Technische Universität München

Fakultät für Medizin

Searching for epistasis in complex traits: Genome-wide and biological knowledge-driven approaches in coronary artery disease

Sylvain Cédric Moser

Vollständiger Abdruck der von der Fakultät für Medizin der Technischen Universität München zur Erlangung eines Doktors der Naturwissenschaften (Dr. rer. nat.) genehmigten Dissertation.

Vorsitz: Prof. Dr. Lars Mägdefessel

Prüfer*innen der Dissertation:

1. apl. Prof. Dr. Bertram Müller-Myhsok
2. Prof. Dr. Julien Gagneur

Die Dissertation wurde am 27.09.2022 bei der Technischen Universität München eingereicht und durch die Fakultät für Medizin am 21.02.2023 angenommen.

Abstract

In the last decade, Genome-wide Association Studies (GWAS) have brought the field of statistical genetics in the big data era and discovered thousands of genetic variants associated with various traits and diseases. While they tremendously helped understanding the genetic architecture of complex traits and to generate new hypothesis of disease etiologies, all those variants still explain only a small proportion of the estimated genetic heritability.

The presence of genetic interactions or epistasis is one of the mechanisms that could explain this "missing heritability". However, because of challenges linked to the combinatorial nature of epistasis, there are up to date very few reported examples of epistasis in human traits. This absence of data, combined with a theoretical controversy on the partition of the genetic variance in additive and non-additive variance have led to high skepticism toward epistasis in human genetics.

Fortunately, improving computational resources start allowing genome-wide interactions associations studies (GWIAS) with the promise of bringing experimental evidence for the presence and nature of genetic interactions in human traits. In this thesis, we report the first hypothesis free GWIAS for coronary artery disease (CAD) and thereby show the feasibility of such studies using to a two-step analysis strategy. This approach successfully identified 17 interacting SNP-pairs located in the SLC22A3-LPAL2-LPA gene cluster and in the CDKN2A-CDKN2B region, which are both previously identified CAD risk loci.

To complement the hypothesis free approach, we conducted two epistasis scans using a-priori biological filters; one targeting regulatory SNPs and a second focusing on SNPs with previous evidence of interactions. The two approaches were however unable to detect any significant interactions, highlighting the difficulty to generate meaningful a-priori filters and thus the advantages of hypothesis-free methods.

In addition, we report two methodological advances tackling current challenges in epistasis studies. Firstly, we developed the epilogitpower R package which allows to compute the power or necessary sample size in GWIAS. Secondly, we established a method to disentangle epistasis signals from haplotype effects and which allowed us to discover interactions involving rs140570886, a rare variant at the LPA locus previously associated with CAD.

In conclusion, we developed computational methods that help making epistasis studies easier to conduct and more robust. Applying these methods on two samples of CAD patients, we brought some of the first evidence regarding the presence and genetic architecture of epistasis in a complex human disease.

Zusammenfassung

In den letzten 10 Jahren haben genomweite Assoziationsstudien (*Genome-wide Association Studies*, GWAS) die Statistische Genetik in die *big data* Ära gebracht und dabei Tausende von genetischen Varianten entdeckt, die mit verschiedensten Eigenschaften und Krankheiten assoziieren. Obwohl dies enorm zum Verständnis der genetischen Grundlagen von komplexen Eigenschaften beigetragen hat und auch neue Hypothesen zu Krankheitetiologien hervorgebracht hat, erklären all dieser Varianten doch nur einen kleinen Teil der geschätzten genetischen Vererbung.

Genetische Interaktionen, auch Epistase genannt, stellen einen Mechanismus dar, der diese „fehlende Vererbung“ (*missing heritability*) erklären könnte. Aufgrund der in der kombinatorischen Natur von Epistasiseffekten begründeten Herausforderungen gibt es bislang nur wenige belegte Beispiele für Epistasis beim Menschen. Dieses Fehlen von Daten verbunden mit der theoretischen Kontroverse bzgl. der Aufteilung der genetischen Varianz in additiv und nicht-additiv hat zu einer starken Skepsis gegenüber Epistase im Bereich der Humangenetik geführt.

Glücklicherweise erlauben besser werdende Computerressourcen inzwischen genomweite Interaktions-Assoziationsstudien (GWIAS) und verheißen experimentelle Evidenz für Vorhandensein und Grundlagen der genetischen Interaktionen in menschlichen Eigenschaften. In dieser Arbeit berichten wir über die erste hypothesenfreie GWIAS in einer großen Stichprobe bei Koronararterienerkrankung (*coronary artery disease*, CAD) und beweisen damit die Durchführbarkeit solcher Studien mittels eines zweistufigen Analysestrategie. Dieser Ansatz identifizierte 17 interagierende SNP-Paare, die im SLC22A3-LPAL2-LPA Gen-Cluster und der CDKN2A-CDKN2B Region lokalisiert sind. Beide Bereiche waren bereits als CAD-Risiko-Loci identifiziert worden.

Um den hypothesenfreien Ansatz zu komplementieren haben wir zwei Epistase-Scans mit Einsatz von *a priori*-Filtern durchgeführt. Der eine zielt auf regulatorische SNPs, der andere auf SNPs mit früherer Evidenz für Interaktionen. Leider konnten beide Ansätze keine signifikanten Interaktionen finden. Dies unterstreicht die Schwierigkeit wirksame *a priori*-Filter zu finden und damit die Vorteile hypothesenfreier Methoden.

Darüber hinaus beschreiben wir zwei methodische Weiterentwicklungen, die aktuelle Herausforderungen bei Epistase-Studien adressieren. Zum einen haben wir das R-Paket epilogitpower entwickelt, das es erlaubt Trennschärfe bzw. benötigte Gruppengröße für GWIAS zu berechnen. Zweitens haben wir eine Methode etabliert, mit der Epistase- von Haplotyp-Effekten unterschieden werden können. Dadurch konnten wir Interaktionen mit rs140570886, einer schon früher mit CAD assoziierten, seltenen Variation im LPA-Lokus, entdecken.

Zusammenfassend haben wir Rechenmethoden entwickelt, mit Hilfe derer Epistase-Studien einfacher und robuster durchgeführt werden können. Mit ihrer Anwendung auf zwei CAD-Patientengruppen fanden wir erste Belege für Existenz und genetische Grundlagen von Epistase in komplexen humanen Krankheiten.

Acknowledgements

I would like to thank my first supervisor, Prof. Müller-Myhsok for his support all along my PhD. I am very thankful to you for giving me the chance to join the Statistical Genetics group and transitioning from the wet lab to the computational side. I would like to express my sincere gratitude for your kind supervision. You did not only share your expertise and helped me being a better scientist, but also offered me, through your caring support, the peace of mind that I needed to succeed in these different projects. And, maybe the most important, you reminded me that science can be fun and productive at the same time.

I would also like to thank my second advisor, Prof. Gagneur, for his guidance and insightful suggestions all along my thesis. A special thank you to Prof. Schunkert, whose support and expertise has been key in these projects. Thank you as well for your interest in epistasis and for giving us access to great quality data.

Moreover, I would like to express my gratitude to my mentor, Dr. Till Andlauer for his help and availability throughout this PhD. The precision and quality of your work and teaching have been and will stay for me an example and a motivation to continue improving my skills. I would also like to thank Dr. Benno Pütz for his patient and meticulous help on all computing related matters but also well beyond. Thank you Benno for looking after all of us. Many thanks to Dr. Nazanin Mirza-Schreiber for introducing me to the core principles of Statistical Genetics and her warm welcome in the group.

This PhD would not have been so fun and exciting without the colleagues and friends from this institute with whom I shared the happy and the difficult moments. Thank you Ane, Ezgi, Lucia, Lucas, Helena, Dunja, Riya and Min.

Thank you Fabrizia for having been by my side every single day of this PhD, for cheering at the successes; patiently listening to my complaints, often; motivating me, in the hard times; making me believe in myself, when needed; and making life exciting, always.

My deepest gratitude goes to my family and my parents for always believing in me and supporting me during the sometimes tortuous path that led me to this thesis. I am forever grateful for your education, the values that you taught me, and the love that you gave me.

This research has been conducted using the UK Biobank Resource under Application Number 25214 (https://www.ukbiobank.ac.uk/enable-your-research/approved-research/investigating-multi-locus-epistatic-interactions-underlying-coronary-artery-disease).

Table of Figures

Table of Tables

List of Acronyms

AIC	Akaike information criterion
ANOVA	Analysis of Variance
CAD	Coronary Artery Diseases
CDCV	Common Disease - Common Variant
CDKN2A	Cyclin Dependent Kinase Inhibitor 2A
CDNK2B	Cyclin Dependent Kinase Inhibitor 2B
CFTR	Cystic fibrosis transmembrane conductance regulator
CH	Compound heterozygosity
CPU	Central processing unit
DNA	Deoxyribonucleic acid
FDR	False Discovery Rate
GPU	Graphics processing unit
GRR	Genotypic Odds Ratio
GTEX	Genotype-Tissue Expression
GWIAS	Genome-Wide Interaction Association Study
GWAS	Genome-Wide Association Study
HPC	High Performance Computing
HWE	Hardy-Weinberg Equilibrium
LD	Linkage-Disequilibrium
LPA	Lipoprotein(A)
LPAL2	Lipoprotein(A) Like 2, Pseudogene
LRT	Likelihood Ratio Test
MAF	Minor Allele Frequency
MC	Monte Carlo
MDS	Multi-Dimensional Scaling
NGS	Next-Generation Sequencing
OR	Odds Ratio
PC	Principal Component
PCA	Principal Component Analysis
PLG	Plasminogen

QC	Quality Control
QTL	Quantitative-Trait Locus
RNA	Ribonucleic Acid
SCL22A1/2/3	Solute Carrier Family 22 Member 1/2/3
SEM	Standard Error of the Mean
SNP	Single Nucleotide Polymorphism
TF	Transcription Factor
TSS	Transcription Start Site
UKBB	UK Biobank

1 Introduction

1.1 Genetic epidemiology and association studies

Since Plato already, humans and philosophers have been trying to disentangle the influence of nature and nurture, of what is innate and what is acquired during life. While the mechanisms underlying the effect of the innate characteristics and their transmission remained hidden for centuries, the DNA discovery by Watson, Crick, and Franklin in 1953 finally enabled scientists to tackle this question with modern technologies. Providing a mechanistic explanation and a physical basis to Mendel's law, the discovery of DNA would open the way for the development of genetic epidemiology.

Epidemiologists who were before trying to understand the relation between readily measurable biological characteristics, such as height, weight, scull dimensions, etc., and traits of interests, such as diseases or cognitive abilities, could thereafter create a new scientific discipline associating genetic differences among people with biological traits and diseases. Morton, one of the founders of genetic epidemiology defines this new discipline as: "a science that deals with the etiology and control of disease in groups of relatives and with inherited causes of disease in populations" (Morton, 1982). With the development of the DNA sequencing method by Sanger, genetic epidemiologist met their first success in establishing direct causation between mutations in single genes and Mendelian diseases such as the causal mutation of CFTR causing cystic fibrosis (Kerem et al., 1989; Riordan et al., 1989; Rommens et al., 1989).

1.1.1 Single-Nucleotide Polymorphisms and Genome-wide Association Analysis

1.1.1.1 Single Nucleotide Polymorphisms (SNPs)

The genomic DNA molecules, organized in 22 autosomal and 2 sex-specific chromosomes in humans, are long double-stranded chains of nucleotides, or bases. Each strand of DNA is made of a repetition of 4 different nucleotides Adenine (A), Thymine (T), Guanine (G), and Cytosine (C) which can bind to another nucleotide on the opposite strand to form a double helix structure. Because of their different chemical properties, an Adenine always binds to a Thymine and a Guanine always binds to a Cytosine. The alternation of these 4 types of bases along the DNA follows a 3-codon code which can be deciphered by messenger RNAs during translation and encodes the assembly instruction for all proteins of the organism.

Single Nuclear Polymorphisms (SNPs) are variable positions in the genome at which a change in base-pair happens in a certain percentage of individuals in a population (Figure 1). Most of the time, an Adenine (A) is exchanged for a Guanine (G) or a Cytosine (C) is changed for a Thymine (T). Since the beginning of the 21st century, increasing attention has been paid to SNPs to capture genetic diversity between individuals. Indeed, as stated by Wang and colleagues who conducted one of the first screening for such polymorphisms, "This genetic diversity is of interest because it explains the basis of heritable variation in disease susceptibility, as well as harbors a record of human migrations" (D. G. Wang et al., 1998).

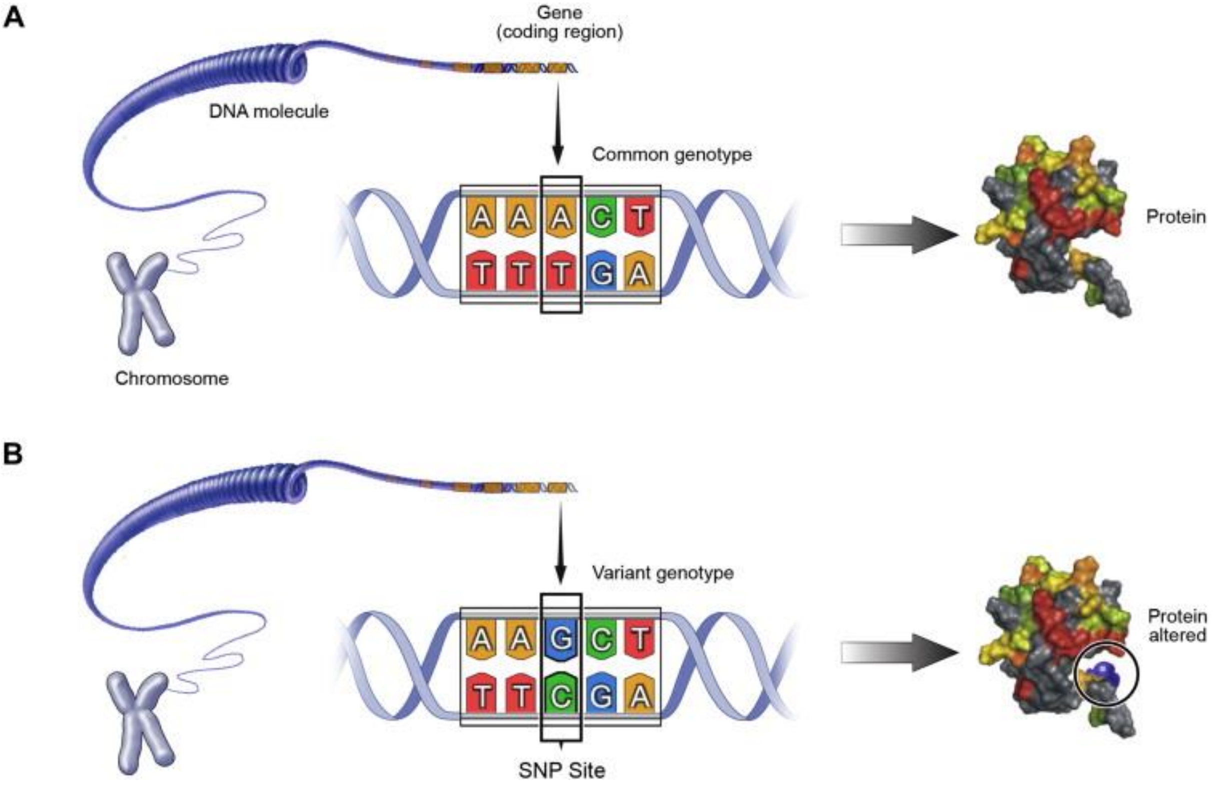


Figure 1: Single Nuclear Polymorphisms
"Single nucleotide polymorphisms (SNPs) are small sequence differences within genes where the DNA sequences of many individuals vary by a single base; not all SNPs result in structural protein changes. For example, some people may have a chromosome with an A at a particular site where others have a chromosome with a G". Figure and legend from reproduced from (Camp & Trujillo, 2014) with the permission of Elsevier Inc.

Today, the largest reference panels contain up to 40 million SNPs covering the whole genome (J. Huang et al., 2015). A large part of the genome, around 98% in mammals, formerly called junk-DNA, does not code for proteins (Shabalina & Spiridonov, 2004). The important role of these non-coding regions has been the focus of extensive research in the last decades which revealed that 80% percent of the genome has a biochemical function (Dunham et al., 2012).

Non-coding DNA regions have two main functional roles. Firstly, they can code for non-translated RNA transcripts which are involved in multiple functions such as transcription regulation, alternative splicing, or alternative splicing (Shabalina & Spiridonov, 2004). Secondly, these non-coding DNA sequences can form motives that are recognized by enhancers and transcription factors (TF) to regulate gene expression, or form topologically associating domains (TAD) which can change the 3D conformation of chromatin. (Spielmann & Mundlos, 2016). Depending on the location of the SNPs in the genome, these base substitutions can therefore perform their action by changing the sequence of amino acids of a protein or by changing the properties of a non-coding regulatory sequence.

1.1.1.2 GWAS and linkage disequilibrium

Before the wide availability of SNP data using microarrays, or more-recently next-generation sequencing (NGS), genetic associations were mostly investigated using linkage analysis. In this technique, disease-causing genes were identified by their co-inheritance, and therefore close genetic distance, with genetic markers of known chromosomal location and the phenotype of interest. Such approaches could identify genes responsible for Mendelian diseases (Visscher, Brown, McCarthy, & Yang, 2012) but were not successful in complex disorders as they could only scan small portions of the genome (the loci close to a known genetic marker).

The first successes in identifying genetic bases of complex disorders came with the advent of SNP-based Genome-Wide Association Analysis (GWAS). In a GWAS, the association of a large number of single SNPs with the phenotype is tested in a univariate manner. In an analogy to the linkage analysis, each of these SNPs can be used as a genetic marker, to test if a region of the genome is associated with the phenotype of interest, because of linkage disequilibrium (LD).

