksonen, J Y Lee, L Lind, W H Ouwehand, S Parish, J E Park, N L Pedersen, A Peters, T Quertermous, D J Rader, V Salomaa, E Schadt, S H Shah, J Sinisalo, K Stark, K Stefansson, D A Tregouet, J Virtamo, L Wallentin, N Wareham, M E Zimmermann, M S Nieminen, C Hengstenberg, M S Sandhu, T Pastinen, A C Syvanen, G K Hovingh, G Dedoussis, P W Franks, T Lehtimaki, A Metspalu, P A

Zalloua, A Siegbahn, S Schreiber, S Ripatti, S S Blankenberg, M Perola, R Clarke, B O Boehm, C O'Donnell, M P Reilly, W Marz, R Collins, S Kathiresan, A Hamsten, J S Kooner, U Thorsteinsdottir, J Danesh, C N Palmer, R Roberts, H Watkins, H Schunkert, and N J Samani. Large-scale association analysis identifies new risk loci for coronary artery disease. *Nat Genet*, 45(1):25–33, 2013.

[17] H J Cordell. epistasis, what it means, what it doesn't mean, and statistical methods to detect it in humans. *Hum Mol Genet*, 2002.

[18] H J Cordell. Detecting gene-gene interactions that underlie human diseases. *Nat Rev Genet*, 10(6):392–404, 2009.

[19] G Cordova, A Rochard, C Riquelme-Guzman, C Cofre, D Scherman, P Bigey, and E Brandan. SMAD3 and SP1/SP3 Transcription Factors Collaborate to Regulate Connective Tissue Growth Factor Gene Expression in Myoblasts in Response to Transforming Growth Factor beta. *J Cell Biochem*, 116(9):1880–1887, 2015.

[20] dbGaP. The database of genotypes and phenotypes. `http://www.ncbi.nlm.nih.gov/gap`, 2015.

[21] P S de Vries, M Kavousi, S Ligthart, A G Uitterlinden, A Hofman, O H Franco, and A Dehghan. Incremental predictive value of 152 single nucleotide polymorphisms in the 10-year risk prediction of incident coronary heart disease: the Rotterdam Study. *Int J Epidemiol*, 44(2):682–688, 2015.

[22] O Delaneau, J F Zagury, and J Marchini. Improved whole-chromosome phasing for disease and population genetic studies. *Nat Methods*, 10(1):5–6, 2013.

[23] N Dhaouadi, J Y Li, P Feugier, M P Gustin, H Dab, K Kacem, G Bricca, and C Cerutti. Computational identification of potential transcriptional regulators of TGF-ss1 in human atherosclerotic arteries. *Genomics*, 103(5-6):357–370, 2014.

[24] e:AtheroSysMed. Systemmedizin der koronaren herzkrankheit und des schlaganfalls. http://www.sys-med.de/de/konsortien/eatherosysmed/, 2013.

[25] M T Ebbert, P G Ridge, and J S Kauwe. Bridging the gap between statistical and biological epistasis in Alzheimer's disease. *Biomed Res Int*, 2015:870123, 2015.

[26] S L Edwards, J Beesley, J D French, and A M Dunning. Beyond GWASs: illuminating the dark road from association to function. *Am J Hum Genet*, 93(5):779–797, 2013.

[27] E E Eichler, J Flint, G Gibson, A Kong, S M Leal, J H Moore, and J H Nadeau. Missing heritability and strategies for finding the underlying causes of complex disease. *Nat Rev Genet*, 11(6):446–450, 2010.

[28] J Erdmann, A Grosshennig, P S Braund, I R Konig, C Hengstenberg, A S Hall, P Linsel-Nitschke, S Kathiresan, B Wright, D A Tregouet, F Cambien, P Bruse, Z Aherrahrou, A K Wagner, K Stark, S M Schwartz, V Salomaa, R Elosua, O Melander, B F Voight, C J O'Donnell, L Peltonen, D S Siscovick, D Altshuler, P A Merlini, F Peyvandi, L Bernardinelli, D Ardissino, A Schillert, S Blankenberg, T Zeller, P Wild, D F Schwarz, L Tiret, C Perret, S Schreiber, N E El Mokhtari, A Schafer, W Marz, W Renner, P Bugert, H Kluter, J Schrezenmeir, D Rubin, S G Ball, A J Balmforth, H E Wichmann, T Meitinger, M Fischer, C Meisinger, J Baumert, A Peters, W H Ouwehand, Thrombosis Italian Atherosclerosis, Group Vascular Biology Working, Consortium Myocardial Infarction Genetics, Consortium Wellcome Trust Case Control, Consortium Cardiogenics, P Deloukas, J R Thompson, A Ziegler, N J Samani, and H Schunkert. New susceptibility locus for coronary artery disease on chromosome 3q22.3. *Nat Genet*, 41(3):280–282, 2009.

[29] S Eyre, J Bowes, D Diogo, A Lee, A Barton, P Martin, A Zhernakova, E Stahl, S Viatte, K McAllister, C I Amos, L Padyukov, R E Toes, T W Huizinga, C Wi-

jmenga, G Trynka, L Franke, H J Westra, L Alfredsson, X Hu, C Sandor, P I de Bakker, S Davila, C C Khor, K K Heng, R Andrews, S Edkins, S E Hunt, C Langford, D Symmons, Genetics Biologics in Rheumatoid Arthritis, Syndicate Genomics Study, Consortium Wellcome Trust Case Control, P Concannon, S Onengut-Gumuscu, S S Rich, P Deloukas, M A Gonzalez-Gay, L Rodriguez-Rodriguez, L Arlsetig, J Martin, S Rantapaa-Dahlqvist, R M Plenge, S Raychaudhuri, L Klareskog, P K Gregersen, and J Worthington. High-density genetic mapping identifies new susceptibility loci for rheumatoid arthritis. *Nat Genet*, 44(12):1336–1340, 2012.

[30] Xin-Hua Feng, X Lin, and Rik Derynck. Smad2, Smad3 and Smad4 cooperate with Sp1 to induce p15Ink4B transcription in response to TGF-beta. *The EMBO J*, 19(19):5178–5193, 2000.

[31] J Fortin and D J Bernard. SMAD3 and EGR1 physically and functionally interact in promoter-specific fashion. *Cell Signal*, 22(6):936–943, 2010.

[32] A Franke, D P McGovern, J C Barrett, K Wang, G L Radford-Smith, T Ahmad, C W Lees, T Balschun, J Lee, R Roberts, C A Anderson, J C Bis, S Bumpstead, D Ellinghaus, E M Festen, M Georges, T Green, T Haritunians, L Jostins, A Latiano, C G Mathew, G W Montgomery, N J Prescott, S Raychaudhuri, J I Rotter, P Schumm, Y Sharma, L A Simms, K D Taylor, D Whiteman, C Wijmenga, R N Baldassano, M Barclay, T M Bayless, S Brand, C Buning, A Cohen, J F Colombel, M Cottone, L Stronati, T Denson, M De Vos, R D'Inca, M Dubinsky, C Edwards, T Florin, D Franchimont, R Gearry, J Glas, A Van Gossum, S L Guthery, J Halfvarson, H W Verspaget, J P Hugot, A Karban, D Laukens, I Lawrance, M Lemann, A Levine, C Libioulle, E Louis, C Mowat, W Newman, J Panes, A Phillips, D D Proctor, M Regueiro, R Russell, P Rutgeerts, J Sanderson, M Sans, F Seibold, A H Steinhart, P C Stokkers, L Torkvist, G Kullak-Ublick, D Wilson, T Walters,

S R Targan, S R Brant, J D Rioux, M D'Amato, R K Weersma, S Kugathasan, A M Griffiths, J C Mansfield, S Vermeire, R H Duerr, M S Silverberg, J Satsangi, S Schreiber, J H Cho, V Annese, H Hakonarson, M J Daly, and M Parkes. Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nat Genet*, 42(12):1118–1125, 2010.

[33] Consortium Genomes Project, G R Abecasis, D Altshuler, A Auton, L D Brooks, R M Durbin, R A Gibbs, M E Hurles, and G A McVean. A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061–1073, 2010.