LD is defined as the non-random association of alleles at different loci. According to the principle of chromosome linkage, two loci situated on the same chromosomes should be inherited jointly and therefore in LD. However, recombination events can happen, which exchange two fragments of DNA between sister chromatids and therefore "unlink" two alleles. Generation after generation, the occurrence of recombination keeps fragmenting the blocks of jointly associated loci (or LD-blocks) to produce a more fine-grained LD structure (Figure 2). Given that the LD blocks become smaller, the rate of LD decreases with the number of generations and ultimately, after an infinite number of generations, all loci would reach linkage equilibrium. However, in the present population of humans, LD persists and can be used efficiently for gene mapping. The association of one "target SNP" therefore indicates that this particular SNP is causally associated with the phenotype of interest or that another SNPs, in LD with the tag-SNP is. Theoretically, the presence of one SNP in each LD-block would allow to potentially discover associations anywhere in the genome. In practice, several factors such as the high number of SNPs needed to properly cover all LD-block of the genome, the rare frequencies of tag-SNP and causal SNP, or the presence of population stratification greatly complicates the tasks. Most of these challenges are shared with the interaction association studies and will be discussed in more detail later in this chapter.

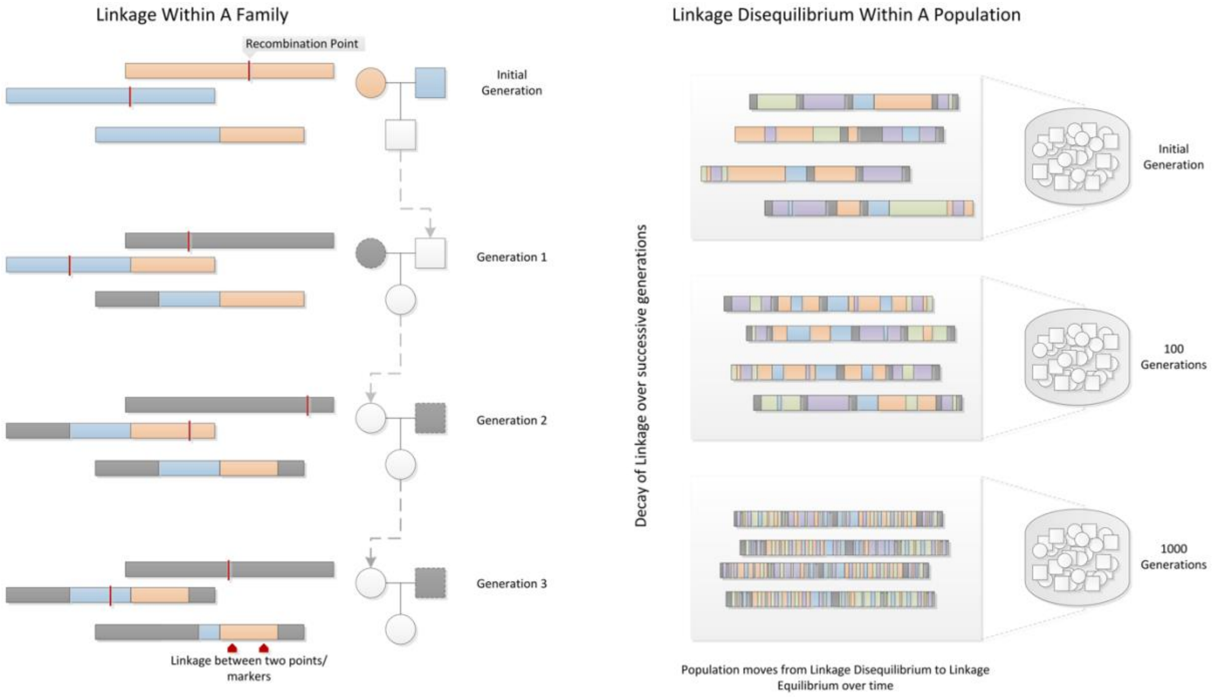

Linkage Within A Family
Recombination Point
Initial Generation
Generation 1
Generation 2
Generation 3
Linkage between two points/ markers
Linkage Disequilibrium Within A Population
Decay of Linkage over successive generations
Initial Generation
100 Generations
1000 Generations
Population moves from Linkage Disequilibrium to Linkage Equilibrium over time

Figure 2: Linkage and Linkage Disequilibrium
"Within a family, linkage occurs when two genetic markers (points on a chromosome) remain linked on a chromosome rather than being broken apart by recombination events during meiosis, shown as red lines. In a population, contiguous stretches of founder chromosomes from the initial generation are sequentially reduced in size by recombination events. Over time, a pair of markers or points on a chromosome in the population move from linkage disequilibrium to linkage equilibrium, as recombination events eventually occur between every possible point on the chromosome." Figure and legend from (Bush & Moore, 2012).

1.1.2 The "common disease – common variant" hypothesis and the missing heritability

With the failure of linkage analysis to identify genetic loci associated with non-Mendelian disorders and the first GWAS results for complex disorders, a new hypothesis emerged concerning the genetic basis of common, complex disorders. Namely, it started to appear that common diseases are associated with genetic variants which have high minor allele frequency (MAF) and relatively low-risk contribution. As a consequence of the high MAF of these variants, as compared to the frequency of the variants or mutation causing Mendelian diseases, their genetic effect must be small. Indeed, if they are highly frequent and had a strong effect, then the common disease would be much more frequent than they actually are. Nonetheless, common disorders were shown to be heritable to some extent by family studies, suggesting that multiple variants act together to modulate the disease risk (Bush & Moore, 2012). Accordingly, the "common disease – common variant" (CDCV) hypothesis "argues that genetic variations with appreciable frequency in the population at large, but relatively low 'penetrance' (or the probability that a carrier of the relevant variants will express the disease), are the major contributors to genetic susceptibility to common diseases" (Schork, Murray, Frazer, & Topol, 2009).

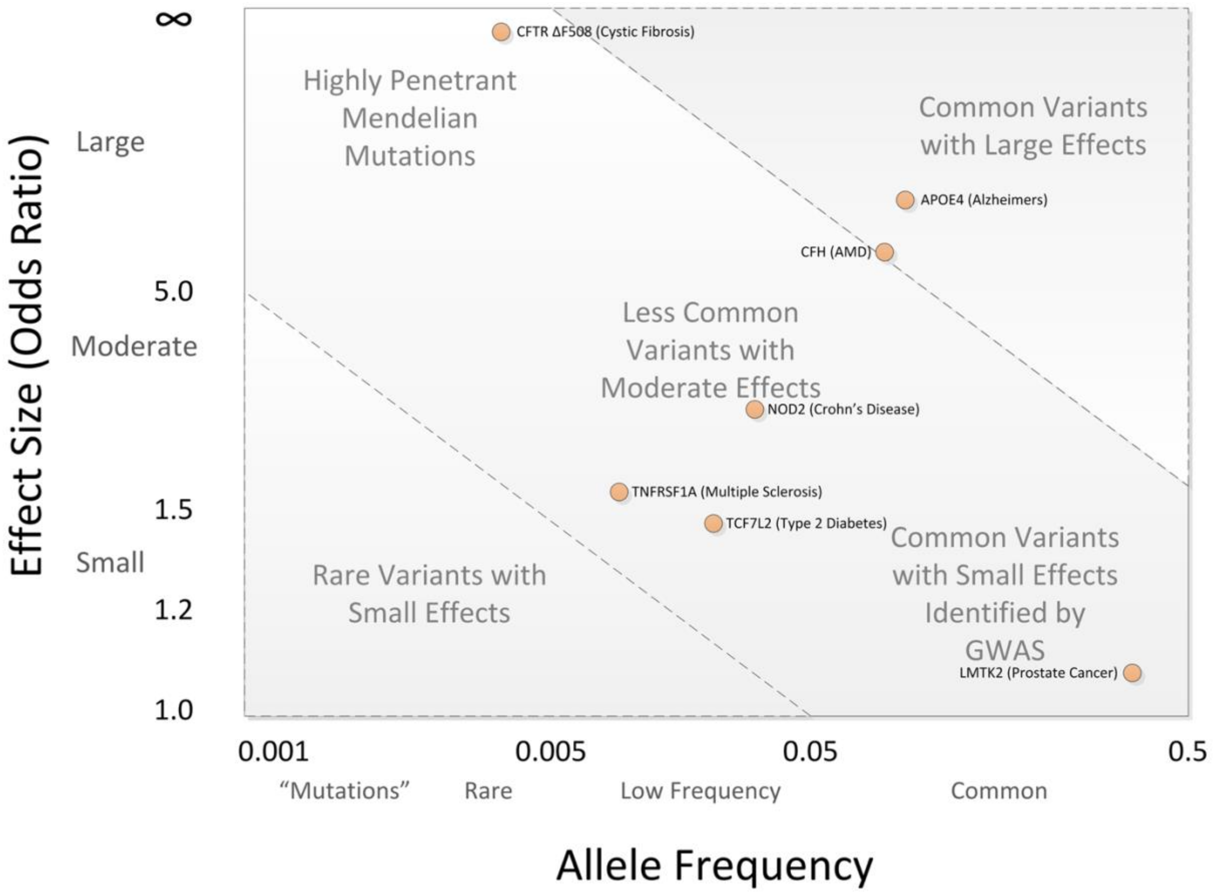


Figure 3: Common Disease – Common variant hypothesis
"Disease associations are often conceptualized in two dimensions: allele frequency and effect size. Highly penetrant alleles for Mendelian disorders are extremely rare with large effect sizes (upper left), while most GWAS findings are associations of common SNPs with small effect sizes (lower right). The bulk of discovered genetic associations lies on the diagonal denoted by the dashed lines". Figure and legend from (Bush & Moore, 2012)

More than a decade of GWAS studies have now brought overwhelming evidence in favor of the CDCV hypothesis and discovered more than 10’000 strong associations between variants and complex traits. More precisely it showed that complex traits are highly polygenic but also that there is a high degree of pleiotropy, meaning that causal variants are shared between complex traits (Visscher et al., 2017).

With ever-decreasing genotyping costs and therefore increasing sample sizes, more and more variants, with lower MAF and lower effect sizes keep being discovered every day. Consequently, the proportion of the estimated heritability of complex traits which can be explained by the cumulative action of the discovered variants keeps rising. However, the proportion of the heritability explained by all discovered variants, h^2_{SNP}, still represents only a fraction of the estimate of pedigree heritability, $\mathrm{h}^2_{\mathrm{ped}}$, for most complex traits. For example, in the case of height, a trait that has been shown to be very heritable with an estimated heritability of near 80%, the ca. 700 discovered variants by 2014 could only explain ca. 20% of $\mathrm{h}^2_{\mathrm{ped}}$. These discrepancies led to the formulation of the concept of “missing heritability”.

Multiple factors have been suggested to contribute to this “missing heritability”. Among them, are rare variants with possibly larger effects, structural variants poorly captured by existing arrays or inadequate accounting for shared environment between relatives in the pedigree estimates (Manolio et al., 2009). Alternatively, or additionally, it has been suggested that interactions between genes (epistasis) could contribute to the genetic architecture of complex diseases and to the missing heritability.

Indeed, biological pathways and mechanisms are made of interactions between biomolecules. It is therefore expected that interactions between genes contribute to the genetic variance of complex traits. In agreement with this hypothesis, large-scale experiments have shown that gene interactions play a crucial role in genotype-phenotype relationships in yeast (Costanzo et al., 2016; Kuzmin et al., 2018) for example. Experimental evidence of epistasis is not only provided in unicellular organisms but also in invertebrate animal models (see (Mackay, 2014) for a review). For example, epistasis has been shown to be involved in odor-guided behavior in drosophila (Anholt et al., 2003), a highly complex and important biological system.

1.2 Epistasis in complex human diseases

The role of epistasis in complex human traits has been very debated in the last decades. Although the important role of epistasis in unicellular and pluricellular animal models has been demonstrated, part of the human genetics community strongly believes that all genetic contribution to the phenotypic variance in humans is of additive nature (Hill, Goddard, & Visscher, 2008). Partly because of theoretical results regarding genetic variance partitioning (Mackay & Moore, 2014), and partly because only few studies could convincingly demonstrate epistasis in human traits, such as cardiovascular disease (Zeng et al., 2022) or Alzheimer's disease (Combarros, Cortina-Borja, Smith, & Lehmann, 2009) for example.

The difficulty to demonstrate epistasis in human disorders can most likely be explained by two factors. Firstly, the term epistasis is commonly used with two slightly different definitions, related to a biological concept or to a statistical concept. Although these definitions are often used interchangeably, and although statistical epistasis is used to investigate biological epistasis, these two concepts are not equivalent. Secondly, the search for statistical epistasis at the genome-wide levels faces several critical computational and statistical challenges which have not been fully solved yet.

1.2.1 Two definitions of epistasis

Although the term epistasis refers broadly to genetic interactions, there has been over the years several different definitions of epistasis, leading to some confusion among geneticists.

1.2.1.1 Biological Epistasis

Bateson introduced the word epistasis in 1907 with a definition of what we nowadays could call biological epistasis. It is defined as a masking effect whereby the effect of a variant or allele at one locus depends on a variant at another locus. This corresponds to the more restricted definition of "compositional epistasis" as coined by Phillips in an attempt to disambiguate the meaning of the word epistasis (Phillips, 2008). The presence of compositional epistasis does imply that the two alleles, or their transcribed forms (proteins or non-coding RNAs), interact at the molecular level, in what Phillips would call „functional epistasis". Extending this definition to accommodate higher-order interactions, between three or more loci, led to the definition of epistasis prevailing in modern system biology. These interactions between molecules in the same pathways, or between pathways, at the level of individual organisms are what Genome-wide interactions analyses (GWIAs) try to discover, as they can bring new insights into the etiology of the disease as well as generate new therapeutical candidates.

1.2.1.2 Statistical epistasis

Contrary to animal models, where biological epistasis can de directly investigated with the help simple dihybrid cross, for example, it cannot be directly assessed at scale in humans. In human genetics, epistasis needs to be investigated in observational studies using statistical models. In Fisher's original definition, statistical epistasis is the phenotypic variance that cannot be explained by the additive or dominant component of the genotypic variance (Sackton & Hartl, 2016). In modern statistical genetics, statistical epistasis is often defined as the deviation from an additive linear model where two or more alleles predict a phenotypic outcome (Cordell, 2009). While this definition of statistical epistasis seems to be a good mathematical translation of biological epistasis, these two definitions are not equivalent. Indeed, as pointed out by Cheverud and Routman, "statistical epistasis is a population phenomenon depending on allele frequencies present in a specific population whereas physiological epistasis is a genotypic phenomenon, independent of allele frequencies at the loci in question." (Cheverud & Routman, 1995). This leads to the fact that there can be biological epistasis in absence of statistical epistasis and vice-versa (J. H. Moore & Williams, 2005). Even though one needs to be cautious when interpreting statistical epistasis findings in regard to their biological implication, statistical methods remain the only solution to investigate the presence of genetic interaction at the scale of the whole genome in human populations and disorders.

1.2.2 Challenges in detecting statistical epistasis

Most of the challenges in investigating statistical epistasis at the genome-wide levels are common with GWAS analysis but their effect is amplified in GWIAs. Among these challenges, we find the relation between MAF and power, the role of LD, the high number of statistical tests performed, and population stratification.

1.2.2.1 Linkage Disequilibrium

LD, defined as the non-random association of alleles at different loci due to population history (as described in 1.1.1.2), has several implications on GWAS and GWIAs.

Firstly, different sub-population histories lead to different LD structures. Indeed, we can observe for example that in African-descent populations, LD regions are smaller than in European-descent or Asiatic-descent populations, because African-descent populations are older and more recombinations accumulated. In a sample with individuals from potentially different ancestry, these different population genetic substructures could possibly confound associations and need to be accounted for.

Secondly, LD implies some degree of collinearity between SNPs. This has two particular implications for statistical testing.

On the one hand, it induces dependencies between the interactions hypothesis tests. Because of these dependencies, multiple testing correction using Bonferroni, which is typically used in GWAS and on which the canonical genome-wide significant threshold of 5×10^{-8} has been derived, tends to be over-conservative and lead to increase rates of type II error (Van Steen & Moore, 2019). As suggested by Van Steen and Moore, False Discovery Rate (FDR) control methods such as the one proposed by Benjamini and Yukutelli should be able to account for dependencies between tests. However, these methods seem to not work properly in the case of GWIAs with large-scale data and highly complex correlations between test (Van Steen & Moore, 2019).

On the other hand, dependencies between variants might also give rise to harmful multicollinearity in regression models, leading to over-estimation of the standard regression of regression parameter and therefore to deflation of the statistical significance (Van Steen & Molenberghs, 2012).

Together, the two previous arguments suggest that SNPs should be LD-pruned for GWIAs to remove collinearities. However, the whole principle of GWAS and GWIAs relies on the fact that the SNPs used in the analysis perform the role of tag or marker for untested variants in their LD block because of their correlation. While the amount of shared signal decreases linearly with the LD for a variant with respect to its tag-SNP in a GWAS, it decreases quadratically for the interaction term between two tagged variants with respect to their two tag-SNPs in an epistasis analysis. Consequently, stronger LD pruning, leading to a sparser "tag-SNP grid" might lead to missing interactions between untested variants. This counter-balances the previous arguments suggesting that strong LD pruning is needed in GWIAs. For these reasons, LD and LD-pruning remain important parameters in GWIAs for which there is currently no consensus concerning the best practices.

1.2.2.2 Population stratification:

Common genetic background among certain individuals in a sample can lead to confounding associations between genetic markers and a trait of interest. Indeed, if variant A is more common in individuals with a certain common ancestry, and that, because of causal genetic variant B or environmental risk factors the trait studied is also more common in this sub-population, we will observe a spurious association with variant A. The impact of population structure, or stratification, on epistasis studies, does not seems to have been particularly well studied but Gusavera and colleagues suggest that it would be even more pronounced than in GWAS (Gusareva & Van Steen, 2014). They also point out that population structure might decrease the power of finding genetic interactions.

In GWAS, population structure is most of the time dealt with by including Principal Components (PC) of the genetic variance as covariates in regression models. These PCs are usually computed on the whole study sample and capture the largest proportion of the genetic variance, which is typically due to different population descent (Bush & Moore, 2012). By extension, this method has also been widely used in GWIAs when using regression models as methods.