[34] Consortium Genomes Project, G R Abecasis, A Auton, L D Brooks, M A DePristo, R M Durbin, R E Handsaker, H M Kang, G T Marth, and G A McVean. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):56–65, 2012.

[35] Consortium Genomes Project, A Auton, L D Brooks, R M Durbin, E P Garrison, H M Kang, J O Korbel, J L Marchini, S McCarthy, G A McVean, and G R Abecasis. A global reference for human genetic variation. *Nature*, 526(7571):68–74, 2015.

[36] D J Grainger. TGF-beta and atherosclerosis in man. *Cardiovasc Res*, 74(2):213–222, 2007.

[37] Y Gruenbaum, R D Goldman, R Meyuhas, E Mills, A Margalit, A Fridkin, Y Dayani, M Prokocimer, and A Enosh. The nuclear lamina and its functions in the nucleus. *Int Rev Cytol*, 226:1–62, 2003.

[38] S He, J M Sun, L Li, and J R Davie. Differential Intranuclear Organization of Transcription Factors Sp1 and Sp3. *Mol Biol Cell*, 16:4073–4083, 2005.

[39] A Helgadottir, G Thorleifsson, A Manolescu, A Kong, and K Stefansson. A Common Variant on Chromosome 9p21 Affects the Risk of Myocardial Infarction. *Science*, 316, 2007.

[40] G Hemani, S Knott, and C Haley. An Evolutionary Perspective on Epistasis and the Missing Heritability. *PLoS Genet*, 2013.

[41] A S Hinrichs, D Karolchik, R Baertsch, G P Barber, G Bejerano, H Clawson, M Diekhans, T S Furey, R A Harte, F Hsu, J Hillman-Jackson, R M Kuhn, J S Pedersen, A Pohl, B J Raney, K R Rosenbloom, A Siepel, K E Smith, C W Sugnet, A Sultan-Qurraie, D J Thomas, H Trumbower, R J Weber, M Weirauch, A S Zweig, D Haussler, and W J Kent. The UCSC Genome Browser Database: update 2006. *Nucleic Acids Res*, 34(Database issue):D590–8, 2006.

[42] B N Howie, P Donnelly, and J Marchini. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet*, 5(6):e1000529, 2009.

[43] R P Huang, Y Fan, Z Ni, D Mercola, and E D Adamson. Reciprocal modulation between Sp1 and Egr-1. *J Cell Biochem*, 66(4):489–499, 1997.

[44] W Huang da, B T Sherman, and R A Lempicki. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc*, 4(1):44–57, 2009.

[45] D J Hunter. Gene-environment interactions in human diseases. *Nat Rev Genet*, 6(4):287–298, 2005.

[46] H Jansen, C Willenborg, W Lieb, **L Zeng**, P G Ferrario, and ... Rheumatoid arthritis and coronary artery disease: Genetic analyses do not support a causal relation. *J Rheumatol*, 2016.

[47] K Jungert, A Buck, M Buchholz, M Wagner, G Adler, T M Gress, and V Ellenrieder. Smad-Sp1 complexes mediate TGFbeta-induced early transcription of oncogenic Smad7 in pancreatic cancer cells. *Carcinogenesis*, 27(12):2392–2401, 2006.

[48] T Kam-Thong, C A Azencott, L Cayton, B Putz, A Altmann, N Karbalai, P G Samann, B Scholkopf, B Muller-Myhsok, and K M Borgwardt. GLIDE: GPU-based linear regression for detection of epistasis. *Hum Hered*, 73(4):220–236, 2012.

[49] D Karolchik, A S Hinrichs, T S Furey, K M Roskin, C W Sugnet, D Haussler, and W J Kent. The UCSC Table Browser data retrieval tool. *Nucleic Acids Res*, 32(Database issue):D493–6, 2004.

[50] S Kathiresan, B F Voight, S Purcell, K Musunuru, D Ardissino, P M Mannucci, S Anand, J C Engert, N J Samani, H Schunkert, J Erdmann, M P Reilly, D J Rader, T Morgan, J A Spertus, M Stoll, D Girelli, P P McKeown, C C Patterson, D S Siscovick, C J O'Donnell, R Elosua, L Peltonen, V Salomaa, S M Schwartz, O Melander, D Altshuler, D Ardissino, P A Merlini, C Berzuini, L Bernardinelli, F Peyvandi, M Tubaro, P Celli, M Ferrario, R Fetiveau, N Marziliano, G Casari, M Galli, F Ribichini, M Rossi, F Bernardi, P Zonzin, A Piazza, P M Mannucci, S M Schwartz, D S Siscovick, J Yee, Y Friedlander, R Elosua, J Marrugat, G Lucas, I Subirana, J Sala, R Ramos, S Kathiresan, J B Meigs, G Williams, D M Nathan, C A MacRae, C J O'Donnell, V Salomaa, A S Havulinna, L Peltonen, O Melander, G Berglund, B F Voight, S Kathiresan, J N Hirschhorn, R Asselta, S Duga, M Spreafico, K Musunuru, M J Daly, S Purcell, B F Voight, S Purcell, J Nemesh, J M Korn, S A McCarroll, S M Schwartz, J Yee, S Kathiresan, G Lucas, I Subirana, R Elosua, A Surti, C Guiducci, L Gianniny, D Mirel, M Parkin, N Burtt, S B Gabriel, N J Samani, J R Thompson, P S Braund, B J Wright, A J Balmforth, S G Ball, A Hall, H Schunkert, J Erdmann, P Linsel-Nitschke, W Lieb, A Ziegler, I Konig, C Hengstenberg, M Fischer, K Stark, A Grosshennig, M Preuss, H E Wichmann, S Schreiber, H Schunkert, N J Samani, J Erdmann, W Ouwehand, C Hengstenberg, P Deloukas, M Scholz, F Cambien, M P Reilly,

M Li, Z Chen, R Wilensky, W Matthai, A Qasim, H H Hakonarson, J Devaney, M S Burnett, A D Pichard, K M Kent, L Satler, J M Lindsay, R Waksman, C W Knouff, D M Waterworth, M C Walker, V Mooser, S E Epstein, D J Rader, T Scheffold, K Berger, M Stoll, A Huge, D Girelli, N Martinelli, O Olivieri, R Corrocher, T Morgan, J A Spertus, P McKeown, C C Patterson, H Schunkert, E Erdmann, P Linsel-Nitschke, W Lieb, A Ziegler, I R Konig, C Hengstenberg, M Fischer, K Stark, A Grosshennig, M Preuss, H E Wichmann, S Schreiber, H Holm, G Thorleifsson, U Thorsteinsdottir, K Stefansson, J C Engert, R Do, C Xie, S Anand, S Kathiresan, D Ardissino, P M Mannucci, D Siscovick, C J O'Donnell, N J Samani, O Melander, R Elosua, L Peltonen, V Salomaa, S M Schwartz, and D Altshuler. Genome-wide association of early-onset myocardial infarction with single nucleotide polymorphisms and copy number variants. *Nat Genet*, 41(3):334–341, 2009.

[51] S Kathiresan, C J Willer, G M Peloso, S Demissie, K Musunuru, E E Schadt, L Kaplan, D Bennett, Y Li, T Tanaka, B F Voight, L L Bonnycastle, A U Jackson, G Crawford, A Surti, C Guiducci, N P Burtt, S Parish, R Clarke, D Zelenika, K A Kubalanza, M A Morken, L J Scott, H M Stringham, P Galan, A J Swift, J Kuusisto, R N Bergman, J Sundvall, M Laakso, L Ferrucci, P Scheet, S Sanna, M Uda, Q Yang, K L Lunetta, J Dupuis, P I de Bakker, C J O'Donnell, J C Chambers, J S Kooner, S Hercberg, P Meneton, E G Lakatta, A Scuteri, D Schlessinger, J Tuomilehto, F S Collins, L Groop, D Altshuler, R Collins, G M Lathrop, O Melander, V Salomaa, L Peltonen, M Orho-Melander, J M Ordovas, M Boehnke, G R Abecasis, K L Mohlke, and L A Cupples. Common variants at 30 loci contribute to polygenic dyslipidemia. *Nat Genet*, 41(1):56–65, 2009.

[52] J Kind, L Pagie, S S de Vries, L Nahidiazar, S S Dey, M Bienko, Y Zhan, B Lajoie, C A de Graaf, M Amendola, G Fudenberg, M Imakaev, L A Mirny, K Jalink,

J Dekker, A van Oudenaarden, and B van Steensel. Genome-wide maps of nuclear lamina interactions in single human cells. *Cell*, 163(1):134–147, 2015.

[53] D S Kuo, C Labelle-Dumais, and D B Gould. COL4A1 and COL4A2 mutations and disease: insights into pathogenic mechanisms and potential therapeutic targets. *Hum Mol Genet*, 21(R1):R97–110, 2012.

[54] D S Kuo, C Labelle-Dumais, M Mao, M Jeanne, W B Kauffman, J Allen, J Favor, and D B Gould. Allelic heterogeneity contributes to variability in ocular dysgenesis, myopathy and brain malformations caused by Col4a1 and Col4a2 mutations. *Hum Mol Genet*, 23(7):1709–1722, 2014.

[55] T LaFramboise. Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances. *Nucleic Acids Res*, 37(13):4181–4193, 2009.

[56] E S Lander. The New Genomics: Global Views of Biology. *Science*, 274, 1996.

[57] H Lango Allen, K Estrada, G Lettre, S I Berndt, M N Weedon, F Rivadeneira, C J Willer, A U Jackson, S Vedantam, S Raychaudhuri, T Ferreira, A R Wood, R J Weyant, A V Segre, E K Speliotes, E Wheeler, N Soranzo, J H Park, J Yang, D Gudbjartsson, N L Heard-Costa, J C Randall, L Qi, A Vernon Smith, R Magi, T Pastinen, L Liang, I M Heid, J Luan, G Thorleifsson, T W Winkler, M E Goddard, K Sin Lo, C Palmer, T Workalemahu, Y S Aulchenko, A Johansson, M C Zillikens, M F Feitosa, T Esko, T Johnson, S Ketkar, P Kraft, M Mangino, I Prokopenko, D Absher, E Albrecht, F Ernst, N L Glazer, C Hayward, J J Hottenga, K B Jacobs, J W Knowles, Z Kutalik, K L Monda, O Polasek, M Preuss, N W Rayner, N R Robertson, V Steinthorsdottir, J P Tyrer, B F Voight, F Wiklund, J Xu, J H Zhao, D R Nyholt, N Pellikka, M Perola, J R Perry, I Surakka, M L Tammesoo, E L Altmaier, N Amin, T Aspelund, T Bhangale, G Boucher, D I

Chasman, C Chen, L Coin, M N Cooper, A L Dixon, Q Gibson, E Grundberg, K Hao, M Juhani Junttila, L M Kaplan, J Kettunen, I R Konig, T Kwan, R W Lawrence, D F Levinson, M Lorentzon, B McKnight, A P Morris, M Muller, J Suh Ngwa, S Purcell, S Rafelt, R M Salem, E Salvi, S Sanna, J Shi, U Sovio, J R Thompson, M C Turchin, L Vandenput, D J Verlaan, V Vitart, C C White, A Ziegler, P Almgren, A J Balmforth, H Campbell, L Citterio, A De Grandi, A Dominiczak, J Duan, P Elliott, R Elosua, J G Eriksson, N B Freimer, E J Geus, N Glorioso, S Haiqing, A L Hartikainen, A S Havulinna, A A Hicks, J Hui, W Igl, T Illig, A Jula, E Kajantie, T O Kilpelainen, M Koiranen, I Kolcic, S Koskinen, P Kovacs, J Laitinen, J Liu, M L Lokki, A Marusic, A Maschio, T Meitinger, A Mulas, G Pare, A N Parker, J F Peden, A Petersmann, I Pichler, K H Pietilainen, A Pouta, M Ridderstrale, J I Rotter, J G Sambrook, A R Sanders, C O Schmidt, J Sinisalo, J H Smit, H M Stringham, G Bragi Walters, E Widen, S H Wild, G Willemsen, L Zagato, L Zgaga, P Zitting, H Alavere, M Farrall, W L McArdle, M Nelis, M J Peters, S Ripatti, J B van Meurs, K K Aben, K G Ardlie, J S Beckmann, J P Beilby, R N Bergman, S Bergmann, F S Collins, D Cusi, M den Heijer, G Eiriksdottir, P V Gejman, A S Hall, A Hamsten, H V Huikuri, C Iribarren, M Kahonen, J Kaprio, S Kathiresan, L Kiemeney, T Kocher, L J Launer, T Lehtimaki, O Melander, T H Mosley Jr., A W Musk, M S Nieminen, C J O'Donnell, C Ohlsson, B Oostra, L J Palmer, O Raitakari, P M Ridker, J D Rioux, A Rissanen, C Rivolta, H Schunkert, A R Shuldiner, D S Siscovick, M Stumvoll, A Tonjes, J Tuomilehto, G J van Ommen, J Viikari, A C Heath, N G Martin, G W Montgomery, M A Province, M Kayser, A M Arnold, L D Atwood, E Boerwinkle, S J Chanock, P Deloukas, C Gieger, H Gronberg, P Hall, A T Hattersley, C Hengstenberg, W Hoffman, G M Lathrop, V Salomaa, S Schreiber, M Uda, D Waterworth, A F Wright, T L Assimes, I Barroso, A Hofman, K L Mohlke, D I Boomsma, M J Caulfield, L A Cupples, J Erdmann, C S Fox,

V Gudnason, U Gyllensten, T B Harris, R B Hayes, M R Jarvelin, V Mooser, P B Munroe, W H Ouwehand, B W Penninx, P P Pramstaller, T Quertermous, I Rudan, N J Samani, T D Spector, H Volzke, H Watkins, J F Wilson, L C Groop, T Haritunians, F B Hu, R C Kaplan, A Metspalu, K E North, D Schlessinger, N J Wareham, D J Hunter, J R O'Connell, D P Strachan, H E Wichmann, I B Borecki, C M van Duijn, E E Schadt, U Thorsteinsdottir, L Peltonen, A G Uitterlinden, P M Visscher, N Chatterjee, R J Loos, M Boehnke, M I McCarthy, E Ingelsson, C M Lindgren, G R Abecasis, K Stefansson, T M Frayling, and J N Hirschhorn. Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature*, 467(7317):832–838, 2010.

[58] M LeBlanc, V Zuber, B K Andreassen, A Witoelar, **L Zeng**, F Bettella, Y Wang, L K McEvoy, W K Thompson, A J Schork, S Reppe, E Barrett-Connor, S Ligthart, A Dehghan, K M Gautvik, C P Nelson, H Schunkert, N J Samani, CARDIoGRAM Consortium, P M Ridker, D I Chasman, P Aukrust, S Djurovic, A Frigessi, R S Desikan, A M Dale, and O A Andreassen. Identifying Novel Gene Variants in Coronary Artery Disease and Shared Genes With Several Cardiovascular Risk Factors. *Circ Res*, 118(1):83–94, 2016.

[59] C Loley, M Alver, T L Assimes, A Bjonnes, A Goel, ..., **L Zeng**, and ... No Association of Coronary Artery Disease with X-Chromosomal Variants in Comprehensive International Meta-Analysis. *Scientific Reports*, 2016.

[60] C Loley, A Ziegler, and I R Konig. Association tests for X-chromosomal markers–a comparison of different test statistics. *Hum Hered*, 71(1):23–36, 2011.

[61] G Lucas, C Lluis-Ganella, I Subirana, M D Musameh, J R Gonzalez, C P Nelson, M Senti, Consortium Myocardial Infarction Genetics, Consortium Wellcome Trust Case Control, S M Schwartz, D Siscovick, C J O'Donnell, O Melander, V Salomaa, S Purcell, D Altshuler, N J Samani, S Kathiresan, and R Elosua. Hypothesis-based

analysis of gene-gene interactions and risk of myocardial infarction. *PLoS One*, 7(8):e41730, 2012.

[62] L Ma, A Keinan, and A G Clark. Biological Knowledge-Driven Analysis of Epistasis in Human GWAS with Application to Lipid Traits. In *Epistasis. Methods and Protocols.* 2014.