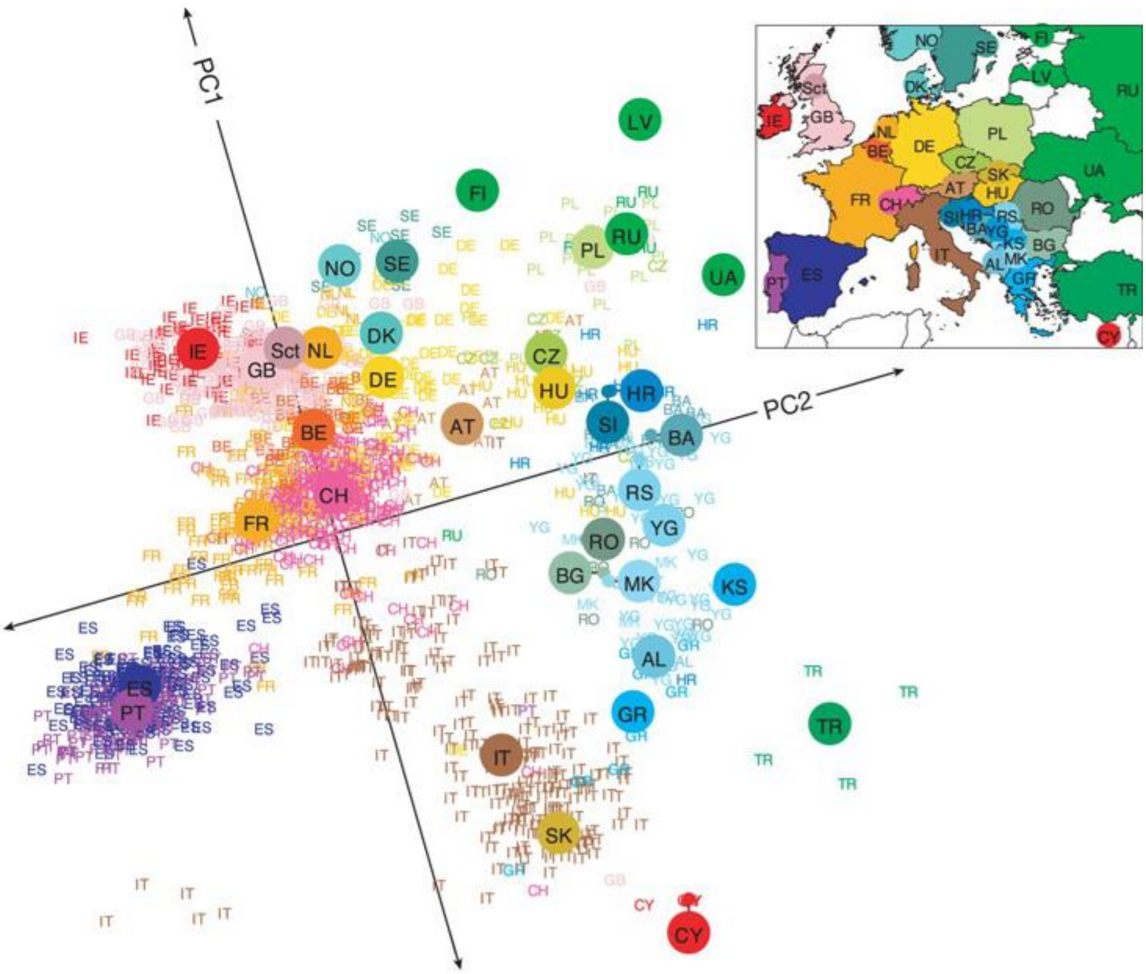


Figure 4: Principal Component of genetic variance
Principal Component Analysis (PCA) of the genetic variance of 3000 European individuals reproduces very closely the geographical map of Europe, thereby suggesting that PCA captures well the population structure. Figure adapted from (Novembre et al., 2008) with the permission of Springer Nature.

1.2.2.3 Minor Allele Frequencies:

The frequency of the rarest allele at a particular locus is called the minor allele frequency (MAF) and has profound implications regarding the power to detect an association at this locus. Indeed, very logically, the rarer this allele is, the rarer is also the disease manifestation, and therefore the number of cases in a population if the variant is causal. The power to detect rare variants is therefore low. Moreover, the potentially low resulting number of minor allele homozygous individuals violates the large sample assumption of the linear models and can lead to erroneous test results. In particular, it has been reported that the Wald test, commonly used to assess the statistical significance of the marginal effect or interaction effect of SNPs, behaves aberrantly for rare variants. It would report highly non-significant p-values if a “causal variant with high penetrance is present at low frequency in the cases and nearly absent from the controls“ (Xing, Lin, Wooding, & Xing, 2012).

Importantly, there is a double penalty to include these rare variants in GWAS or GWIAs; not only have these variants a very low chance to be significantly associated if they are causal, but they also reduce the power to detect more frequent variance because they increase the number of tests and concomitantly the multiple correction burden.

These problems are even more exacerbated in GWIAs as compared to GWAS because the frequency of multi-loci minor-allele homozygous combinations is the product of the frequency of minor-allele homozygous at each locus. Namely if one considers SNP A and SNP B, both with a MAF of 0.1, the minor-allele homozygous frequencies for both SNP A and SNP B (so a/a and b/b genotypes) are 0.01. However, the frequency of individuals who are minor allele homozygous at the locus A and B (a/a;b/b) is 0.0001. Therefore, it is important to define appropriate MAF thresholds, as function of the sample size, expected effect size, and genetic transmission model, by doing proper power calculation (as discussed more in detail in Section 2.3.2).

1.2.2.4 High number of pairwise combinations

One of the biggest challenges in studying epistasis is the combinatorial nature of genetic interactions which leads to an extremely high number of statistical tests to perform and consequently to a high multiple testing correction burden and low statistical power. In addition, the technical requirement needed to achieve enough computational power for epistasis study were not common in research institutes for a long time. However, in the last years, advances in computational methods and hardware have made it possible to consider conducting epistasis studies at the level of the whole genome.

Genome-wide interaction studies can be performed using two different approaches to take care of the problem arising from the high number of SNP pairs. In the "hypothesis-free" approach, researchers aim to exhaustively test the whole interaction space for associations. This most of the time requires a two-step strategy where a first testing step is used to reduce the search space for a second more fine-grained testing step. Such approaches also most often rely on fast testing algorithms and parallelized computation or computation on graphical processing units (GPUs). Their main advantage is to not rely on any prior biological knowledge and therefore allow to find genetic interactions between genes or pathways that might not have already been associated with the disorder. They potentially represent high hopes for the development of new insights into the disease and innovative therapies. However, these studies are still confronted with a high multiple testing correction burden and typically will only have a high power to detect interacting SNP pairs with a moderate to strong effect size and a relatively high MAF in big samples. In Section 3, we will present one of the first attempts to run a hypothesis-free genome-wide epistasis screening for Coronary Artery Diseases (CAD) using a two-step strategy.

On the other side of this trade-off between novelty potential and test number, filtering methods have been developed to reduce the search space using prior knowledge. While allowing to detect smaller effects in smaller samples, they are dependent on present knowledge about biological interactions or the etiology of the disease. There exist two different classes of "filtering" approaches. The first one proposes to consider variants whose marginal effects are associated with the disease. The second one is to consider SNPs related to genes showing putative functional interaction by, e.g., being involved in the same pathways or biological processes.

Marginal effect filtering relies on the idea that variants whose marginal effects have been associated with the disease are also more likely to be involved in interactions affecting this disease. Either because the apparent additive effect is actually resulting from an underlying true epistatic effect, or simply because they play a role in pathways that are involved in the disease's etiology. This filtering technique was used in, for example, the attached publication (Zeng et al., 2022) described in chapter 5.

Functional variant filtering is based on the observation that most of the variants discovered by GWAS studies are located in non-coding DNA regions and are thus likely to regulate gene expression, one can limit the epistasis search space to such regulatory variants. Another approach is to consider genes that are likely to biologically interact at different levels of cell processes.

The former idea can be implemented by scrutinizing publicly available eQTL databases to find regulatory variants. Additionally, newly developed deep learning models can predict with high accuracy which variants are likely to influence gene expression (Zhou et al., 2018).

The latter idea is based on the assumption that genes involved in the same pathways, or implicated in the same biological processes are likely to influence each other's effect. This has indeed been confirmed by a huge effort to characterize the full network of interaction between the 6000 genes of the yeast (Kuzmin et al., 2018). Practical implementation of this method requires cross-checking of different databases, such as Gene-Ontology, gene-gene interactions, and protein-protein interactions databases. While doing this manually would be a daunting task, there exist softwares, such as Biofilter from Bush and colleagues, which were precisely developed to infer plausible interaction from various knowledge sources and can generate a list of SNPs to test in interactions studies (Bush, Dudek, & Ritchie, 2008).

In the absence of better characterization of the genetic architecture of complex traits, for example about whether or not the genes involved in epistatic interactions also have a marginal effect, it is impossible to predict which of the filtering strategies will yield the more biologically and clinically interesting results, or if they complement each other. Researchers might therefore have to try several of them and assess which one yields the most interesting results for their disease and context of interest.

1.2.2.5 Statistical Power

Frequentist statistical methods, such as the one used in linear or logistic regression models in GWAS and GWIAs, rely on hypothesis testing. It typically implies the computation of a test statistic whose distribution is known and typically centered at zero in the case where the variable of interest has no effect on the outcome (Figure 5 A). This is referred to as the Null hypothesis. The actual value of the test statistic under the alternative hypothesis, which is the value computed in the observed sample, can be compared to its distribution under the null hypothesis. This gives information about how unlikely such a test statistic is under the null hypothesis. The percentile of this null distribution which corresponds to the observed test statistic is called the p-value and can be interpreted as the likelihood to observe such an extreme finding if there was no effect (i.e., if the null hypothesis was true). A small p-value represents low statistical evidence for the null hypothesis, and this hypothesis is typically rejected (and the effect considered present) if the p-value is lower than 5%. For a given effect in the population, the distribution of the test statistic under the alternative hypothesis also has a known distribution, which is shifted towards its true value in the population (Figure 5 B). The concept of power is closely linked to the separation between these two distributions. Indeed, a true observed effect, distributed according to the alternative distribution, is statistically significant at a threshold α if it does not lie within the 1– α a percentile of the null distribution. The further away the two distributions, the bigger the area of the alternative distribution which does not fall in this region (statistically significant), and the higher the power. The power, therefore, corresponds to the area under the alternative distribution (which is 1) minus its overlap with the last α percent of the null distribution (called β) (Figure 5 B). The β region represents cases when the alternative hypothesis was true, but the null hypothesis cannot be rejected. These cases are called false negatives. On the contrary, a true null effect which is in the extreme α percent of the null distribution could be falsely considered significant.

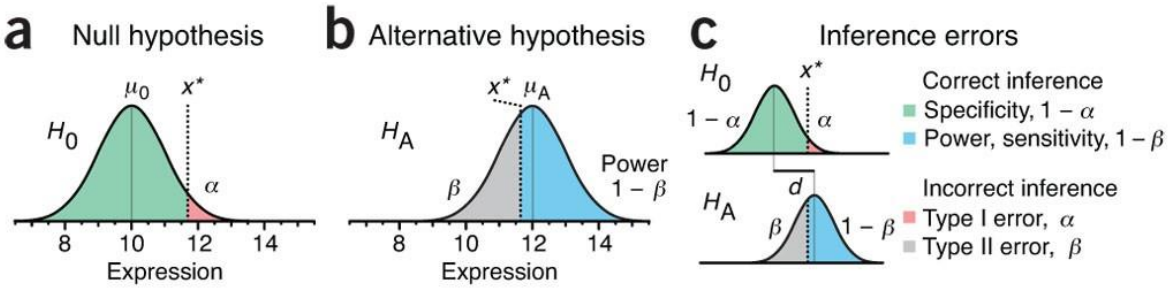


Figure 5: Statistical Power - The Null and the Alternative hypothesis
“(a) Observations are assumed to be from the null distribution (H_0) with mean μ_0. We reject H_0 for values larger than $x^*$with an error rate α(red area). (b) The alternative hypothesis (H_A) is the competing scenario with a different mean μ_A. Values sampled from H_A smaller than x^*do not trigger rejection of H_0 and occur at a rate β. Power (sensitivity) is 1 – β (blue area). (c) Relationship of inference errors to x^*.” Figure and legend from (Krzywinski & Altman, 2013) with the permission of Springer Nature.

We, therefore, see that there is a trade-off between the false positive and false negative rate which is influenced by the threshold α. A larger α threshold will decrease the rate of false-negative but increase the rate of false positive (Figure 6).

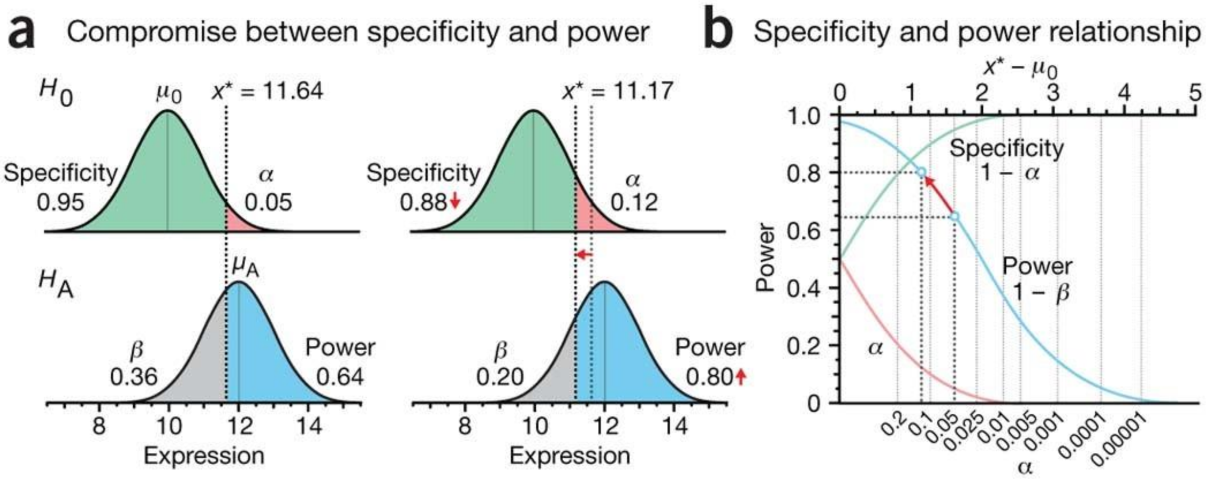


Figure 6: Statistical Power - Power vs false positives trade-off
“Decreasing specificity increases power. H_0 and H_A are assumed normal with σ = 1. (a) Lowering specificity decreases the H_0 rejection cutoff x^*, capturing a greater fraction of H_A beyond x^*, and increases the power from 0.64 to 0.80. (b) The relationship between specificity and power as a function of x^*.” The open circles correspond to the scenarios in a. Figure and legend from (Krzywinski & Altman, 2013) with the permission of Springer Nature.

It appears from Figure 7 that the more separated the two distributions are, i.e., the stronger the observed effect is, the higher the power. Importantly, for the same separation, the narrower the two distribution, the higher the power as well. The width of the two distributions is proportional to the standard error of the mean (SEM). Because the SEM is computed as $\frac{\sigma}{\sqrt{n}}$ it appears that higher sample sizes lead to a smaller SEM and therefore to a greater separation of the distribution and concomitantly to a higher power. In summary, the power of a study is influenced both by the strength of the effect and by the sample size (Figure 7).

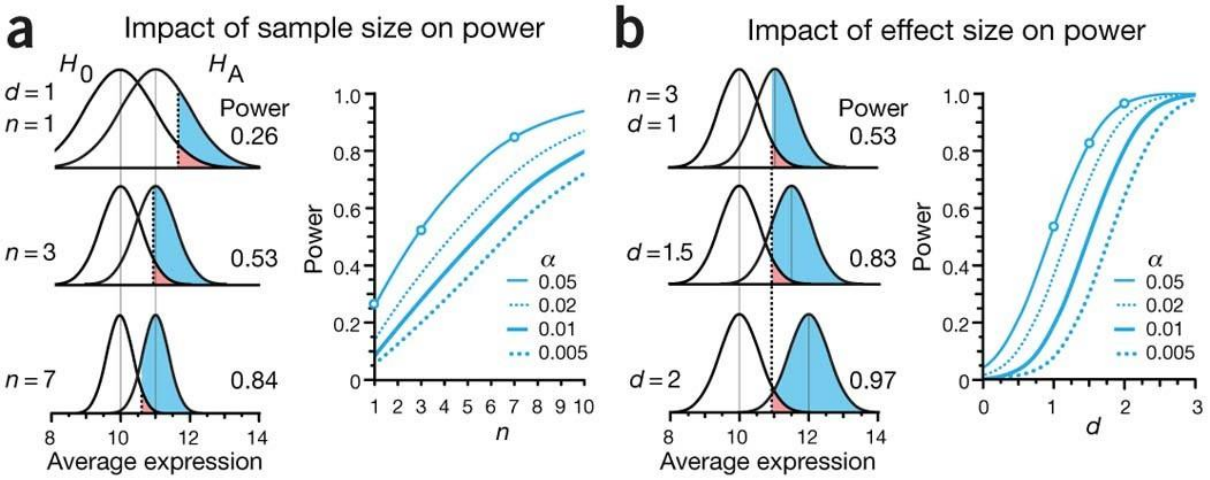


Figure 7: Statistical Power – Sample-size vs power relation
"H_0 and H_A are assumed normal with σ = 1. (a) Increasing n decreases the spread of the distribution of sample averages in proportion to $\frac{1}{\sqrt{n}}$. Shown are scenarios at n = 1, 3 and 7 for d = 1 and α = 0.05. Right, power as a function of n at four different a values for d = 1. The circles correspond to the three scenarios. (b) Power increases with d, making it easier to detect larger effects. The distributions show effect sizes d = 1, 1.5 and 2 for n = 3 and α = 0.05. Right, power as a function of d at four different a values for n=3." Figures and legend from (Krzywinski & Altman, 2013) with the permission of Springer Nature.