[63] T F Mackay. Epistasis and quantitative traits: using model organisms to study gene-gene interactions. *Nat Rev Genet*, 15(1):22–33, 2014.

[64] T A Manolio, F S Collins, N J Cox, D B Goldstein, L A Hindorff, D J Hunter, M I McCarthy, E M Ramos, L R Cardon, A Chakravarti, J H Cho, A E Guttmacher, A Kong, L Kruglyak, E Mardis, C N Rotimi, M Slatkin, D Valle, A S Whittemore, M Boehnke, A G Clark, E E Eichler, G Gibson, J L Haines, T F Mackay, S A McCarroll, and P M Visscher. Finding the missing heritability of complex diseases. *Nature*, 461(7265):747–753, 2009.

[65] M W Marcus, O Y Raji, S W Duffy, R P Young, R J Hopkins, and J K Field. Incorporating epistasis interaction of genetic susceptibility single nucleotide polymorphisms in a lung cancer risk prediction model. *Int J Onc*, 49(1), 2016.

[66] A Mathelier, O Fornes, D J Arenillas, C Y Chen, G Denay, J Lee, W Shi, C Shyr, G Tan, R Worsley-Hunt, A W Zhang, F Parcy, B Lenhard, A Sandelin, and W W Wasserman. JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res*, 44(D1):D110–5, 2016.

[67] B Mayer, J Erdmann, and H Schunkert. Genetics and heritability of coronary artery disease and myocardial infarction. *Clin Res Cardiol*, 96(1):1–7, 2007.

[68] R McPherson, A Pertsemlidis, N Kavaslar, A Stewart, R Roberts, D R Cox, and J C Cohen. A Common Allele on Chromosome 9 Associated with Coronary Heart Disease. *Science*, 316, 2007.

[69] S Mendis, L H Lindholm, S G Anderson, A Alwan, R Koju, B J Onwubere, A M Kayani, N Abeysinghe, A Duneas, S Tabagari, W Fan, N Sarraf-Zadegan, P Nordet, J Whitworth, and A Heagerty. Total cardiovascular risk approach to improve efficiency of cardiovascular prevention in resource constrain settings. *J Clin Epidemiol*, 64(12):1451–1462, 2011.

[70] R L Milne, J Herranz, K Michailidou, J Dennis, J P Tyrer, M P Zamora, J I Arias-Perez, A Gonzalez-Neira, G Pita, M R Alonso, Q Wang, M K Bolla, K Czene, M Eriksson, K Humphreys, H Darabi, J Li, H Anton-Culver, S L Neuhausen, A Ziogas, C A Clarke, J L Hopper, G S Dite, C Apicella, M C Southey, G Chenevix-Trench, Investigators KConFab, Group Australian Ovarian Cancer Study, A Swerdlow, A Ashworth, N Orr, M Schoemaker, A Jakubowska, J Lubinski, K Jaworska-Bieniek, K Durda, I L Andrulis, J A Knight, G Glendon, A M Mulligan, S E Bojesen, B G Nordestgaard, H Flyger, H Nevanlinna, T A Muranen, K Aittomaki, C Blomqvist, J Chang-Claude, A Rudolph, P Seibold, D Flesch-Janys, X Wang, J E Olson, C Vachon, K Purrington, R Winqvist, K Pylkas, A Jukkola-Vuorinen, M Grip, A M Dunning, M Shah, P Guenel, T Truong, M Sanchez, C Mulot, H Brenner, A K Dieffenbach, V Arndt, C Stegmaier, A Lindblom, S Margolin, M J Hooning, A Hollestelle, J M Collee, A Jager, A Cox, I W Brock, M W Reed, P Devilee, R A Tollenaar, C Seynaeve, C A Haiman, B E Henderson, F Schumacher, L Le Marchand, J Simard, M Dumont, P Soucy, T Dork, N V Bogdanova, U Hamann, A Forsti, T Rudiger, H U Ulmer, P A Fasching, L Haberle, A B Ekici, M W Beckmann, O Fletcher, N Johnson, I dos Santos Silva, J Peto, P Radice, P Peterlongo, B Peissel, P Mariani, G G Giles, G Severi, L Baglietto, E Sawyer,

I Tomlinson, M Kerin, N Miller, F Marme, B Burwinkel, A Mannermaa, V Kataja, V M Kosma, J M Hartikainen, D Lambrechts, B T Yesilyurt, G Floris, K Leunen, G G Alnaes, V Kristensen, A L Borresen-Dale, M Garcia-Closas, S J Chanock, J Lissowska, J D Figueroa, M K Schmidt, A Broeks, S Verhoef, E J Rutgers, H Brauch, T Bruning, Y D Ko, Genica Network, F J Couch, A E Toland, Tnbcc, D Yannoukakos, P D Pharoah, P Hall, J Benitez, N Malats, and D F Easton. A large-scale assessment of two-way SNP interactions in breast cancer susceptibility using 46,450 cases and 42,461 controls from the breast cancer association consortium. *Hum Mol Genet*, 23(7):1934–1946, 2014.

[71] J H Moore and S M Williams. Traversing the conceptual divide between biological and statistical epistasis: systems biology and a more modern synthesis. *Bioessays*, 27(6):637–646, 2005.

[72] K S Moorefield, H Yin, T D Nichols, C Cathcart, S O Simmons, and J M Horowitz. Sp2 localizes to subnuclear foci associated with the nuclear matrix. *Mol Biol Cell*, 17(4):1711–1722, 2006.

[73] W Murk, M B Bracken, and A T DeWan. Confronting the missing epistasis problem: on the reproducibility of gene-gene interactions. *Hum Genet*, 134(8):837–849, 2015.

[74] M D Musameh, W Y Wang, C P Nelson, C Lluis-Ganella, R Debiec, I Subirana, R Elosua, A J Balmforth, S G Ball, A S Hall, S Kathiresan, J R Thompson, G Lucas, N J Samani, and M Tomaszewski. Analysis of gene-gene interactions among common variants in candidate cardiovascular genes in coronary artery disease. *PLoS One*, 10(2):e0117684, 2015.

[75] C P Nelson, S E Hamby, D Saleheen, J C Hopewell, **L Zeng**, T L Assimes, S Kanoni, C Willenborg, S Burgess, P Amouyel, S Anand, S Blankenberg, B O

Boehm, R J Clarke, R Collins, G Dedoussis, M Farrall, P W Franks, L Groop, A S Hall, A Hamsten, C Hengstenberg, G K Hovingh, E Ingelsson, S Kathiresan, F Kee, I R Konig, J Kooner, T Lehtimaki, W Marz, R McPherson, A Metspalu, M S Nieminen, C J O'Donnell, C N Palmer, A Peters, M Perola, M P Reilly, S Ripatti, R Roberts, V Salomaa, S H Shah, S Schreiber, A Siegbahn, U Thorsteinsdottir, G Veronesi, N Wareham, C J Willer, P A Zalloua, J Erdmann, P Deloukas, H Watkins, H Schunkert, J Danesh, J R Thompson, N J Samani, and C ARDIoGRAM+C4D Consortium. Genetically determined height and coronary artery disease. *N Engl J Med*, 372(17):1608–1618, 2015.

[76] NIH. What was the international hapmap project? `https://www.genome.gov/11511175/about-the-international-hapmap-project-fact-sheet/`, 2016.