The issue of power is fundamental in hypothesis testing and has profound implications for the search for statistical epistasis. Indeed, the power of a study informs about what would be the chances of detecting an effect of a certain strength in the study dataset if that effect did actually exist. If the effect size or sample size are too low, statistical hypothesis testing could lead to the conclusion that an effect is not present although it actually is. Knowledge about the power of a study is needed to be able to interpret properly its results. For example, if an epistasis study had a 95% to detect even a very weak effect and that the GWIAs yielded no significant results, one could conclude with high confidence that epistatic interactions do not play a role in this trait. On the contrary, when a GWIAs has 60% power to detect a strong effect, it is erroneous to use failure to find any effect as an argument against epistasis in this trait. Indeed, interactions with a small effect size had nearly zero chance to be detected, if they were present. In addition, different modes of genetic inheritance, such as recessive inheritance, which have been most of the time ignored to the benefit of additive models, might substantially lower the power to detect interactions and should be taken in account (see Section 2.3.4)

Power calculations in GWIAs are indispensable because the very high number of tests makes them very likely to be underpowered. Indeed, even the biggest publicly available genetic data sets, such as the UKBB, can only reliably detect moderate to strong effects between SNPs with moderate to high MAFs (see Section 2.3.2). Unfortunately, whereas a wide variety of tools are available to compute power in GWAS, specific tools to compute the power to detect genetic interaction effects are, to my best knowledge, missing. In Section 2, we present an R package, epilogitpower, aiming at providing fast and well-documented power computation for epistasis studies.

1.2.2.6 Haplotype effect and rare-variant tagging:

As introduced earlier, association studies rely on the principle that not all genetic variants need to be tested for association, but rather a certain number of "tag-SNPs" are used to assess the association of all variants in high LD with them. The causal variant can then be any variant in LD with the tag variant.

Along the same lines, a particular combination of tag-SNPs present on the same chromosome between which an interaction has been detected (cis-epistasis) can actually identify a particular haplotype that is driving the association signal. Whereas it has been shown that linkage between variants involved in interactions can be favored by evolutionary mechanisms (Lappalainen, Montgomery, Nica, & Dermitzakis, 2011), and therefore represent genuine interaction signal of interest, it is also possible that the haplotype contains a single causal variant associated with the disease.

As an example, in 2014 Hemani and colleagues reported having identified "501 significant pairwise interactions between common SNPs influencing the expression of 238 genes" (Hemani et al., 2014). After the publication of this article, Wood and colleagues could actually show that most of the interaction signals could actually be explained by a single third variant present in sequencing data but absent from the first analysis (Wood et al., 2014).

This highlights the need to control for possible confounding by haplotype effect in interaction studies. One particular way to proceed is to assess the significance of the interaction term when conditioning on any other SNP in LD with the tag-SNPs. This strategy is able to identify additional variants tagged by the interactions or might lead to the discovery of different or higher order interactions (as exemplified in Section 5 and in (Zeng et al., 2022)). However, this method can only take into account variants for which data is available; that is variants which were genotyped or imputed and passed the quality control steps (QC). Even if the interaction passes this conditional testing strategy with success, one cannot exclude that a rarer variant for example, which would not be part of any imputation panel, or whose very low MAF lead to exclusion during QC, be tagged by the association. Such an association signal tagging a rare variant could also be replicated with success in an independent dataset if the same haplotype structure is provided there as well.

In conclusion, even when controlling for it, cis-epistasis pairs could always be due to the haplotype tagging and not to a genuine interaction effect. Functional validation of the interaction partners and elucidation of the interaction mechanism at the molecular level can however bring more credibility to these interactions.

2 Epilogitpower: Power computation for epistasis studies under different genetic models using a logistic regression framework

2.1 Introduction

Low-power studies which were not able to detect true epistatic pairs or produce high rates of false positives, which then failed replication, led to a high degree of skepticism towards epistasis in human diseases. Accurate, off-the-shelf, and easy-to-use power calculation methods, directly available for researchers are therefore needed to ensure adequate power in epistasis studies and provide data-driven insights into the role of epistasis in human genetics.

In this chapter, we present the R package epilogitpower, which aims at providing easy power calculation for epistasis studies based on logistic regression. Although a wide range of statistical models have been developed for epistasis studies (see (Niel, Sinoquet, Dina, & Rocheleau, 2015) for a review), we decided to focus on logistic regression for multiple reasons: Firstly, logistic models have been used for decades in GWAS and are widely accepted and understood in the statistical genetics community. Secondly, we could take advantage of well-studied statistical theories of generalized linear models to efficiently compute the power. Despite logistic models being very popular, there is to our knowledge no available tool to compute the power of interaction terms in these models, let alone any tool providing genetics-oriented documentation for these computations. The major focus of epilogitpower is therefore to offer user-friendly power computation with different possible sets of user input that are well-defined and understood in the genetic community. Our package also allows to specify different genetic models, namely additive, dominant, recessive, and heterozygous for each SNP. Indeed, previous research has shown that different modes of heritability have important consequences on power the achieved in GWAS (C. M. Moore, Jacobson, & Fingerlin, 2020). Finally, we provide functions to help predict the plausible range of effect sizes for a given epistasis pair given the marginal effect of the individual SNPs and epidemiological data. This information can be used both to help power computation or to prioritize putative epistasis pairs.

2.2 Materials and Methods

2.2.1 Case-control-status dependent genotype frequencies

The different genotypes frequencies in the cases, P(G|D+), and the controls group , P(G|D–), can be computed using Bayes' rule in the following manner:

1) From Bayes' rule:

$\Pr(\mathrm{G}|\mathrm{D}) = \frac{\Pr(\mathrm{D}|\mathrm{G}) \cdot \Pr(\mathrm{G})}{\Pr(\mathrm{D})}$, where Pr(D) is the prevalence of the disease and Pr(G) the frequency of the genotype in the whole population which can be computed from the SNPs MAF under the HWE assumption. We, therefore, need to compute the penetrances of the different genotypes; Pr(D|G) to find the stratified genotypes frequencies Pr(G|D).

2) The genotype penetrances can be easily estimated if one knows the genotypic risk ratios, GRRs. As per the definition of GRRs, $GRR = \frac{\Pr(\mathrm{D}|\mathrm{G})}{\Pr(\mathrm{D}|G_{ref})}$, where $\Pr(\mathrm{D}|G_{ref})$ is the penetrance of the reference genotype which can be estimated in the following manner:

$\Pr(\mathrm{D}) = \Pr(\mathrm{D}|G_{ref}) \cdot \Pr(G_{ref}) + \ldots + \Pr(\mathrm{D}|G_{aabb}) \cdot \Pr(G_{aabb})$ =>

$\frac{\Pr(\mathrm{D})}{\Pr\,(D|G_{ref})} = \Pr\left(G_{ref}\right) \cdot \frac{\Pr(\mathrm{D}|G_{aabb}) \cdot \Pr(G_{aabb})}{\Pr\,(D|G_{ref})}$, recognizing that $\frac{\Pr(\mathrm{D}|G_{aabb})}{\Pr(\mathrm{D}|G_{ref})} = GRR_{aabb}$,

we obtain $\frac{\Pr(\mathrm{D})}{\Pr\,(D|G_{ref})} = \Pr\left(G_{ref}\right) + \cdots + GRR_{aabb} \cdot Pr(G_{aabb})$ =>

$$\Pr\,(D|G_{ref}) = \frac{\mathrm{P(D)}}{\Pr\left(G_{ref}\right) + \cdots + GRR_{aabb} \cdot \Pr(G_{aabb})}$$

2.2) The genotypic risk ratios are not quantities that are typically known or easy to estimate for any given disease without individual-level genotype data. Odds ratios (OR), on the contrary, are available from the summary data of any GWAS. Under the assumption of a given reference genotype penetrance, P_0 , one can compute the GRR from the ORs using the following relation:

$$GRR = \frac{OR}{(1 - P_0) + P_0 \cdot OR}$$

We therefore also provide a function to simulate data and compute power using the ORs and the reference genotype penetrance as input. In this case, the prevalence of the disease is then computed using the other variables with

$$\Pr(\mathrm{D}) = \Pr\left(\mathrm{D}\middle|G_{ref}\right) \cdot \Pr(G_{ref}) + \ldots + \Pr(\mathrm{D}|G_{aabb}) \cdot \Pr(G_{aabb})$$

.

2.2.2 Monte Carlo Simulation

When the frequencies of the different genotypes for the cases and control groups under the alternative hypothesis are known, simulated datasets can be obtained by sampling the desired number of cases and controls from the stratified genotypes probabilities. This simulated dataset can then be used to fit a logistic regression model. The sampling and subsequent logistic model fitting are repeated a given number of times to generate the distribution of the p-values under the alternative hypothesis. The power is computed as the fraction of these p-values smaller than the α threshold.

2.2.3 Power computation using distributions of the coefficients under the null and alternative hypothesis

Given a logistic regression model of the form $P(y) = \frac{e^{\theta_0+\theta X}}{1+e^{\theta_0+\theta X}}$ the Wald-test Statistic for each of the coefficients , computed as, $\frac{(\hat{\theta}_i-\theta_{0i})^2}{Var(\hat{\theta}_i)}$ follows the following distributions:

- Under the null hypothesis H_0 that $\theta_i = 0$: a central χ^2 distribution with 1 degree of freedom
- Under the alternative hypothesis H_1 that $\theta_i \neq 0$: a non-central χ^2 distribution with 1 degree of freedom and non-centrality parameter γ equal to the expectation of the Wald test under the alternative hypothesis (Gudicha, Schmittmann, & Vermunt, 2017; Wald, 1942).

Knowing these two distributions, the power can be computed as: $1-\beta = \Pr(\chi^2(1,\gamma) > \chi^2_\alpha(1)$ where α is the significance threshold.

2.3 Results

2.3.1 Fast power computation

This package makes use of data simulations to perform the power computations. The easiest way to compute power is indeed to perform Monte Carlo simulations by repeatedly simulating a dataset with the desired properties (see Methods), fitting the logistic model of which one wants to assess the power and compute the fraction of time in which the diverse coefficients of the model are significant at the desired threshold. This naïve approach can, however, be very slow, especially for large sample sizes which are typically needed in GWIAs. To circumvent this caveat, we provide an alternative power computation method (referred to as the χ^2method hereafter) which relies on the statistical distribution of the test statistic for the interaction term. These statistical properties allow us to compute the power in a very cheap way, as we only need to fit a single logistic model using the simulated data to find the expected value of the Wald Test statistic for each of the coefficients under the alternative hypothesis.

These two methods are asymptotically equivalent and they do yield almost identical results with the large sample sizes that are typically of interest for GWIAS. As expected, we observed that the difference between the two computations method decreases both when increasing the number of simulations and the number of samples (Figure 8A). Indeed, it is expected that an increased number of simulations leads to a better precision of the MC simulations, whereas the χ^2 method becomes more accurate with larger sample sizes because the statistical properties used are asymptotically true. But even with a hundred simulations in a small sample size of 5000 cases and 5000 controls, the difference between the two methods is negligible (Figure 8 left). On the contrary, the computation time difference for different sample sizes and numbers of simulations are big, and growing exponentially with the number of simulations (Figure 8 right).

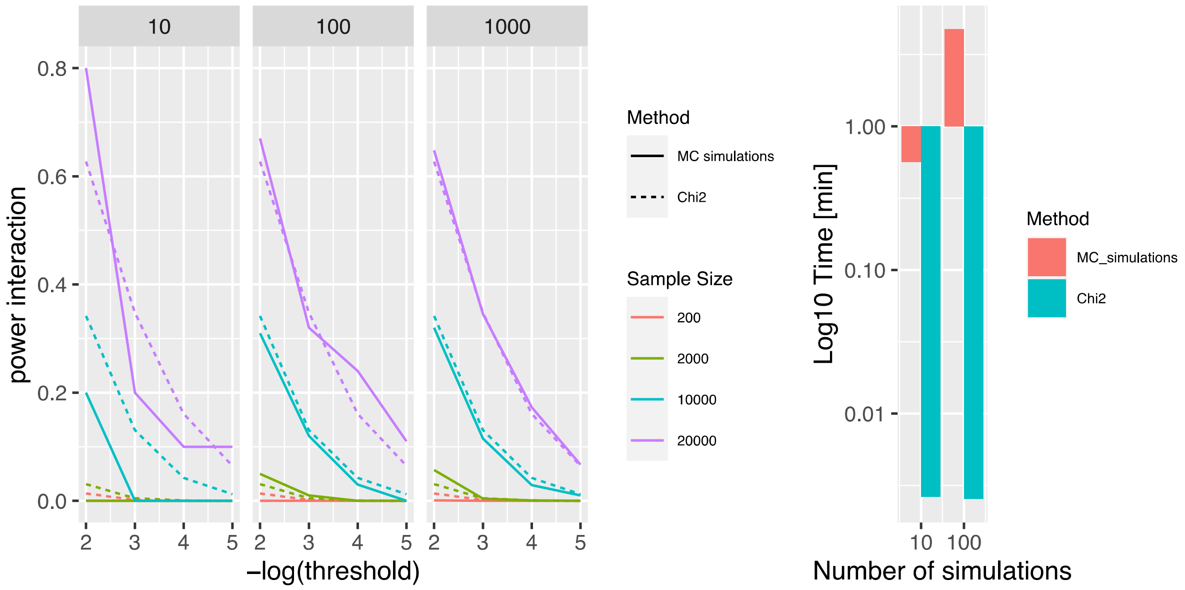


Figure 8: Fast power calculation
Left: Power for the interaction term in the logistic model in function of the αthreshold on the log$_{10}$ scale and the sample size as computed using respectively 10 (left), 100 (middle), or 1000 (right) simulations. The continuous line represents the power computed using Monte Carlo simulations and the dotted line using the χ^2 method. The power was computed assuming a 1:1 case-control ratio, MAFs of 0.1 for both SNPs, marginal ORs of respectively 1.1 and 0.9 for SNP1 and SNP2, an interaction OR of 1.25, and a penetrance of 0.1 for the reference genotype.
Right: Computation time in minutes, displayed on the log10 scale, in function of the number of simulations for a sample size of 10’000 using the Monte Carlo simulations (in red) or the χ^2 method (in turquoise). The power was computed assuming a 1:1 case-control ratio.

2.3.2 Sample size calculation

Sample size calculations should be performed before starting any genetic studies. It allows to compute the minimal sample size to detect an effect of a certain strength with a certain probability if the effect truly exists. For this example, we consider conducting a GWIA on some disease. We want to estimate the number of cases and controls that one needs to measure in order to detect an interaction with a certain effect size. In order to use the power computation function of the epilogitpower package we need to specify the following parameters which impact the power :

- Minor allele frequency (MAF) of the two interacting SNPs: a lower MAF reduces the power. For this example, we set both MAFs to 0.1, a usual threshold in GWAS.
- The additive Odds ratio (OR) of the SNPs forming the putative epistatic pairs. In this example, we hypothesize that epistasis happens between SNPs that do not have a marginal effect and set their OR to 1.

- The OR of the interaction. This is a measure of the effect size of the interaction. Interactions with larger OR are easier to detect. In this case, we estimate the interaction OR to be 1.15, a value in the range of the additive ORs commonly observed in GWAS.
- The penetrance of the reference genotype. This quantity can be easily approximated if one has access to the raw genetic data. If not, in the case of a complex disorder, with small effect sizes for individual loci and relatively rare SNPs, the prevalence of the disease in the population might be a good estimate of the penetrance of the reference genotype (as this genotype is over-represented and therefore weights more in the general population prevalence). In this example, we will assume the penetrance of the reference genotype to be 0.1
- The α threshold. This is the p-value threshold corrected for multiple testing. The most common correction procedure is the Bonferroni correction, computed as the desired α threshold divided by the number of tests. For this study, we will assume 2 million LD independent variants leading to 2×10^{12} tests.

Figure 9 shows the power to detect the interaction in function of the number of cases for different numbers of controls. We observe that a power of 80% can be achieved with about 125’000 cases and between 500’000 to 1 million controls. Alternatively, a similar power can be achieved with 275’000 cases and 200’000 controls, in the case where control recruitment is as difficult as cases recruitment. In comparison, the sample size needed to achieve 80% power in a GWAS with the same settings (MAF 0.1, OR=1,15, α threshold=1×10^{-8}, additive genetic model) is ca 80’000 with a case-control ratio of 1/5 or ca 40’000 with a case-control ratio of 1 (as computed with the genpwr R package (C. M. Moore et al., 2020)).

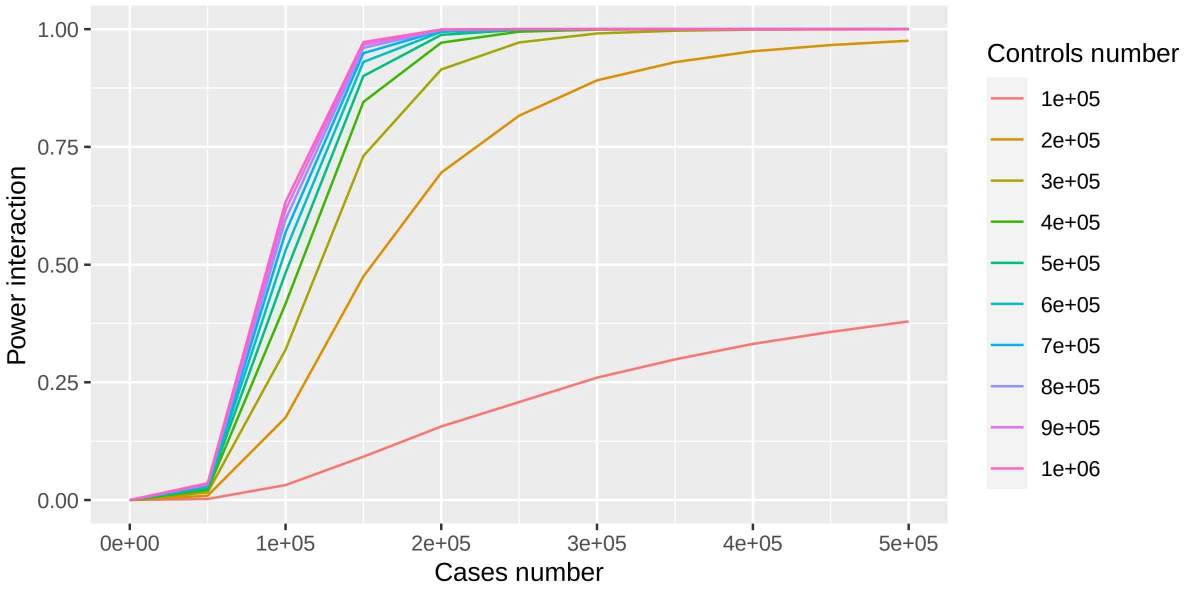


Figure 9: Sample-size computation
Power for the interaction term in the logistic model in function of the number of cases for different numbers of controls. The power was computed assuming a Bonferroni correction for pairwise testing of 2 million SNPs, MAFs of 0.1 for both SNPs, marginal ORs of 1 (no effect) for both SNPs, an interaction OR of 1.15, and a penetrance of 0.1 for the reference genotype.