[77] M Nikpay, A Goel, H H Won, L M Hall, C Willenborg, S Kanoni, D Saleheen, T Kyriakou, C P Nelson, J C Hopewell, T R Webb, **L Zeng**, A Dehghan, M Alver, S M Armasu, K Auro, A Bjonnes, D I Chasman, S Chen, I Ford, N Franceschini, C Gieger, C Grace, S Gustafsson, J Huang, S J Hwang, Y K Kim, M E Kleber, K W Lau, X Lu, Y Lu, L P Lyytikainen, E Mihailov, A C Morrison, N Pervjakova, L Qu, L M Rose, E Salfati, R Saxena, M Scholz, A V Smith, E Tikkanen, A Uitterlinden, X Yang, W Zhang, W Zhao, M de Andrade, P S de Vries, N R van Zuydam, S S Anand, L Bertram, F Beutner, G Dedoussis, P Frossard, D Gauguier, A H Goodall, O Gottesman, M Haber, B G Han, J Huang, S Jalilzadeh, T Kessler, I R Konig, L Lannfelt, W Lieb, L Lind, C M Lindgren, M L Lokki, P K Magnusson, N H Mallick, N Mehra, T Meitinger, F U Memon, A P Morris, M S Nieminen, N L Pedersen, A Peters, L S Rallidis, A Rasheed, M Samuel, S H Shah, J Sinisalo, K E Stirrups, S Trompet, L Wang, K S Zaman, D Ardissino, E Boerwinkle, I B Borecki, E P Bottinger, J E Buring, J C Chambers, R Collins, L A Cupples,

J Danesh, I Demuth, R Elosua, S E Epstein, T Esko, M F Feitosa, O H Franco, M G Franzosi, C B Granger, D Gu, V Gudnason, A S Hall, A Hamsten, T B Harris, S L Hazen, C Hengstenberg, A Hofman, E Ingelsson, C Iribarren, J W Jukema, P J Karhunen, B J Kim, J S Kooner, I J Kullo, T Lehtimaki, R J Loos, O Melander, A Metspalu, W Marz, C N Palmer, M Perola, T Quertermous, D J Rader, P M Ridker, S Ripatti, R Roberts, V Salomaa, D K Sanghera, S M Schwartz, U Seedorf, A F Stewart, D J Stott, J Thiery, P A Zalloua, C J O'Donnell, M P Reilly, T L Assimes, J R Thompson, J Erdmann, R Clarke, H Watkins, S Kathiresan, R McPherson, P Deloukas, H Schunkert, N J Samani, and M Farrall. A comprehensive 1,000 Genomes-based genome-wide association meta-analysis of coronary artery disease. *Nat Genet*, 47(10):1121–1130, 2015.

[78] Leducq Transatlantic Network of Excellence. Understanding coronary artery disease genes. `https://www.fondationleducq.org/network/understanding-coronary-artery-disease-genes/`, 2013.

[79] R S Patel, Y V Sun, J Hartiala, E Veledar, S Su, S Sher, Y X Liu, A Rahman, R Patel, S T Rab, V Vaccarino, A M Zafari, H Samady, W H Tang, H Allayee, S L Hazen, and A A Quyyumi. Association of a genetic risk score with prevalent and incident myocardial infarction in subjects undergoing coronary angiography. *Circ Cardiovasc Genet*, 5(4):441–449, 2012.

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

[81] P C Phillips. Epistasis–the essential role of gene interactions in the structure and evolution of genetic systems. *Nat Rev Genet*, 9(11):855–867, 2008.

[82] B Phipson and G K Smyth. Permutation P-values should never be zero: calculating exact P-values when permutations are randomly drawn. *Stat Appl Genet Mol Biol*, 9:Article39, 2010.

[83] M Pirinen, P Donnelly, and C C Spencer. Including known covariates can reduce power to detect genetic effects in case-control studies. *Nat Genet*, 44(8):848–851, 2012.

[84] A C Poncelet and H W Schnaper. Sp1 and Smad proteins cooperate to mediate transforming growth factor-beta 1-induced alpha 2(I) collagen expression in human glomerular mesangial cells. *J Biol Chem*, 276(10):6983–6992, 2001.

[85] S Purcell, B Neale, K Todd-Brown, L Thomas, M A Ferreira, D Bender, J Maller, P Sklar, P I de Bakker, M J Daly, and P C Sham. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*, 81(3):559–575, 2007.

[86] R. The r project for statistical computing. `https://www.r-project.org`, 2016.

[87] M Ruiz-Ortega, J Rodriguez-Vita, E Sanchez-Lopez, G Carvajal, and J Egido. TGF-beta signaling in vascular fibrosis. *Cardiovasc Res*, 74(2):196–206, 2007.

[88] T B Sackton and D L Hartl. Genotypic Context and Epistasis in Individuals and Populations. *Cell*, 166(2):279–287, 2016.

[89] N Samani, J Erdmann, A S Hall, C Hengstenberg, A Ziegler, J R Thompson, and H Schunkert. Genomewide Association Analysis of Coronary Artery Disease. *N Engl J Med*, 357(5):443–453, 2007.

[90] H Schunkert, I R Konig, S Kathiresan, M P Reilly, T L Assimes, H Holm, M Preuss, A F Stewart, M Barbalic, C Gieger, D Absher, Z Aherrahrou, H Allayee, D Altshuler, S S Anand, K Andersen, J L Anderson, D Ardissino, S G Ball,

A J Balmforth, T A Barnes, D M Becker, L C Becker, K Berger, J C Bis, S M Boekholdt, E Boerwinkle, P S Braund, M J Brown, M S Burnett, I Buysschaert, Cardiogenics, J F Carlquist, L Chen, S Cichon, V Codd, R W Davies, G Dedoussis, A Dehghan, S Demissie, J M Devaney, P Diemert, R Do, A Doering, S Eifert, N E Mokhtari, S G Ellis, R Elosua, J C Engert, S E Epstein, U de Faire, M Fischer, A R Folsom, J Freyer, B Gigante, D Girelli, S Gretarsdottir, V Gudnason, J R Gulcher, E Halperin, N Hammond, S L Hazen, A Hofman, B D Horne, T Illig, C Iribarren, G T Jones, J W Jukema, M A Kaiser, L M Kaplan, J J Kastelein, K T Khaw, J W Knowles, G Kolovou, A Kong, R Laaksonen, D Lambrechts, K Leander, G Lettre, M Li, W Lieb, C Loley, A J Lotery, P M Mannucci, S Maouche, N Martinelli, P P McKeown, C Meisinger, T Meitinger, O Melander, P A Merlini, V Mooser, T Morgan, T W Muhleisen, J B Muhlestein, T Munzel, K Musunuru, J Nahrstaedt, C P Nelson, M M Nothen, O Olivieri, R S Patel, C C Patterson, A Peters, F Peyvandi, L Qu, A A Quyyumi, D J Rader, L S Rallidis, C Rice, F R Rosendaal, D Rubin, V Salomaa, M L Sampietro, M S Sandhu, E Schadt, A Schafer, A Schillert, S Schreiber, J Schrezenmeir, S M Schwartz, D S Siscovick, M Sivananthan, S Sivapalaratnam, A Smith, T B Smith, J D Snoep, N Soranzo, J A Spertus, K Stark, K Stirrups, M Stoll, W H Tang, S Tennstedt, G Thorgeirsson, G Thorleifsson, M Tomaszewski, A G Uitterlinden, A M van Rij, B F Voight, N J Wareham, G A Wells, H E Wichmann, P S Wild, C Willenborg, J C Witteman, B J Wright, S Ye, T Zeller, A Ziegler, F Cambien, A H Goodall, L A Cupples, T Quertermous, W Marz, C Hengstenberg, S Blankenberg, W H Ouwehand, A S Hall, P Deloukas, J R Thompson, K Stefansson, R Roberts, U Thorsteinsdottir, C J O'Donnell, R McPherson, J Erdmann, C ARDIoGRAM Consortium, and N J Samani. Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. *Nat Genet*, 43(4):333–338, 2011.

[91] N Solovieff, C Cotsapas, P H Lee, S M Purcell, and J W Smoller. Pleiotropy in

complex traits: challenges and strategies. *Nat Rev Genet*, 14(7):483–495, 2013.

[92] D Speed, G Hemani, M R Johnson, and D J Balding. Improved heritability estimation from genome-wide SNPs. *Am J Hum Genet*, 91(6):1011–1021, 2012.

[93] N O Stitziel, K Stirrups, N Masca, J Erdmann, P G Ferrario, I R Konig, H Watkins, C J Willer, S Kathiresan, P Deloukas, N Samani, and H Schunkert. Coding Variation in ANGPTL4, LPL, and SVEP1 and the Risk of Coronary Disease. *N Engl J Med*, 374(12):1134–1144, 2016.