The parameters used in this example are typically the parameters that are available in the literature from previous studies and are well understood among researchers. Epilogitpower also accepts a different set of input for the cases where they would be more readily available. In particular, the user can specify Genotypes Relative Risks (GRRs) instead of ORs, along with the general disease prevalence instead of the penetrance of the reference genotype (see Methods for details). The code to perform such analysis using the epilogitpower package is entirely provided in the vignette of the package.

2.3.3 Post-hoc power computation

As a counterpart of computing the power of an epistasis study before performing it, one can also assess the power of a study after its completion to help interpreting the results. For example, if we find in the literature the study from the previous paragraph was performed with 20’000 cases and 20’000 controls, we can compute a-posteriori the power to detect epistatic pairs with different effect sizes for a certain MAF. Given that the study did not find any significant pairs, Figure 10 would help us conclude that there might indeed truly not be any interaction with an OR of 1.6 or higher and a MAF of 0.3 or higher, as the power to detect it was 80%. On the other hand, one should be very careful to conclude the absence of epistasis with an OR smaller than 1.5, as the power to detect it was lower than 50%.

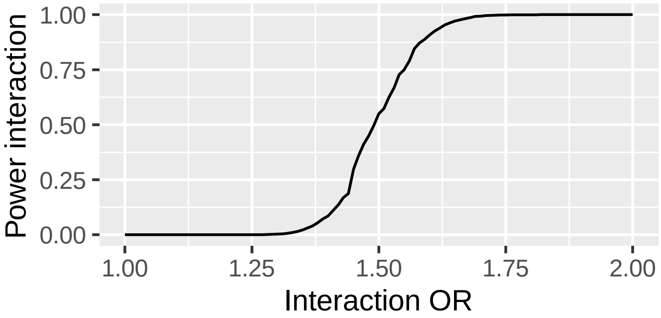


Figure 10: A-posteriori power computation
Power for the interaction term in the logistic model in function of the interaction OR in a sample of 20’000 cases and 20’000 controls. The power was computed assuming a Bonferroni correction for pairwise testing of 2 million SNPs, MAFs of 0.1 for both SNPs, marginal ORs of 1 (no effect) for both SNPs, an interaction OR of 1.15, and a penetrance of 0.1 for the reference genotype.

2.3.4 Effects of genetic models on power computation

All regression-based genetic association studies require the encoding of the genetic risk carried at a single locus into a unique number in a process called genetic encoding. At any locus, given reference allele, A, and alternate or risk allele a, all encodings assume that the AA (homozygous referent) genotype incurs no risk increase or decrease (i.e. sets the baseline) and the aa (homozygous alternate) genotype bears full risk. Depending on the underlying genetics, the heterozygous can present a risk anywhere between 0% and 100%. 0% corresponds to the well-known recessive model of inheritance, where two copies of the risk allele are needed for the phenotype to appear. On the other hand, 100% corresponds to the dominant model of inheritance where one copy of the risk allele is sufficient for the phenotype. Another commonly used model, the additive or dosage model, supposes that the risk increases linearly with the number of risk alleles, yielding a 50% risk for the heterozygous. A fourth model, less commonly used but also possible, is the heterozygous model where the risk is null for both homozygous while being 100% for the heterozygous.

A priori, it is impossible to know which of these 4 models is true for any given SNP in the genome. This is why GWAS were originally performed using all encodings. Over the years the tendency shifted to using only the additive encoding. It is beyond the scope of this package to discuss the advantages or disadvantages of this trend. However, for a given MAF in a sample of individuals, the number of minor allele homozygous (bearing full genetic risk in the recessive model) is lower than the number of major allele homozygous (bearing full risk in the dominant model) for example. The power to detect interaction, or a marginal effect, for a pair of recessively acting SNPs is smaller than for a pair of dominant SNPs with the same interaction

OR and sample size. Until now, all examples in this chapter used the default “dosage-dosage” encoding, for a dosage encoding of both SNPs 1 and 2. However, all functions of this package have an “encoding” argument to specify which encoding to use for the two SNPs of the pair, if it was different from the default one. This argument should be specified accordingly to the information available on the two SNPs.

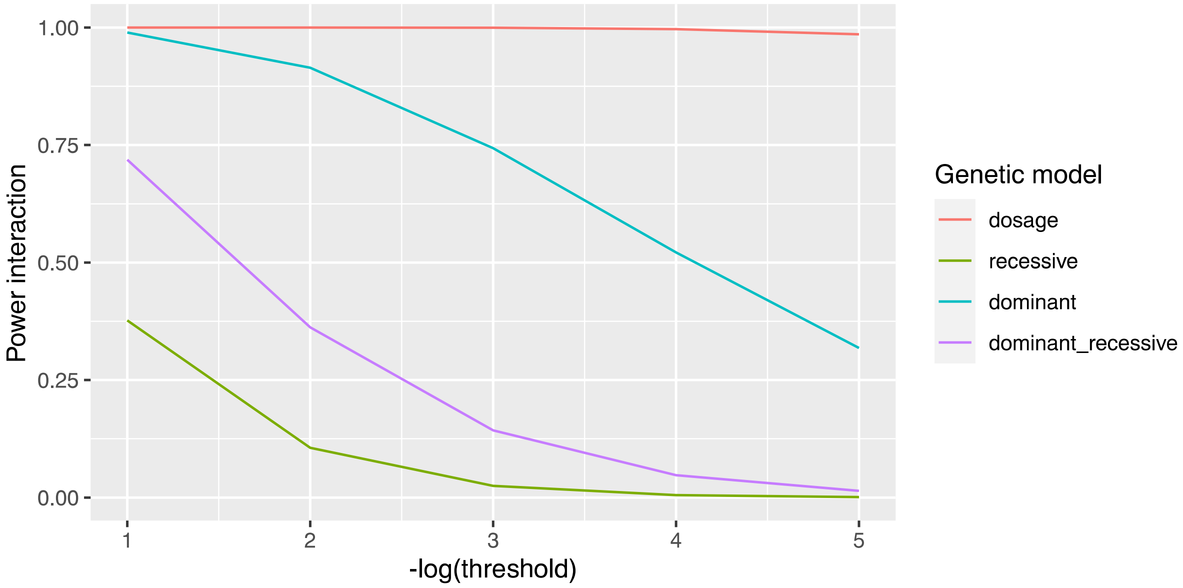


Figure 11: Effects of genetic models on power computation
Power for the interaction term in the logistic model in function of the alpha threshold for different genetic model encoding. The power was computed assuming a sample size of 10’000 cases and 10’000 controls, MAFs of 0.3 for both SNPs, marginal ORs of 1 (no effect) for both SNPs, an interaction OR of 1.25, and a penetrance of 0.3 for the reference genotype.

Figure 11 shows the substantial power differences resulting from different genetic encodings. As expected, when both SNPs are following a recessive model the power drops greatly compared to other models. Taking into account the possible genetic models of both SNPs is therefore of extreme importance when planning a GWIAS or interpreting results. It is also important to pay attention to the genetic encoding used in the studies providing information that is used for the power calculation, for example, the GWAS summary statistic for the marginal ORs of the two SNPs. The reported ORs are associated with the particular genetic encoding used in this study. Using the same OR but with a different genetic model encoding in the power calculation will result in erroneous power calculation as in Figure 11.

2.3.5 Predicting plausible interaction effect sizes

Using the set of equations developed to compute the genotypes frequencies of cases and controls (see Methods), along with information about the two SNPs and their marginal additive effect, we can compute a range of possible effect sizes for their interaction. More precisely, we can find out which range of interaction effect sizes would be compatible with the observed disease prevalence and marginal effects if we know the MAF and additive ORs of the two SNPs, the penetrance of the reference genotype, and with the strong assumption of a monogenic model. In Figure 12, we hypothesized the following parameters; MAF of both SNPs=0.2, significant marginal effect in the GWAS with respective ORs of 1.2 and 1.5, and penetrance of the reference genotype of 0.2 for a fairly common disease. Figure 12 shows the disease prevalence in function of the putative interaction OR. If we, for example, had the supplementary information that the general prevalence of the disease is between 24% and 26%, this would only be compatible with an interaction OR between ca 0.9 and 1.5 Additionally, any ORs between the lower and upper default bound of 0.5 and 2 are compatible with a prevalence between the lower and upper default bound of 0 and 1.

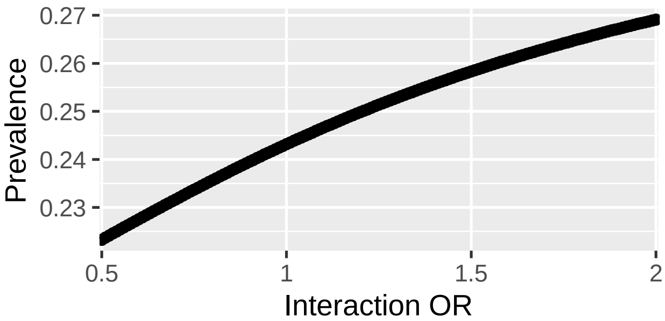


Figure 12: Plausible interaction ORs
Prevalence of the disease predicted in function of the OR of the interaction. The computation was performed assuming the following parameters: MAF of both SNPs=0.2, significant marginal effect in the GWAS with respective ORs of 1.2 and 1.5 for SNP1 and SNP2, and penetrance of the reference genotype=0.2.

Computing the possible range of effect size of interaction in this manner could have two usages. First, it can be used to get an approximation of the interaction effect size to compute the power to detect an epistatic interaction between the two SNPs. Secondly, we can imagine using this possible effect size computation as a filter in Genome-Wide Interaction Association Studies. One could use the data available from a previous GWAS about the MAF and marginal effect size of the SNPs along with the observed disease prevalence to compute the possible effect size of each interaction. We would then have a ranking of the SNPs whose interactions have the potential to have meaningful effect size, and threshold it to generate a list of interactions to test.

2.4 Discussion

We provide and describe a new R package that provides straight-forward power computation for epistatic interactions in a GWIAS setting. This package can be used both to compute the sample size needed for a new study and to compute the power of a published a-posteriori.

Lack of power is a critical problem in epistasis studies (Ritchie & Van Steen, 2018). Apart from the absence, up to now, of power-computation tools for interaction terms in logistic regression in the genetic context, this is certainly due to the very high sample sizes needed for epistasis studies. Our results show indeed that much bigger sample sizes are needed for epistasis studies as compared to univariate GWAS in the same settings. Whereas this should not come as a surprise, the extent of this difference was maybe underestimated. It is important to note that this difference will only get bigger when we start investigating interactions between rarer variants, which will increasingly become the case with the advances in sequencing techniques (Bomba, Walter, & Soranzo, 2017). In this context, we hope that the epilogitpower will help researchers to set their epistasis studies within the range of desired statistical power.

The main functions of the package can take two different sets of commonly available genetic information, in an attempt to make power computation as easy as possible for researchers and foster its use in future studies design. Depending on the availability of prior knowledge, users can use either GRRs or ORs, two quantities that are often mixed up but are only close to equal for rare diseases (prevalence < 10%) (J. Zhang & Yu, 1998). Genotypic Risk Ratios (GRR) can be used in combination with the prevalence of the disease to compute power. Whereas the prevalence is typically known for most diseases, GRRs might not be. This is because GRRs cannot be computed in case-control studies (Ranganathan, Aggarwal, & Pramesh, 2015) such as the GWAS which generated most of the knowledge on genetic association in the last decade. On the other hand, Odds Ratios (OR) are readily available from any GWAS summary statistics. Unfortunately, the penetrance of the disease for the reference genotype (J. Zhang & Yu, 1998) is then needed to compute the power using ORs, and this quantity might not be as easily available as the prevalence.

Previous studies have already shown that genetic model miss-specification could greatly wrong the power calculation in the case of GWAS (C. M. Moore et al., 2020). We show in the present chapter that this is also the case for epistasis. Epistasis studies are even more impacted by the miss-specification of the genetic model, as it can happen on both of the SNPs in the pair. Our package, therefore, allows to specify the genetic model for each of the SNPs independently. As it is not possible to know a priori which genetic model is true for any given SNP when performing a GWIAS, one solution is to calculate an upper and lower bound of the power using the different possible genetic models.

Epilogitpower also provides a function to compute the range of plausible interaction effect sizes given the marginal effect size of the two SNPs, the penetrance of the reference genotype, and the prevalence of the disease. A major limitation of this function is however that it needs a very precise estimate of the prevalence and reference genotype penetrance to give meaningful information on the plausible effect size of the interaction. Indeed, in the example depicted in Figure 12, a prevalence of 23% indicates an OR of approximately 0.65, and therefore a protective effect of the minor alleles interaction whereas a prevalence of 27% indicates an OR of 2 and therefore a detrimental role of the minor allele interactions. We observe here that two prevalences which are possibly closer than the error margin of prevalence estimation lead to opposite conclusions on the effect of the interaction. We, therefore, emphasize that large sample size and precise estimation of the prevalence and penetrance of the reference genotype are needed for this estimation. The univariate model used for this computation is another strong limitation. Indeed, the range of plausible interaction ORs is calculated with the assumption that the prevalence of the disease is influenced solely by the marginal effect of the two SNPs and their interaction.

In conclusion, we provide the first R package to perform power computation for genetic interaction in the logistic regression framework. This package along with the demonstration of its usage presented in the companion vignette should facilitate power computation in epistasis studies and thereby increase the reproducibility of future studies and the trust of the community in epistasis research.

3 Genome-Wide Interaction Association Study for CAD

3.1 Introduction

For the analysis presented in this chapter, we aimed at conducting a well-powered and robustly designed hypothesis-free and genome-wide interaction association study. The debate about the involvement or not of epistatic interaction in human complex traits has predominantly involved theoretical and mathematical arguments, but very few actual data from interaction association analysis. We believe that experimental data from multiple genome-wide interaction study could help bringing practical and maybe decisive evidence to the debate.

Our goal here was therefore first to demonstrate that such a study using one of the largest genetic datasets available for multiple diseases, namely the UKBB, is computationally feasible without having to reduce the search space with an a-priori filter. Indeed, a hypothesis-free method has several advantages over the filter-based one. Firstly, it can allow discovering interactions between genes that were not previously associated with the disease and therefore generate insights into new biological mechanisms underlying the disease. Secondly, the presence of patterns in that undirected search might be able to reveal some basic principles of epistasis in common traits, which in turn could help to target knowledge-based analysis more efficiently.

For this study, we decided to concentrate on Coronary Artery Disease (CAD) because we both had high-quality and large discovery (UKBB) and replication datasets (10 CAD studies, see Section 5.2) at hand. Moreover, CAD presents the crucial advantage that its etiology and genetic basis are relatively well understood for a complex disorder, in comparison to psychiatric disorders for example. And we believe, that being able to relate the putative epistasis findings of this study to a well-understood genetic basis is key for the credibilisation and acceptance of the results by a genetic community that is generally skeptical about epistasis studies.

3.2 Methods

3.2.1 QC and imputation

3.2.1.1 UKBB dataset

The UKBB imputed dataset containing 487'410 samples was downloaded from the UKBB portal. Around 50'000 of these samples were genotyped on the Applied Biosystems UK BiLEVE Axiom Array by Affymetrix while the rest was genotyped on the UK Biobank Axiom Array. Both of these arrays did assess genotype at roughly 800'000 markers and shared 95% of their variants (Bycroft et al., 2018). According to the UKBB manuals, the pre-imputation QC was performed as described in (Bycroft et al., 2018). For this study, we considered only variants located on the 22 autosomal chromosomes.

The genetic dataset was imputed separately using the HRC reference panel on one hand and a merged reference panel of the UK10K and 1000 Genomes phase 3 panels on the other hand. These imputations were then integrated together, with priority on the HRC panel for SNPs present in both imputations. Imputation was performed using Impute4 and the final dataset comprises 93,095,623 autosomal SNPs, short indels, and large structural variants in 487,442 individuals (some of them had to be excluded in our analysis due to withdrawn consent, leading to a final number of 487410 samples in our analysis).

After downloading this dataset from the UKBB server, we applied the following post-imputation QC. Firstly, we kept individuals fulfilling the HARD criteria of CAD (see (Nelson et al., 2017) as cases and individual not fulfilling neither the SOFT nor the HARD CAD criteria as controls. Secondly, we removed related individuals up to the third degree. Starting with a list of pairs of individuals related up to the third degree, we first removed the control individuals in pairs containing both a control and a case individual. For the pairs containing two cases, we built a graph of the related people and removed as few individuals as possible to brake all edges of the graph. Furthermore, we removed all individuals with non-white European Ancestry (as provided in the UKBB phenotypes). While this reduces the applicability of the findings to a narrow population, some important risk factors for CAD , for example, LPA, are known to have different genetic risk factors in different ethnies (Enkhmaa, Anuurad, & Berglund, 2016). As this scan is more of a proof of concept for epistasis genome-wide scans than a direct search for clinically useful variants, it makes sense to reduce the variability at the cost of a narrower population target. We did not remove samples based on extreme Multi-Dimensional Scaling

(MDS) Components of the genetic variance like it is usually done in GWAS, as we observed a very homogenous MDS distribution with no outliers.

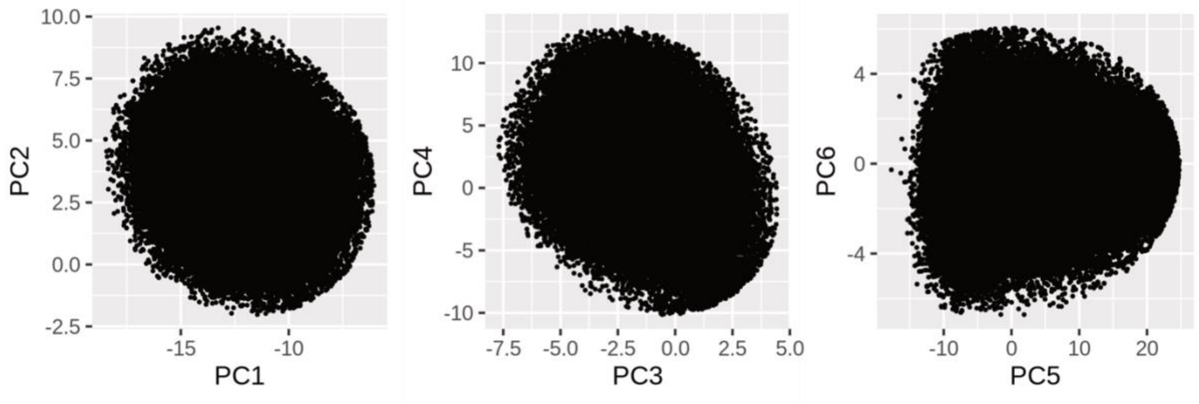


Figure 13: Pairwise scatter plot of the first MDS components of the genetic variance in the QCed UKBB dataset
The first MDS components of the genetic variance, as computed by the UKBB team, indicate if there is a strong population structure in the dataset. In this case, the plots are very homogenous and do not indicate such a population structure.

The final analysis sample was made of 331’956 samples. Among them 314’728 controls and 17’228 cases. With 178’731 women and 153’225 males, the two genders are almost equally represented. The proportion of males and females in controls is roughly the same, however, there are almost four times more men in the cases group. We also observe that cases are older than controls. This indicates that age and sex might be confounders in the genetic associations analysis and will need to be corrected for.

sex	pheno	N	mean age	sd age
Female	case	3,757	62.03939	5.93796
Female	control	174,974	56.40647	7.896362
Male	case	13,471	61.6491	6.041801
Male	control	139,754	56.50197	8.132102

Table 1: Demographics of the UKBB dataset used as discovery sample

We then filtered the imputed variants according to the following criteria:

- MAF: 0.05 or 0.25 for the recessive models (see Section 3.3.1)
- Imputation Score as reported by IMPUTE4 higher than 80 %
- Hardy-Weinberg Equilibrium (HWE) test p-value < 1×10^{-6}
- Missingness of the SNP across all samples < 0.2

After post-imputation QC the final genetic dataset was made of 6'722'728 variants with the MAF threshold of 0.05 and 294'674 variants at the MAF threshold of 0.25.

3.2.2 Analysis Strategy

In the light of the power analysis and the runtime predictions, it appeared feasible to run the genome-wide hypothesis-free interaction scan using a two-step strategy, a MAF threshold of 0.05 for all genetic models except the recessive one, and an r^2 LD pruning threshold of 0.5. The overall strategy is described in Figure 14 and goes as follows. I will first explain the strategy for the dominant, dosage, and heterozygous genetic encoding and then explain the differences for the recessive models.

We start with the QCed dataset of 6'722'728 SNPs with a MAF threshold > 0.05. The SNPs were pruned with an r^2 LD threshold of 0.5 and a sliding window of 10'000kb using Plink (Purcell et al., 2007). The resulting dataset consisted of 537'930 LD-independent SNPs. The genotype at each of these SNPs was then encoded in the four genetic models, dominant, additive (minor allele dosage), heterozygous, and recessive (see Section 2.3.2).

3.2.2.1 Step 1: loose statistical filtering

All pairwise combinations between all three genetic encodings of these SNPs were then tested using the CPU version of the episcan R package (v0.01, Jiang & Pütz (2018) , https://CRAN.R-project.org/package=episcan). The goal of this first step was to provide a fast statistical screening for interesting pairs. We, therefore, decided to use a loose significance threshold of 1×10^{-10}, instead of the suggestive threshold of 6.9×10^{-12} which would be computed as 1 divided by the number of LD-independent SNP pair combinations to account for the fact that causal variants might not be in full LD with any of the pruned SNPs. With this we hoped to capture these possibly pruned causal variants while still generating a number of putative epistatic pairs that can be handled in the second, more stringent, testing step.

For all pairs which successfully passed this first screening step, we performed LD expansion to invert the LD pruning step performed before the episcan. In other words, we search for any SNPs in LD > 0.5 for each SNP of each of the selected pairs, which we will call tagged SNPs. For each of the selected pairs, we then list all combinations of their tagged SNPs and forward them to the second, more rigorous statistical testing step, allowing the fine mapping of the causal signal inside the LD blocks.

For example, the pair rs1:rs2 was selected in step 1. rs3 and rs4 are in LD with rs1 and rs5 in LD with rs2, the following pairs would be tested in step 2.

rs1:rs2

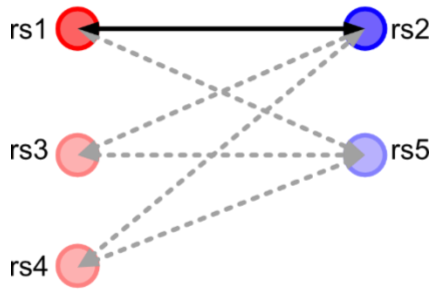
rs1:rs5

rs3:rs2

rs3:rs5

rs4:rs2

rs4:rs5



With this pruning-testing-expanding strategy, we hope to effectively test all SNPs in the genome, while keeping a manageable number of tests.

3.2.2.2 Step 2: Fine-mapping and validation of interaction using logistic regression

For each of the SNP pairs forwarded to this second step and for each of the genotype encoding combinations, we ran the following logistic regression model with the case status as the outcome, the SNPs in the respective genetic encoding as predictors, and sex, age as well as the 10 first MDS components of the genetic variance as covariates to account for population stratification:

Case/control status ~ $SNP_1 + SNP_2 + SNP_1{:}SNP_2 + sex + age + MDS_1 + MDS_2 + \ldots + MDS_{10}$

At this step, we applied the suggestive multiple testing correction thresholds of 6.9×10^{-12} (1/ number of LD independent SNP pair combinations). We could have applied a more stringent correction at this step. However, the threshold of 6.9×10^{-12} did not yield more than 100 pairs and the computational cost of replicating them was very low as compared to the one of the previous steps. We, therefore, decided to use this looser threshold to generate more candidates for replication, at the cost of the control of the type 1 error. But we argue that replication in an independent dataset provides in this case higher evidence for the epistatic pairs than stronger control of the type 1 error. For each of the original SNPs pairs from step 1, we selected the pairs with the lowest p-value in step 2 among the pairs that their LD expansion generated and forwarded it to the replication step. If multiple genetic encoding combinations for this pair had passed the significance threshold, we replicated using the most significant one.

3.2.2.3 Replication

We replicated the significant pairs from step 2 using a fixed-effect inverse variance weighted meta-analysis of 10 independent CAD studies amounting to a total of 29’755 individuals. The QC and imputation of these 10 studies are discussed in more detail in (Zeng et al., 2022). These 10 studies were however imputed on a different, smaller, imputation panel and therefore not all SNPs from our analysis were available for replication. The first step in the replication process was therefore to find which of the significant SNPs were absent from the replication dataset and to find their tag-SNPs, i.e., SNPs present in the replication and in LD with them. For each SNP to replicate which was absent in the replication we looked for all tag-SNP in LD > 0.5 within a 10000kb window in the UKBB dataset using Plink. We then crossed this list with the list of SNPs present in the replication and chose the one with the higher LD as final tag. However, for ca. 40% of the SNPs-pairs, we could not find any tag present in the replication, and could not replicate these pairs.

For the replicable pairs, we used a fixed-effect inverse variance weighted meta-analysis using the metagen function from the meta package (v4.9-9, Balduzzi, Rücker, Schwarzer (2015),https://CRAN.R-project.org/package=meta). We used the following logistic model in the meta-analysis:

Case/control status $\sim SNP_1 + SNP_2 + SNP_1{:}SNP_2 + sex + MDS_1 + MDS_2 + \ldots + MDS_{10}$

We considered an interaction replicated if the coefficient for the interaction term was of the same sign as in the discovery and significant at the nominal one-sided threshold of 0.05.

3.2.2.4 Conditional Analysis

To rule out haplotype effects or the tagging of a rarer causal variant, we ran a conditional analysis for all the replicated SNP pairs. The principle of the conditional analysis is to re-run the logistic model for the interaction term but including the additive effect of all neighboring SNP, one at a time. Practically for each conditioning SNP, we compute the likelihood-ratio test (LRT) for one model (null model) with the additive effect of the 2 interaction SNPs and the conditioning and one model with the interaction effect in addition (full model).

null model :

$CCS \sim SNP_1 + SNP_2 + SNP_c + sex + age + MDS_1 + MDS_2 + \ldots + MDS_{10}$

full model :

$CCS \sim SNP_1 + SNP_2 + SNP_c + SNP_1{:}SNP_2 + sex + age + MDS_1 + MDS_2 + \ldots + MDS_{10}$

Where CCS is the Case/Control status and SNP_c the conditioning SNP.

If the p-value from the LRT is significant at the nominal threshold of 0.05 we conclude that the more complex model with the interaction term is significantly better and that therefore the association signal is at least partly due to epistasis. If this is the case for all of the conditioning SNPs, the epistasis pair passes the conditional analysis and is considered a true epistasis signal.

3.2.2.5 integration of recessive model results

All steps from the fast episcan in step 1 until the fine-mapping step 2 were conducted in parallel for the recessive model using a MAF threshold of 0.25 for the SNPs having the recessive encoding (the other SNP in the pair encoded in another genetic model came from the dataset thresholded at MAF 0.05). The recessive pairs which passed the fine-mapping step were added to the list of results from the other models if no other pair (or the pair itself) in the same LD block had a lower p-value using a different genetic model combination.

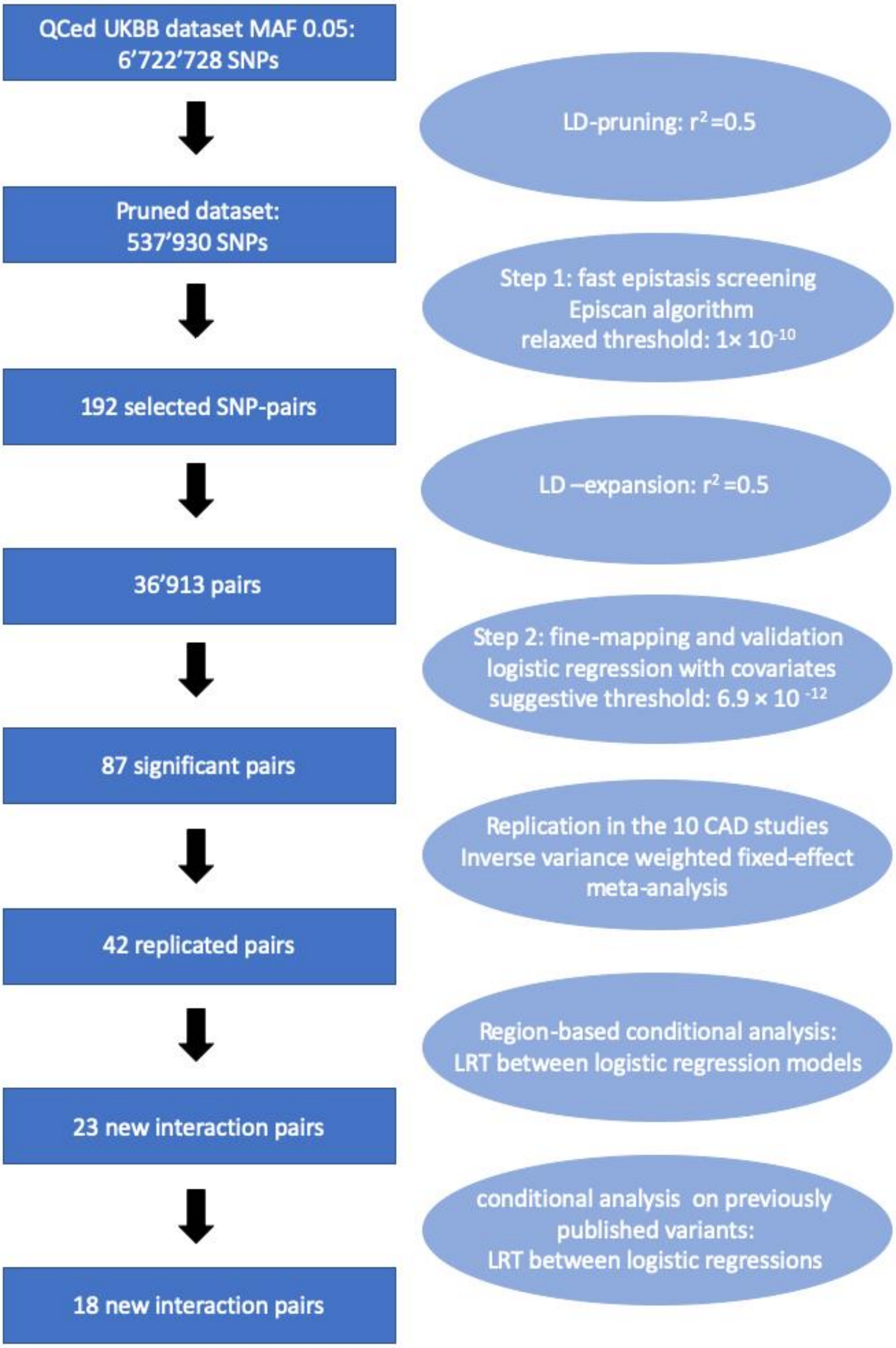

QCed UKBB dataset MAF 0.05:
6'722'728 SNPs
LD-pruning: r^2=0.5
Pruned dataset:
537'930 SNPs
Step 1: fast epistasis screening
Episcan algorithm
relaxed threshold: 1× 10^{-10}
192 selected SNP-pairs
LD –expansion: r^2 =0.5
36'913 pairs
Step 2: fine-mapping and validation
logistic regression with covariates
suggestive threshold: 6.9 × 10^{-12}
87 significant pairs
Replication in the 10 CAD studies
Inverse variance weighted fixed-effect
meta-analysis
42 replicated pairs
Region-based conditional analysis:
LRT between logistic regression models
23 new interaction pairs
conditional analysis on previously
published variants:
LRT between logistic regressions
18 new interaction pairs

Figure 14: Schematic representation of the analysis step
The dark blue boxes represent the data and the pale blue ellipses represent the analysis through which they go.

3.3 Results

3.3.1 Power computation and Episcan validation

3.3.1.1 Power computation

Power is a critical parameter in GWIAs because it determines how one can interpret the findings of the study. Indeed, the power tells us what were the chances to find an effect of a certain strength given the sample size, and in the case of genetic studies, the minor allele frequency of the variants of interest. Given the extremely high number of tests that one has to correct for in GWIAs without an a priori filter, the likelihood to have very low power to detect reasonable effect is high. As developed in the introduction, this can critically undermine the credibility of epistasis in human disorders and should be avoided. We, therefore, decided to conduct a power analysis before the GWIAs using the epilogitpower package that we developed and introduced earlier (see Section 2).

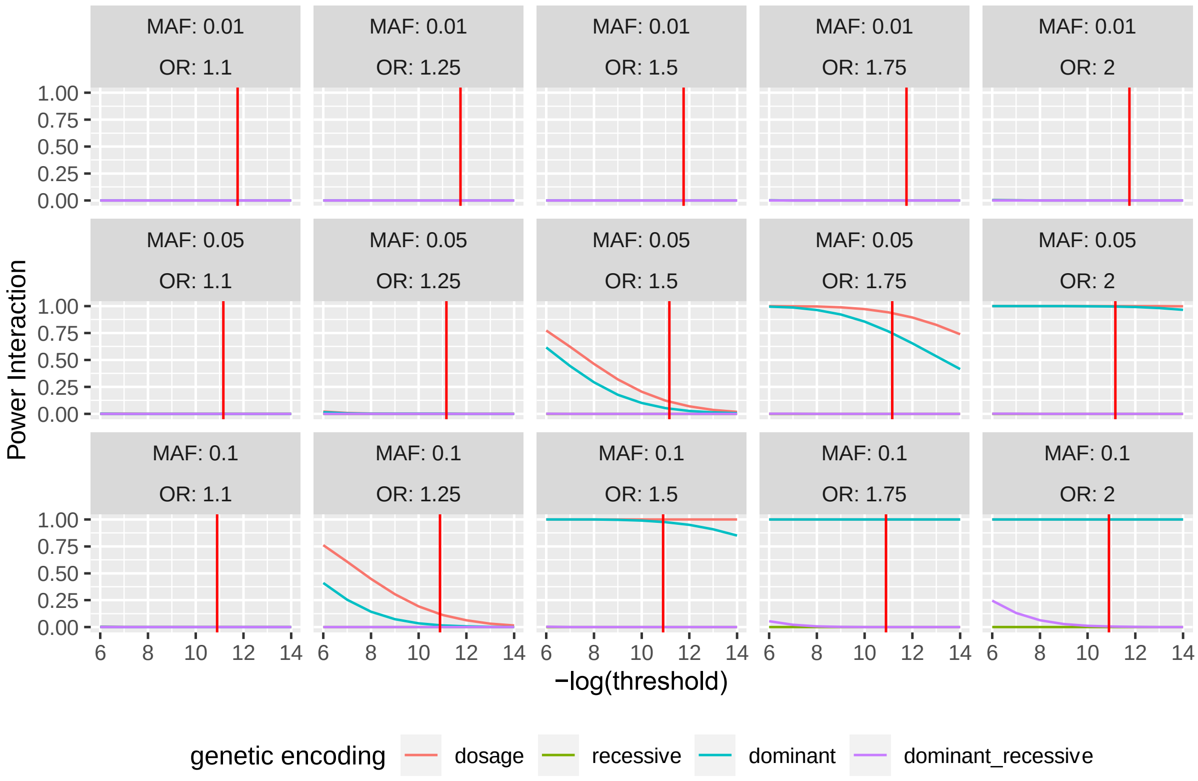


Figure 15: Power computation for the GWIAs in the UKBB dataset
Power curves for different genetic models encoding (colors) in function of the α threshold for different combinations of SNPs MAFs and interactions effect sizes. We assume here for simplicity that both SNPs have the same MAFs. The power was computed with the Epilogitpower R package described in chapter 2.