[94] P H Sudmant, T Rausch, E J Gardner, R E Handsaker, A Abyzov, J Huddleston, Y Zhang, K Ye, G Jun, M Hsi-Yang Fritz, M K Konkel, A Malhotra, A M Stutz, X Shi, F Paolo Casale, J Chen, F Hormozdiari, G Dayama, K Chen, M Malig, M J Chaisson, K Walter, S Meiers, S Kashin, E Garrison, A Auton, H Y Lam, X Jasmine Mu, C Alkan, D Antaki, T Bae, E Cerveira, P Chines, Z Chong, L Clarke, E Dal, L Ding, S Emery, X Fan, M Gujral, F Kahveci, J M Kidd, Y Kong, E W Lameijer, S McCarthy, P Flicek, R A Gibbs, G Marth, C E Mason, A Menelaou, D M Muzny, B J Nelson, A Noor, N F Parrish, M Pendleton, A Quitadamo, B Raeder, E E Schadt, M Romanovitch, A Schlattl, R Sebra, A A Shabalin, A Untergasser, J A Walker, M Wang, F Yu, C Zhang, J Zhang, X Zheng-Bradley, W Zhou, T Zichner, J Sebat, M A Batzer, S A McCarroll, Consortium Genomes Project, R E Mills, M B Gerstein, A Bashir, O Stegle, S E Devine, C Lee, E E Eichler, and J O Korbel. An integrated map of structural variation in 2,504 human genomes. *Nature*, 526(7571):75–81, 2015.

[95] Tg, National Heart Lung Hdl Working Group of the Exome Sequencing Project, Institute Blood, J Crosby, G M Peloso, P L Auer, D R Crosslin, N O Stitziel, L A Lange, Y Lu, Z Z Tang, H Zhang, G Hindy, N Masca, K Stirrups, S Kanoni, R Do, G Jun, Y Hu, H M Kang, C Xue, A Goel, M Farrall, S Duga, P A Merlini, R Asselta, D Girelli, O Olivieri, N Martinelli, W Yin, D Reilly, E Speliotes, C S

Fox, K Hveem, O L Holmen, M Nikpay, D N Farlow, T L Assimes, N Franceschini, J Robinson, K E North, L W Martin, M DePristo, N Gupta, S A Escher, J H Jansson, N Van Zuydam, C N Palmer, N Wareham, W Koch, T Meitinger, A Peters, W Lieb, R Erbel, I R Konig, J Kruppa, F Degenhardt, O Gottesman, E P Bottinger, C J O'Donnell, B M Psaty, C M Ballantyne, G Abecasis, J M Ordovas, O Melander, H Watkins, M Orho-Melander, D Ardissino, R J Loos, R McPherson, C J Willer, J Erdmann, A S Hall, N J Samani, P Deloukas, H Schunkert, J G Wilson, C Kooperberg, S S Rich, R P Tracy, D Y Lin, D Altshuler, S Gabriel, D A Nickerson, G P Jarvik, L A Cupples, A P Reiner, E Boerwinkle, and S Kathiresan. Loss-of-function mutations in APOC3, triglycerides, and coronary disease. *N Engl J Med*, 371(1):22–31, 2014.

[96] D A Tregouet, I R Konig, J Erdmann, A Munteanu, P S Braund, A S Hall, A Grosshennig, P Linsel-Nitschke, C Perret, M DeSuremain, T Meitinger, B J Wright, M Preuss, A J Balmforth, S G Ball, C Meisinger, C Germain, A Evans, D Arveiler, G Luc, J B Ruidavets, C Morrison, P van der Harst, S Schreiber, K Neureuther, A Schafer, P Bugert, N E El Mokhtari, J Schrezenmeir, K Stark, D Rubin, H E Wichmann, C Hengstenberg, W Ouwehand, Consortium Wellcome Trust Case Control, Consortium Cardiogenics, A Ziegler, L Tiret, J R Thompson, F Cambien, H Schunkert, and N J Samani. Genome-wide haplotype association study identifies the SLC22A3-LPAL2-LPA gene cluster as a risk locus for coronary artery disease. *Nat Genet*, 41(3):283–285, 2009.

[97] A W Turner, M Nikpay, A Silva, P Lau, A Martinuk, T A Linseman, S Soubeyrand, and R McPherson. Functional interaction between COL4A1/COL4A2 and SMAD3 risk loci for coronary artery disease. *Atherosclerosis*, 242(2):543–552, 2015.

[98] R Virmani, F D Kolodgie, A P Burke, A Farb, and S M Schwartz. Lessons from

sudden coronary death: a comprehensive morphological classification scheme for atherosclerotic lesions. *Arterioscler Thromb Vasc Biol*, 20(5):1262–1275, 2000.

[99] P M Visscher, M A Brown, M I McCarthy, and J Yang. Five years of GWAS discovery. *Am J Hum Genet*, 90(1):7–24, 2012.

[100] S Volkel, B Stielow, F Finkernagel, T Stiewe, A Nist, and G Suske. Zinc finger independent genome-wide binding of Sp2 potentiates recruitment of histone-fold protein Nf-y distinguishing it from Sp1 and Sp3. *PLoS Genet*, 11(3):e1005102, 2015.

[101] K Wang, M Li, and H Hakonarson. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res*, 38(16):e164, 2010.

[102] H Watkins and M Farrall. Genetic susceptibility to coronary artery disease: from promise to progress. *Nat Rev Genet*, 7(3):163–173, 2006.

[103] W H Wei, G Hemani, A Gyenesei, V Vitart, P Navarro, C Hayward, C P Cabrera, J E Huffman, S A Knott, A A Hicks, I Rudan, P P Pramstaller, S H Wild, J F Wilson, H Campbell, N D Hastie, A F Wright, and C S Haley. Genome-wide analysis of epistasis in body mass index using multiple human populations. *Eur J Hum Genet*, 20(8):857–862, 2012.

[104] 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):661–678, 2007.

[105] D Welter, J MacArthur, J Morales, T Burdett, P Hall, H Junkins, A Klemm, P Flicek, T Manolio, L Hindorff, and H Parkinson. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res*, 42(Database issue):D1001–6, 2014.

[106] Wikipedia. Human genetic variation — wikipedia, the free encyclopedia. https://en.wikipedia.org/w/index.php?title=Human_genetic_variation&oldid=723066697, 2016.

[107] Wikipedia. International hapmap project — wikipedia, the free encyclopedia. https://en.wikipedia.org/w/index.php?title=International_HapMap_Project&oldid=725414359, 2016.

[108] B R Winkelmann, W Marz, B O Boehm, R Zotz, J Hager, P Hellstern, and J Senges. Rationale and design of the LURIC study–a resource for functional genomics, pharmacogenomics and long-term prognosis of cardiovascular disease. *Pharmacogenomics*, 2(1 Suppl 1):S1–73, 2001.

[109] T W Winkler, F R Day, D C Croteau-Chonka, A R Wood, A E Locke, R Magi, T Ferreira, T Fall, M Graff, A E Justice, J Luan, S Gustafsson, J C Randall, S Vedantam, T Workalemahu, T O Kilpelainen, A Scherag, T Esko, Z Kutalik, I M Heid, R J Loos, and Consortium Genetic Investigation of Anthropometric Traits. Quality control and conduct of genome-wide association meta-analyses. *Nat Protoc*, 9(5):1192–1212, 2014.

[110] J Yang, T Ferreira, A P Morris, S E Medland, ANthropometric Traits Consortium Genetic Investigation of, D IAbetes Genetics Replication, Consortium Meta-analysis, P A Madden, A C Heath, N G Martin, G W Montgomery, M N Weedon, R J Loos, T M Frayling, M I McCarthy, J N Hirschhorn, M E Goddard, and P M Visscher. Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat Genet*, 44(4):369–75, S1–3, 2012.

[111] J Yang, S H Lee, M E Goddard, and P M Visscher. GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet*, 88(1):76–82, 2011.

[112] **L Zeng**, T Kessler, H Schunkert, and ... Genetic risk prediction with multiple independent signals at known risk loci of coronary artery disease. 2016.

[113] **L Zeng**, N Mirza-Schreiber, B Müller-Myhsok, and H Schunkert. Identification of novel trans-epistasis effects in coronary artery disease. 2016.

[114] P Zhang, K M Tchou-Wong, and M Costa. Egr-1 mediates hypoxia-inducible transcription of the NDRG1 gene through an overlapping Egr-1/Sp1 binding site in the promoter. *Cancer Res*, 67(19):9125–9133, 2007.