Figure 15 shows the results of the power computation in function of the MAF, and the OR of the interaction term which is a measure of the strength of the effect of the interaction on the disease risk. The vertical red line shows the multiple testing correction α threshold, computed as 1 divided by the number of SNPs passing QC at the corresponding MAF threshold. The plot shows clearly, that with this sample size the power to detect interactions when both SNPs have a MAF of 1% is null even with a high OR of 2, which is already very high for a genetic predisposition to complex disease. We, therefore, conclude that our sample size did not allow us to consider SNPs with a MAF of 0.1%.

On the contrary, we report a power of 1 to detect interactions between SNPs with a MAF of 5%, an OR of 2 and a dominant or additive mode of transmission. The power goes down to respectively 75% and ca. 90% for these two genetic models when the interaction OR is 1.75 and to ca. 10 and 20% for an OR of 1.5. Although increasing the MAF threshold to 10% would allow us to have a very good power also with an OR of 1.5 we decided to conduct the analysis with a MAF of 5% for the following reason. A MAF of 10% would provide higher power for relatively strong effect sizes but would not be able to increase power to interesting levels for OR in the range of 1.25. At the same time, it would prevent discovering any association for SNPs with a MAF between 5% and 10%. In this case, because we are using one of the largest datasets available to date for genetic CAD studies, we argue that the cost of reducing the search space outweights the gain of power offered by using a higher MAF threshold and chose a threshold of 5%. Indeed, one of the goals of this study is to generate novel mechanistic hypotheses for the etiology of CAD, and no other studies would be able to do it with a higher power. However, we need to keep in mind the actual power of the study with these settings when interpreting the results.

This power computation also clearly shows that much higher sample sizes will be needed to provide high power GWIAs for effect sizes lower than 1.2 and for hetero-dominant and recessive transmission modes.

3.3.1.2 Episcan validation and limitations

The first step of the analysis pipeline is a fast screen for SNP pairs showing evidence for epistasis using the episcan R package. This package implements the method published by Kam-Thong and colleagues (Kam-Thong et al., 2011) to detect pairwise epistasis using the difference in correlation coefficient between two SNPs in the cases and control group:

$$\Delta\rho\ (X^{(A,B)}, Y) = \frac{1}{n_1}\sum_{i:y_i=1} X_i^A X_i^B - \frac{1}{n_0}\sum_{i:y_i=0} X_i^A X_i^B$$

where, X_i^A and X_i^B are the scaled encoded genotypes at in the two interacting SNPs for in the cases (y_i=1) and control (y_i=0).

The idea behind this algorithm is that if two SNPs do not interact to increase the risk of disease, they should be uncorrelated in both groups because of the random segregation law, or equally correlated in both groups if they are in LD. However, if they interact to increase the risk of disease, one would expect that they are correlated in the cases group but not in the control group. These differences in correlation coefficient have been shown to be normally distributed (Gretton, Borgwardt, Rasch, Schölkopf, & Smola, 2007; Wellek & Ziegler, 2009) and Kam-Thong and colleagues showed in 2011 that the p-value derived from them is a very close approximation of the p-value from a logistic regression testing the interaction of the two SNPs. The episcan algorithm, taking advantage of matrix algebra is however much faster than logistic model computation and offers the possibility to pre-screen epistasis pairs which then need to be validated using a logistic model with all important covariates.

The original publication and subsequent study always used epiblaster on minor allele dosage. Here we want to compute interactions between SNPs encoded in the 4 genetic models (additive (minor allele dosage), dominant, recessive, and heterozygous). Some of these models typically lead to lower frequencies for the interaction term and we, therefore, decided to assess the reliability of episcan in this configuration. To do this, we ran the episcan algorithm for every combination of 1100 randomly sampled SNPs across the whole MAF range of the dataset. We ran it for each of the 10 unique genetic-model combinations possible. Very interestingly, observation of the resulting QQ-plots shows the expected uniform distribution (expected under the null hypothesis) in most of the model combinations (Figure 16). However, we also observe strong inflation for the combination of the recessive model with the dosage and heterozygous models and with itself. These are indeed, the genetic model which produce the lower interaction term frequency. This suggests already that a higher MAF threshold should be used for these models to avoid spurious associations.

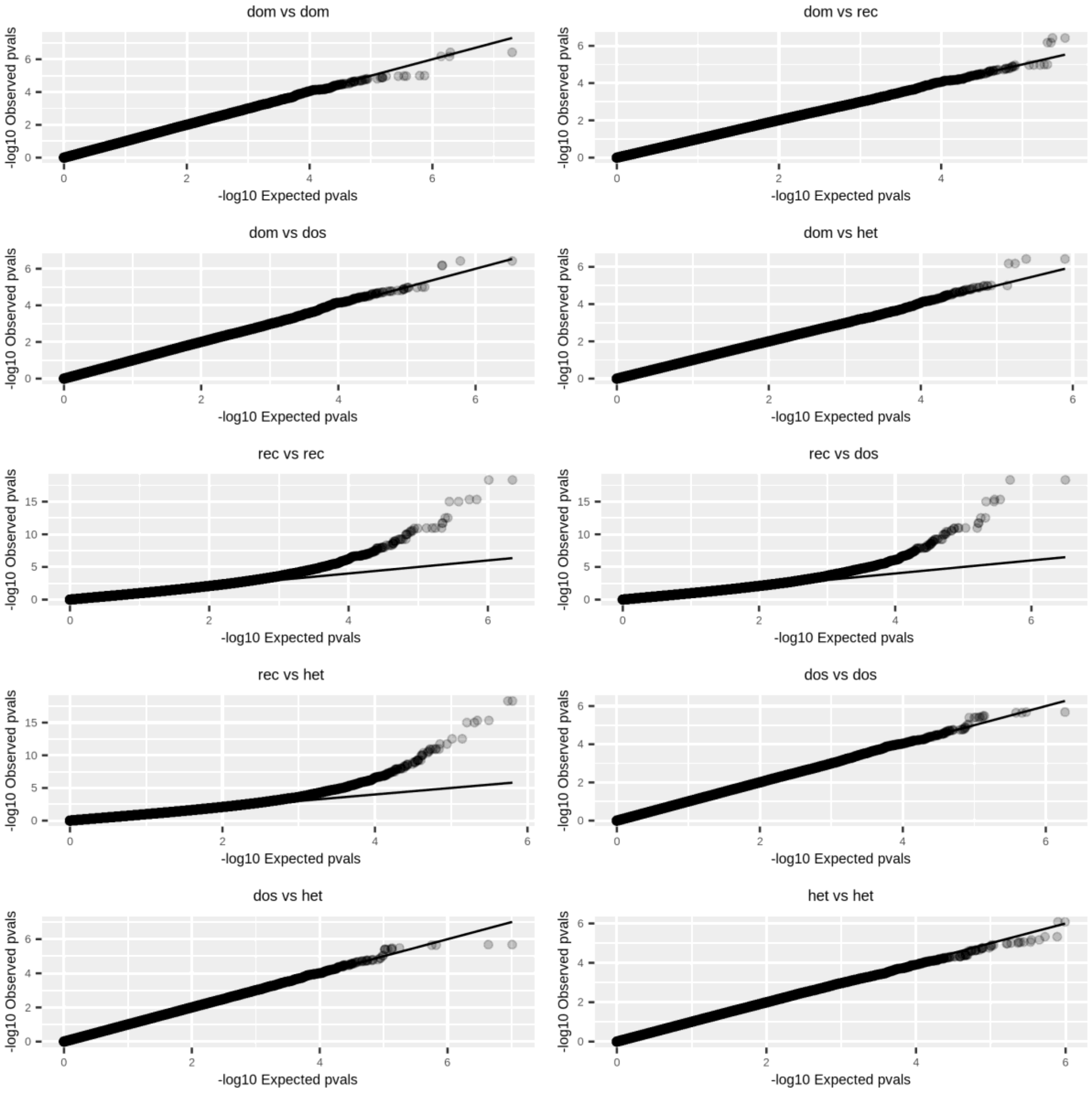


Figure 16: QQ-plots from episcan analysis using 1100 random SNPs
These QQ-plots represents the quantile of the observed p-value distribution when running the episcan using 1100 randomly sampled SNPs with a minimum MAF of 0.05 and encoded in the different genetic models plotted against the quantiles of the expected uniform distribution under the null hypothesis. Early deviation from the slope 1 line (solid black line) indicates p-value inflation in model combinations including the recessive genetic encoding.

To further examine the hypothesis that low MAF combined with these 3 genetic models yielding lower frequency leads to inflated p-value we plotted the distribution of the p-values according to the smaller MAF of the pair, to the largest MAF of the pair, and to the product of the two MAFs in the dominant-dominant model combination and in the recessive-recessive model combination.

In the dominant-dominant combination, we do already observe a trend of p-value tending to be higher with smaller minimum MAF or smaller interaction terms (Figure 17). However, the trend seems to be superficial. This is indeed coherent with the absence of inflation observed on the QQplot in Figure 16.

On the other hand, for the recessive-recessive model combination, we do observe a strong inflation of the p-value when the minimum MAF or product of the two MAF is below 5%. This is also coherent with the inflation of p-values observed on the QQplots in Figure 16.

These two analyses show that episcan is producing reliable results for SNPs with a MAF of 5% in all pairwise combinations of the dosage, dominant, and heterozygous genetic models, confirming that we can use it as a first screening step in our analysis. However, it does produce inflated results when one of the SNPs is encoded in the recessive models. We, therefore, decided to rise the MAF threshold to 25% when using the recessive encoding. With this, we ensure that the frequency of the interaction term will be at least 6.25% in the recessive-recessive combination and 1.25% in the recessive-other encoding combination.

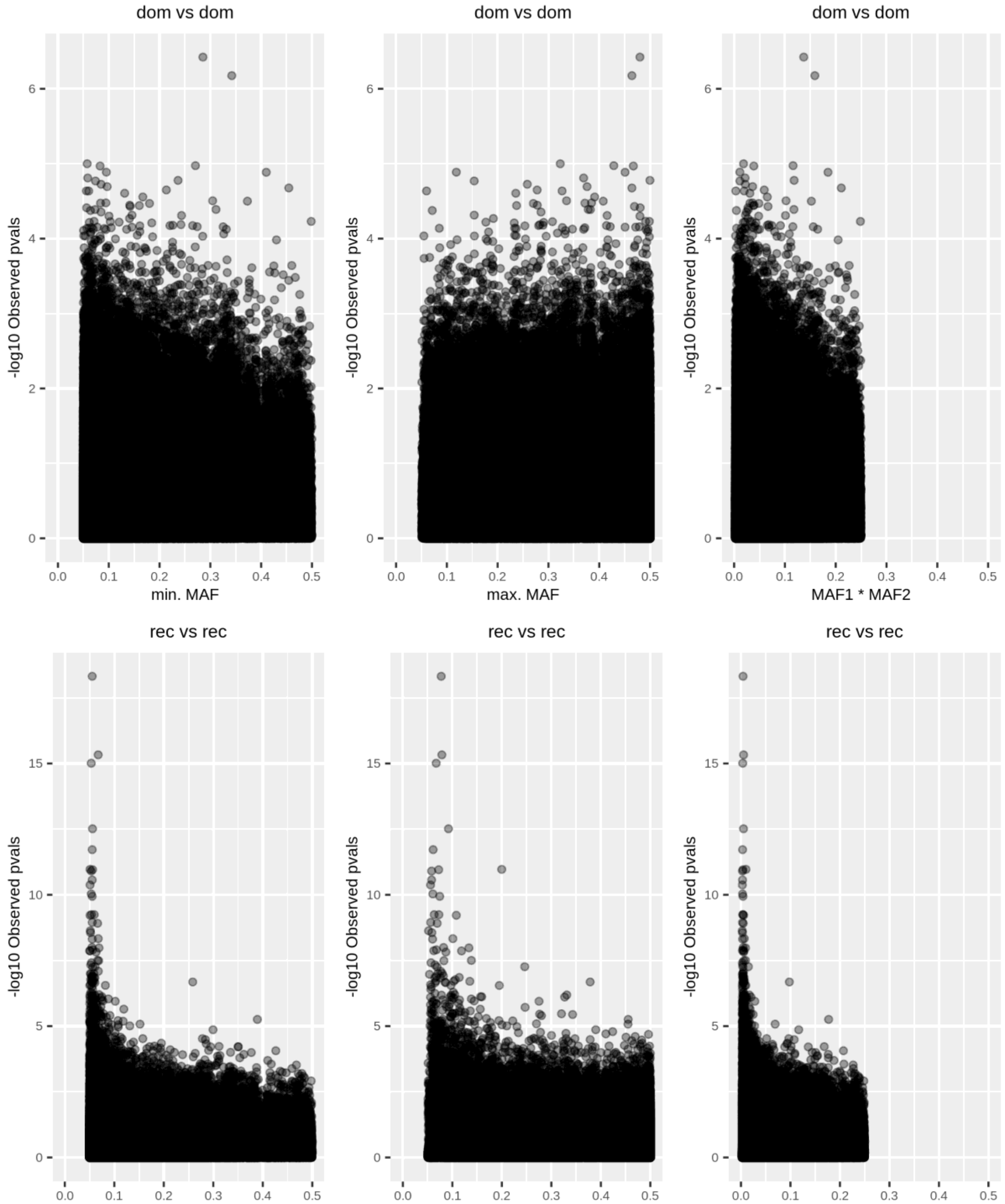


Figure 17: Episcan p-value distributions as a function of the MAF for the dominant-dominant and the recessive-recessive genetic encoding

P-values are plotted against the MAF of the rarest SNP of the pair (left), the MAF of the most common SNP of the pair (center), and of the product of the two MAFs (right). The upper row shows the p-values for the most powerful genetic encoding (dominant-dominant) and the lower row for the least powerful genetic encoding (recessive-recessive). Note the different y-scales between the two rows of plots.

3.3.1.3 Determination of the best episcan computing strategy

The SNPs QC at a MAF threshold of 0.05 filtered our dataset to 6'722'728 variants. This number of variants would already be high for a normal GWAS but is totally prohibitive in the case of a pairwise interaction study. For this reason, and to reduce collinearity between variants, we pruned all variants with an r^2 threshold of 0.5 using plink to arrive at 537'930 LD-independent variants. The number of pairwise comparisons to be tested can therefore be computed as (537'930 * (537'930-1))/2 and amounts to 1.446841×10^{11}. In terms of computational operations, the episcan algorithm will therefore have to compute 1.446841×10^{11} correlation coefficients between 314728 controls and 1.446841×10^{11} correlation coefficients between 17228 cases. Even with efficient matrix algebra this represents a very large number of computations and might take a very long time.

A graphical processing unit (GPU) accelerated version of the episcan package was also developed in our group by Jiang and colleagues. This version takes advantage of the architecture of GPUs and now available general purpose GPU (GPGPU) programming tools to run massively parallel computation of the correlation coefficients via matrix multiplication. This GPU implementation has the potential to substantially speed up the first statistical filtering step with episcan. Both CPU and GPU version are memory limited in the sense that the genetic data has to be loaded in memory for the computation. In case where the data exceeds the limit of the physical memory, the data has to be chunked into blocks of SNPs, which are treated one after another. In this respect the GPU version is at a disadvantage as memory available to a GPU is generally much smaller than RAM on an HPC machine.

Depending on the architecture of the computational resources at hand and the sample size of the dataset the CPU or GPU version might be faster. For example, on our cluster, each node has 2 GPUs with 4GB of RAM each. In the case of big datasets, the dataset will be chunked into numerous chunks because only part of it can fit into the GPU memory, and these chunks will be run sequentially thereby reducing the overall speed gain. On the other hand, each CPU node may have 40 parallel threads and between 256 and 758 GB of shared RAM (allowing an allocation of approx. 6–20 GB of RAM to every process) and could run 40 single-threaded episcan jobs in parallel. Even if each job is running slower than on a GPU, as many more can fit into the bigger available RAM, this might be quicker.

To choose the best computing strategy to run the genome-wide epistasis scan and estimate the total running time, we benchmarked the CPU and GPU versions of episcan in the following way. Using a simulated dataset we recorded the time needed for the computation of one SNP-block of different sizes and with different sample sizes using the CPU and GPU version of episcan. Figure 18 shows that in small datasets, the time increases at a rate close to 1 with the number of SNPs combinations between blocks (which is expected without parallelization). However, the GPU version shows only a very limited increase in computation time with increasing block sizes. It appears that the GPU version of episcan becomes advantageous for block sizes higher than 10000 SNPs. The last row of Figure 18 shows missing values for all block size larger than 1000. This is because the GPU could not accommodate larger block sizes with the sample size of 50'000 samples.