[115] G Zheng, J Joo, C Zhang, and N L Geller. Testing association for markers on the X chromosome. *Genet Epidemiol*, 31(8):834–843, 2007.

[116] A Ziegler and I R Konig. *A Statistical Approach to Genetic Epidemiology*. 2 edition, 2012.

[117] O Zuk, E Hechter, S R Sunyaev, and E S Lander. The mystery of missing heritability: Genetic interactions create phantom heritability. *Proc Natl Acad Sci U S A*, 109(4):1193–1198, 2012.

Appendix A

Supplementary Tables

Gene	Tissue	CAD	p (lm)	p.perm (lm)	p (anova)	p.perm (anova)	p.cor
TMEM176A	mon	cases	0.00136	0.000581	0.000892	0.000581	1.24E-05
RHOQ	mon	controls	0.000119	0.000581	7.60E-05	0.000581	0.000152
SPTBN2	mon	all	0.0017	0.000581	0.0012	0.000581	0.000187
SPTAN1	mon	cases	0.000138	0.000581	0.000317	0.000581	0.000217
PPP2R5A	mac	all	1.85E-06	0.000581	4.31E-06	0.000581	0.000226
PPP2R5A	mac	controls	0.000154	0.000581	0.000122	0.000581	0.000228
PDE8A	mac	all	3.40E-05	0.000581	1.20E-05	0.000581	0.00023
PGD	mon	all	0.000596	0.000581	0.000344	0.000581	0.000257
SIGLEC10	mac	all	0.00021	0.000581	0.00041	0.000581	0.000296
SDC1	mon	all	0.000579	0.000581	0.000359	0.000581	0.000314
RFPL3S	mon	all	0.00516	0.000581	0.00783	0.002496	0.000592
STK32C	mac	controls	0.000354	0.000581	0.00019	0.000581	0.000728

Continued on next page

Table A.1 – *Continued from last page*

Gene	Tissue	CAD	p (lm)	p.perm (lm)	p (anova)	p.perm (anova)	p.cor
SPTBN2	mon	cases	0.000747	0.000581	0.00124	0.001501	0.000786
C15orf48	mac	controls	0.00233	0.000581	0.00647	0.003496	0.000794
SNX13	mon	all	0.00118	0.000581	0.0021	0.000581	0.000953
C14orf1	mon	cases	0.000138	0.000581	0.000158	0.000581	0.000957
GSG1	mon	controls	0.000698	0.000581	0.00105	0.000581	0.00137
PMPCA	mon	controls	3.00E-04	0.000581	0.000256	0.000581	0.00149
DIRC2	mon	controls	0.000179	0.000581	0.000175	0.000581	0.00159
FHIT	mac	cases	0.000374	0.000581	0.000484	0.000581	0.00185
SLC9A8	mac	controls	0.000541	0.000581	0.000192	0.000581	0.00209
ATG2B	mac	controls	0.000581	0.000581	0.000729	0.000581	0.00247
KLF1	mon	controls	0.000422	0.000581	0.000686	0.000581	0.00283
CD69	mac	cases	0.000774	0.000581	0.000347	0.000581	0.00298
TPRKB	mon	cases	0.000397	0.000581	0.000191	0.000581	0.00301
MAP2K7	mac	controls	0.00164	0.000581	0.000962	0.000581	0.00341
PEBP1	mac	controls	0.000217	0.000581	0.000179	0.001501	0.00342
VPS13A	mon	controls	0.00131	0.000581	0.00162	0.001501	0.00384
UBE2H	mon	all	0.000921	0.000581	0.0011	0.000581	0.00388
EPB41	mon	controls	0.000959	0.000581	0.00181	0.004495	0.00461
TMEM176A	mac	cases	0.000132	0.001501	7.99E-05	0.001501	5.55E-06
TMEM176B	mon	cases	0.00155	0.001501	0.00108	0.000581	3.73E-05

Continued on next page

Table A.1 – *Continued from last page*

Gene	Tissue	CAD	p (lm)	p.perm (lm)	p (anova)	p.perm (anova)	p.cor
C20orf24	mon	all	0.00101	0.001501	0.000698	0.001501	0.000102
LARGE	mon	cases	0.00211	0.001501	0.00152	0.001501	0.000115
UTP23	mon	cases	0.00145	0.001501	0.0013	0.001501	0.00021
NLRP3	mon	all	0.00219	0.001501	0.000898	0.000581	0.000211
RARB	mon	cases	0.0025	0.001501	0.00431	0.001501	0.000291
ZCRB1	mon	controls	0.000513	0.001501	0.000346	0.001501	0.000357
ABL2	mon	controls	2.37E-05	0.001501	1.31E-05	0.001501	0.000527
TNXB	mac	all	0.00309	0.001501	0.00428	0.002496	0.000689
NUCB2	mac	controls	0.00111	0.001501	0.00135	0.001501	0.00081
MAEA	mon	all	0.00332	0.001501	0.0017	0.000581	0.000929
CDC7	mac	controls	0.000802	0.001501	0.000774	0.002496	0.0012
EIF2C4	mac	all	0.00084	0.001501	0.000372	0.000581	0.00122
PSMA1	mon	all	0.00259	0.001501	0.00398	0.002496	0.0013
FLJ11795	mon	all	0.000553	0.001501	0.000416	0.002496	0.00131
RIT1	mon	all	0.00389	0.001501	0.0031	0.001501	0.00132
GPR139	mac	cases	0.00454	0.001501	0.00597	0.001501	0.00141
MGC18216	mac	controls	0.000399	0.001501	0.000635	0.000581	0.00144
PGBD4	mon	controls	0.00163	0.001501	0.00217	0.002496	0.00155
SCNM1	mac	cases	0.00221	0.001501	0.00181	0.001501	0.00161
THUMPD3	mac	controls	0.000786	0.001501	0.000913	0.002496	0.00247

Continued on next page

Table A.1 – *Continued from last page*

Gene	Tissue	CAD	p (lm)	p.perm (lm)	p (anova)	p.perm (anova)	p.cor
NAIP	mon	all	0.00107	0.001501	0.0014	0.001501	0.00253
PADI4	mac	all	0.00064	0.001501	0.000869	0.001501	0.00262
PROK1	mon	all	0.00281	0.001501	0.0032	0.002496	0.00283
C19orf15	mon	controls	0.00239	0.001501	0.00308	0.002496	0.00285
WTAP	mon	all	0.000973	0.001501	0.000736	0.001501	0.00301
UBFD1	mon	all	0.000856	0.001501	0.00099	0.001501	0.00306
PROCR	mac	controls	0.00272	0.001501	0.00138	0.000581	0.00321
MT1M	mac	all	0.00245	0.001501	0.00142	0.000581	0.0033
MURC	mac	all	0.00299	0.001501	0.00304	0.003496	0.00332
C20orf197	mon	all	0.00158	0.001501	0.00281	0.002496	0.00406
MOV10	mon	all	0.00174	0.001501	0.0018	0.001501	0.00454
TMEM176B	mac	cases	6.90E-05	0.002496	3.52E-05	0.002496	1.63E-06
ATP5F1	mon	all	0.00246	0.002496	0.00234	0.002496	1.20E-05
THAP10	mon	controls	0.00169	0.002496	0.000833	0.000581	1.38E-05
PIBF1	mon	controls	0.00137	0.002496	0.00225	0.003496	0.000305
SEC22C	mon	all	0.00331	0.002496	0.00245	0.002496	0.000351
PSRC1	mon	controls	0.00198	0.002496	0.000846	0.001501	0.000356
SSSCA1	mon	all	0.00148	0.002496	0.000931	0.001501	0.000463
ZNF280C	mon	all	0.00111	0.002496	0.000458	0.001501	0.000567
LACTB	mac	cases	0.00223	0.002496	0.00162	0.002496	0.000757

Continued on next page

Table A.1 – *Continued from last page*

Gene	Tissue	CAD	p (lm)	p.perm (lm)	p (anova)	p.perm (anova)	p.cor
STXBP1	mon	cases	0.00119	0.002496	0.00118	0.002496	0.000818
GAPDH	mon	cases	0.00276	0.002496	0.00146	0.002496	0.000833
SYNJ2BP	mon	controls	0.00747	0.002496	0.00548	0.002496	0.000936
TNFRSF10A	mac	cases	0.00353	0.002496	0.00132	0.000581	0.00104
PDE8A	mac	cases	0.0014	0.002496	0.00089	0.001501	0.00199
CCM2	mon	all	0.000587	0.002496	0.000588	0.002496	0.00215
PRKCE	mac	all	0.00118	0.002496	0.000813	0.002496	0.00225
C14orf129	mon	all	0.00346	0.002496	0.00241	0.002496	0.00251
CCNF	mon	cases	0.00103	0.002496	0.000786	0.001501	0.00274
C21orf66	mon	controls	0.00214	0.002496	0.00187	0.002496	0.00291
LEO1	mon	all	0.00656	0.002496	0.00617	0.002496	0.00305
ZNF10	mac	controls	0.00126	0.002496	0.000827	0.002496	0.00311
NHEDC2	mac	all	0.00342	0.002496	0.00425	0.002496	0.00423
PRKCZ	mon	cases	0.00523	0.002496	0.00519	0.003496	0.00469
GTF2A2	mon	controls	0.00327	0.003496	0.00188	0.003496	9.27E-05
SIGLEC1	mon	controls	0.00459	0.003496	0.00332	0.001501	0.000132
PLXNA3	mon	cases	0.00385	0.003496	0.00441	0.004495	0.000383
CD160	mac	cases	0.00246	0.003496	0.00388	0.003496	0.000626
TNPO1	mac	all	0.00193	0.003496	0.00108	0.001501	0.000673
MSRA	mon	all	0.00279	0.003496	0.00234	0.003496	0.000737

Continued on next page

Table A.1 – *Continued from last page*

Gene	Tissue	CAD	p (lm)	p.perm (lm)	p (anova)	p.perm (anova)	p.cor
GZMB	mac	controls	0.00311	0.003496	0.00235	0.003496	0.000893
DLD	mon	all	0.00437	0.003496	0.00373	0.003496	0.00123
XAF1	mon	all	0.00577	0.003496	0.00297	0.001501	0.00199
TBC1D22B	mon	all	0.0054	0.003496	0.00583	0.003496	0.00231
ABCC12	mon	controls	0.0038	0.003496	0.00231	0.001501	0.00251
SEC22C	mon	all	0.00163	0.003496	0.00104	0.001501	0.00257
GNAS	mac	all	0.00343	0.003496	0.00145	0.001501	0.00282
C10orf4	mac	cases	0.00499	0.003496	0.00534	0.004495	0.00315
CDC42EP3	mon	cases	0.005	0.003496	0.00296	0.002496	0.00323
TMEM185B	mon	controls	0.00212	0.003496	0.00155	0.001501	0.00344
IFIH1	mac	controls	0.00229	0.003496	0.00264	0.002496	0.00371
CPEB2	mac	cases	0.00215	0.003496	0.00178	0.003496	0.00375
MBNL3	mon	cases	0.00468	0.003496	0.00193	0.002496	0.00379
PHF11	mon	controls	0.0102	0.003496	0.00705	0.002496	0.00451
TNKS1BP1	mon	all	0.00221	0.004495	0.00154	0.003496	0.00102
IL19	mon	controls	0.00347	0.004495	0.00282	0.003496	0.00118
BNC1	mon	all	0.0105	0.004495	0.0042	0.002496	0.00119
PDE8A	mon	all	0.00277	0.004495	0.00292	0.004495	0.00131
CLN8	mac	all	0.00351	0.004495	0.00109	0.001501	0.00138
FGFBP2	mon	all	0.00505	0.004495	0.00307	0.003496	0.00209

Continued on next page

Table A.1 – *Continued from last page*

Gene	**Tissue**	**CAD**	**p (lm)**	**p.perm (lm)**	**p (anova)**	**p.perm (anova)**	**p.cor**
C4orf18	mon	cases	0.00556	0.004495	0.00274	0.001501	0.00227
PLD1	mac	all	0.00218	0.004495	0.00122	0.001501	0.00246
JAZF1	mon	controls	0.00323	0.004495	0.00101	0.000581	0.00294
C17orf64	mon	all	0.00287	0.004495	0.00375	0.003496	0.00456

Table A.1: Potential genes intermediate between epistasis pair and CAD. The tissues are either monocytes(mon) or macrophages(mac); the individual groups are CAD cases, controls, or combined them all. P-values for linear regression and ANOVA test are listed both as original and as permutation adjusted; p-values for the correlation test ($<5 \times 10^{-3}$) between the expression level and the putative CAD odds ratio are listed here.

Appendix B

List of Own Publications

Thesis work

1. Nikpay M, Goel A, Won HH, et al, **Zeng L**. A comprehensive 1,000 Genomes-based genome-wide association meta-analysis of coronary artery disease. Nat Genet 47, 1121-30 (2015).

2. Loley C, Alver M, Assimes TL, et al, **Zeng L**. No Association of Coronary Artery Disease with X-Chromosomal Variants in Comprehensive International Meta-Analysis. Scientific Reports 2016. (under review)

3. **Zeng L**, et al. Genetic risk prediction with multiple independent signals at known risk loci of coronary artery disease. (2016). (manuscript in preparation)

4. Nelson CP, Hamby SE, Saleheen D, et al, **Zeng L**. Genetically determined height and coronary artery disease. N Engl J Med 2015;372:1608-18.

5. Jansen H, Willenborg C, Lieb W, **Zeng L**, et al. Rheumatoid arthritis and coronary artery disease: Genetic analyses do not support a causal relation. J Rheumatol 2016. (accepted)

6. **Zeng L**, Mirza-Schreiber N. et al. Identification of novel trans-epistasis effects in coronary artery disease. 2016. (manuscript in prepartation)

Further work during PhD research

7. **Zeng L**⋆, Dang TA⋆, Schunkert H. Genetics links between transforming growth factor β pathway and coronary disease. Atherosclerosis 2016. pii: S0021-9150(16)31290-4.
8. Stitziel NO, Stirrups K, Masca N, et al, **Zeng L**. Coding Variation in ANGPTL4, LPL, and SVEP1 and the Risk of Coronary Disease. N Engl J Med 2016;374:1134-44.
9. Braenne I, Civelek M, Vilne B, et al, **Zeng L**. Prediction of Causal Candidate Genes in Coronary Artery Disease Loci. Arterioscler Thromb Vasc Biol 2015;35:2207-17.
10. Interleukin 1 Genetics Consortium et al, **Zeng L**. Cardiometabolic effects of genetic upregulation of the interleukin 1 receptor antagonist: a Mendelian randomisation analysis. The Lancet Diabetes Endocrinology 2015;3:243-53.
11. LeBlanc M, Zuber V, Andreassen BK, et al, **Zeng L**. Identifying Novel Gene Variants in Coronary Artery Disease and Shared Genes With Several Cardiovascular Risk Factors. Circ Res 2016;118:83-94.
12. Braenne I, **Zeng L**, et al. Genomic correlates of glatiramer adverse cardiovascular effects lead to a novel locus mediating coronary risk. (manuscript in prepartation)
13. von Scheidt M, Zhao Y, Kurt Z, et al, **Zeng L**. Applications and Limitations of Mouse Models for Understanding Human Atherosclerosis. Cell Metabolism 2016. (under review)

Earlier publications

14. Tao ZH, Wan JL, **Zeng LY**, et al. miR-612 suppresses the invasive-metastatic cascade in hepatocellular carcinoma. J Exp Med 2013;210:789-803.
15. **Zeng L**, Yu J, Huang T, et al. Differential combinatorial regulatory network analysis related to venous metastasis of hepatocellular carcinoma. BMC Genomics 2012;13 Suppl 8:S14.
16. Yuan W, Huang T, Yu J, et al, **Zeng L**. Comparative analysis of viral protein interaction networks in Hepatitis B virus and Hepatitis C virus infected HCC. Biochim Biophys Acta 2014;1844:271-9.
17. Yu J, Xing X, **Zeng L**, et al. SyStemCell: a database populated with multiple levels of experimental data from stem cell differentiation research. PLoS One 2012;7:e35230.