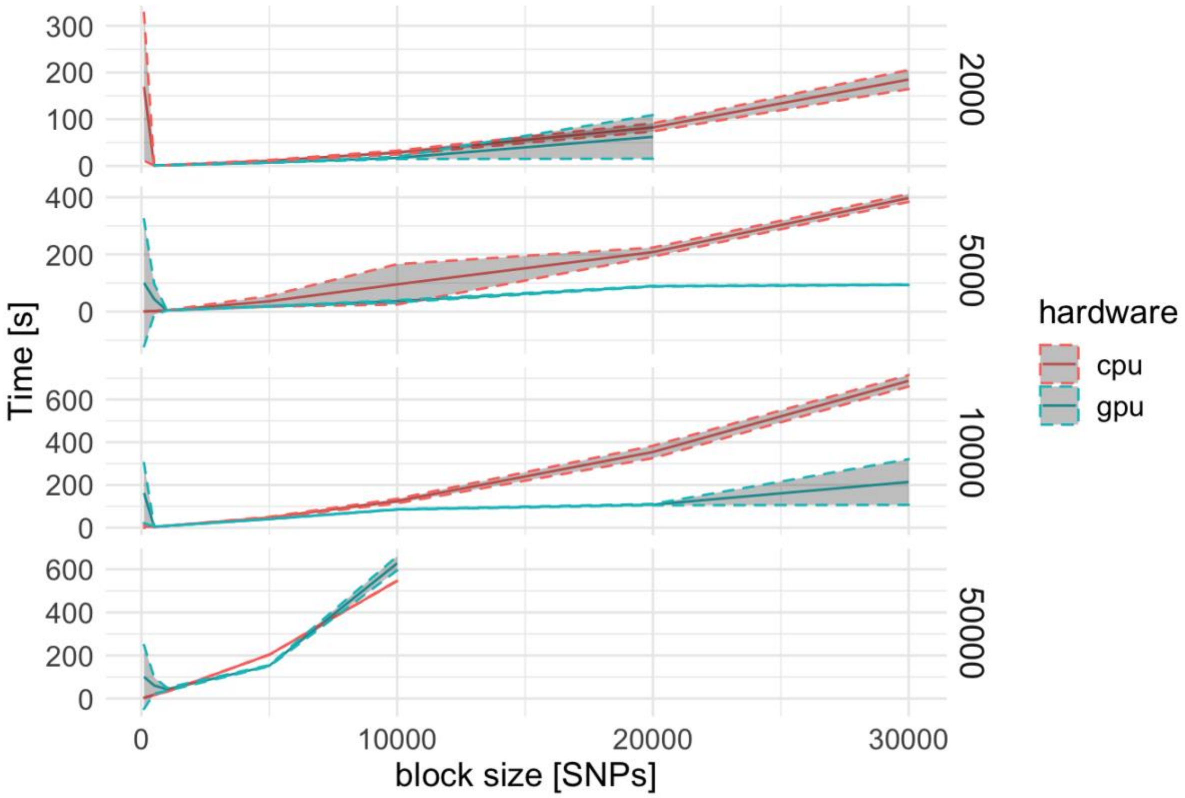


Figure 18: Computation time for one single SNP block of increasing size using the CPU or GPU version of episcan in four different sample sizes.

Indeed, Figure 19 shows the GPU memory requirements as function of the block and sample sizes. The maximum block size of 10000 for a sample size of 50000 is due to the fact that the 4GB RAM limit is reached. Moreover, in the case of the UKBB dataset, the memory limit will be reached with a block size of 1000 SNPs (red diamond on Figure 19).

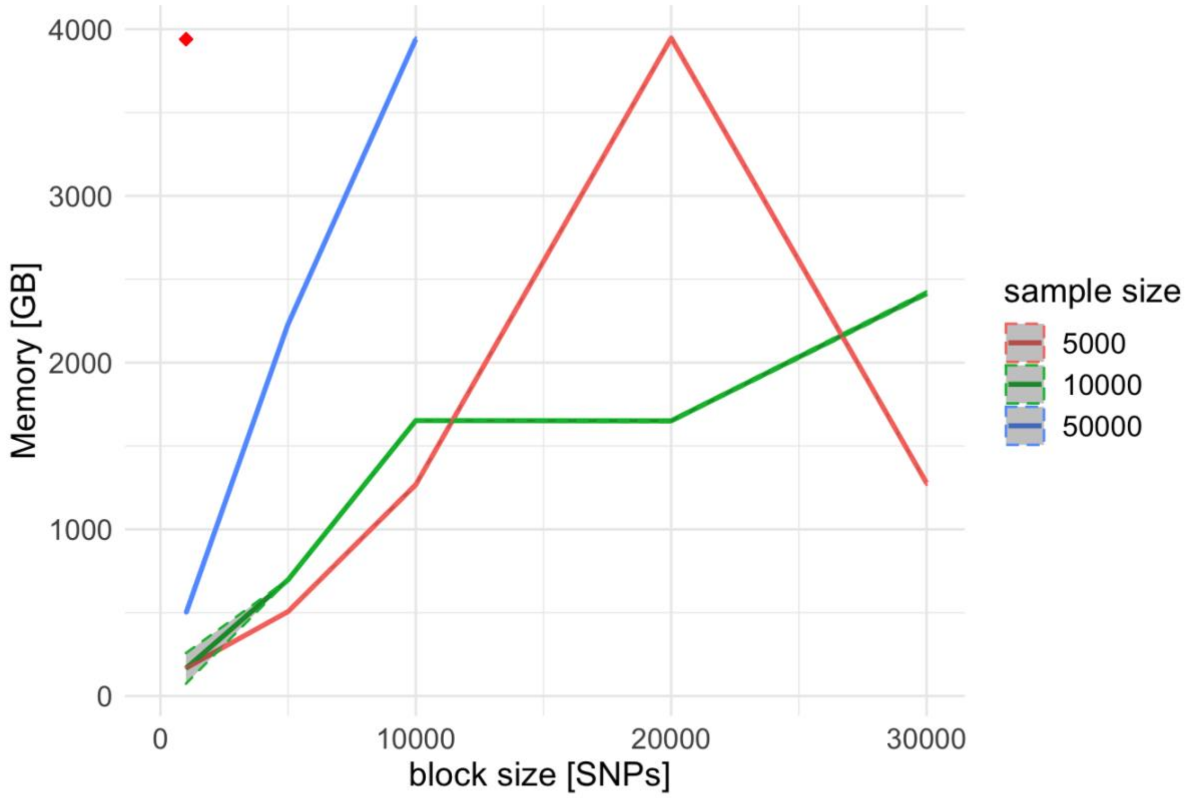


Figure 19: GPU memory usage as function of the block size in increasing sample sizes
The red diamond represents the UKBB dataset with ca 330‘000 individuals. For this dataset, there is no data for block sizes greater than 1000 because the GPU memory limit was reached with 1000 SNPs.

We can summarize the findings as follows: 1) GPU offers speed gain for block-size greater than 10'000 SNPs and 2) GPU is limited to block sizes smaller than 1000 SNPs for the UKBB dataset. It therefore appears that using single-threaded CPU episcan jobs which can be spawned in parallel to many CPU nodes will be the computationally fastest solution.

After establishing that we would use the CPU version of episcan, we still needed to investigate the optimal block size and parallelization strategy on the different cluster nodes. Figure 20 shows the evolution of computation time for different block size in different increasing sample size up to the UKBB sample size. We can appreciate that computation time increases almost linearly with the block size. This means that a when the block size doubles the computation time is multiplied approximately by 2 while 4 times more combinations have been computed. Larger block size would therefore speed up computations if unlimited ressources are available. However, Figure 21 shows that the required memory also increases almost linearly with the block size and quickly reaches several hundred GB for UKBB sample sizes. Given a particular HPC architecture, like the one that we were using, where each nodes has several cores (40 in our case) but less than 1TB of RAM (500GB in our case), only few large-block-size jobs can be run in parallel. This will results in most of the cores not being used. On the other hand, jobs with smaller block-size use less RAM and more of them can be run in parallel , thereby possibly speeding up the computation. Optimal computation time will therefore be reached by optimizing the block size with respect to one's computational ressources.

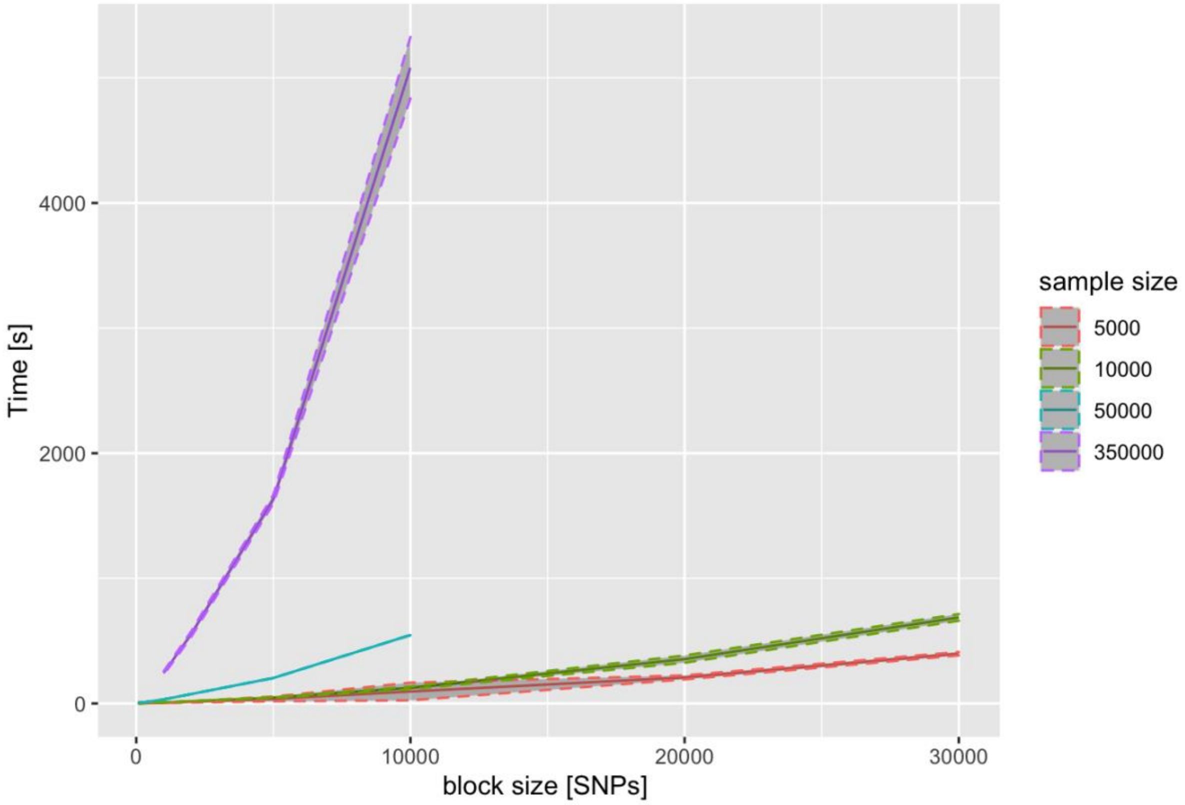


Figure 20: Computation time for single SNP block of different sizes in increasing sample size
The purple line represents a dataset of similar size as the UKBB dataset used in the analysis.

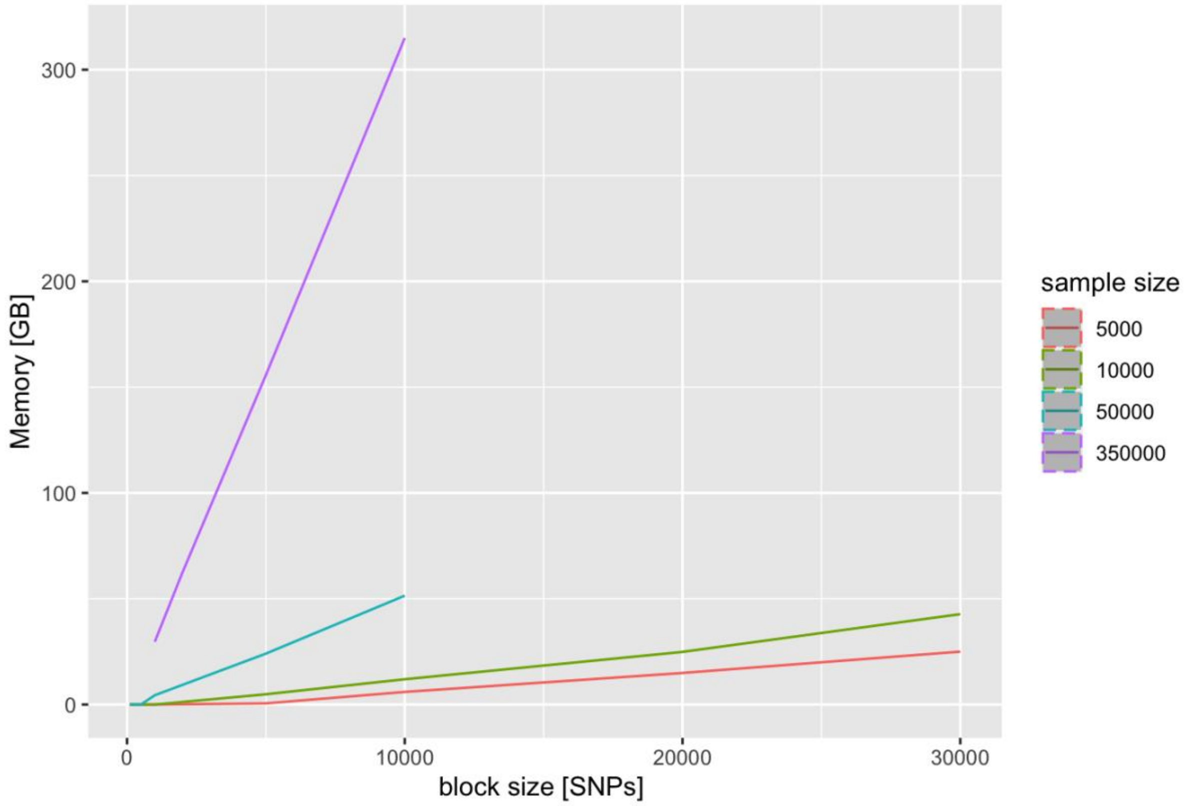


Figure 21: CPU memory usage for single SNP block of different sizes in increasing sample size
The purple line represents a dataset of similar size as the UKBB dataset used in the analysis.

I therefore used the measured computation time and memory requirements for the different block sizes to estimate the total computation time for different numbers of SNPs (corresponding to different MAF thresholds) using the following resources reasonably available on our cluster: 4 nodes with 40 threads and 500 GB RAM each. We can see that indeed, smaller block sizes and therefore less resource-heavy jobs yield shorter total computation time (Figure 22), certainly because they can maximize resource allocation on the cluster. Based on this, I selected a block size of 1000 SNPs.

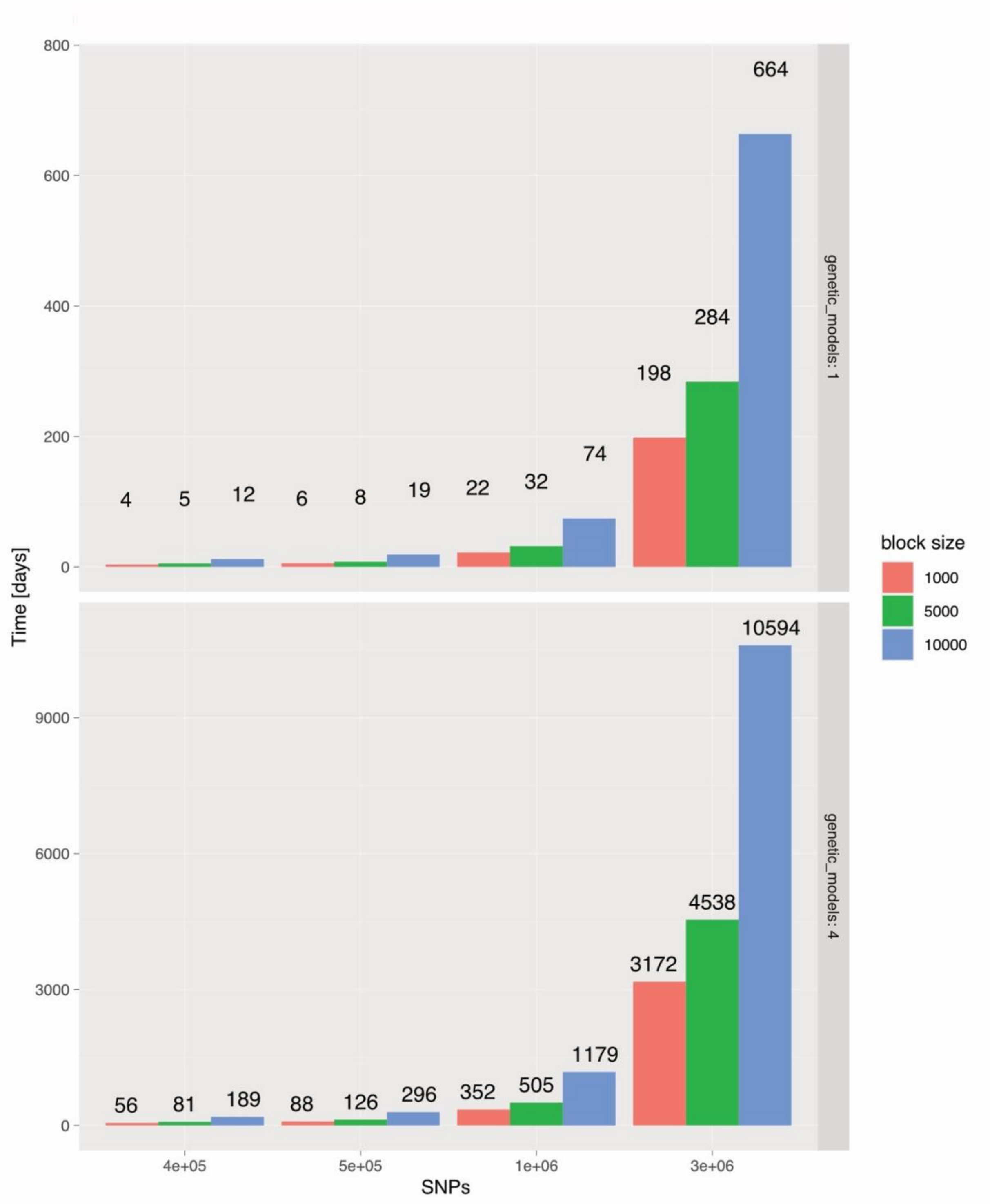


Figure 22: Estimated computation time in days as function of the block size for different total numbers of SNPs

The bar plots shows the estimated computation time in days for different number of SNPs roughly corresponding to different the different MAF threshold of 0.1, 0.05, 0.01, and 0.001 respectively for three different block size. The upper panel represents the computation time if all SNPs are encoded only in the dosage model and the lower panel if all SNPs are encoded in the four different genetic models. Computation time was estimated based on the following hardware resources: 4 HPC nodes with 40 threads and 500 GB RAM each.

Finally, using the selected block size I could estimate the overall computation time for different MAF thresholds (yielding different SNPs number) for the case when one genetic model is considered and the case when all combinations of the 4 genetic models are considered. Figure 23 shows that a MAF threshold of 0.05 allows to run the epistasis scan with all genetic models in roughly 3 months. Lowering the MAF threshold to 0.01 increases the run time prohibitively to almost one year.

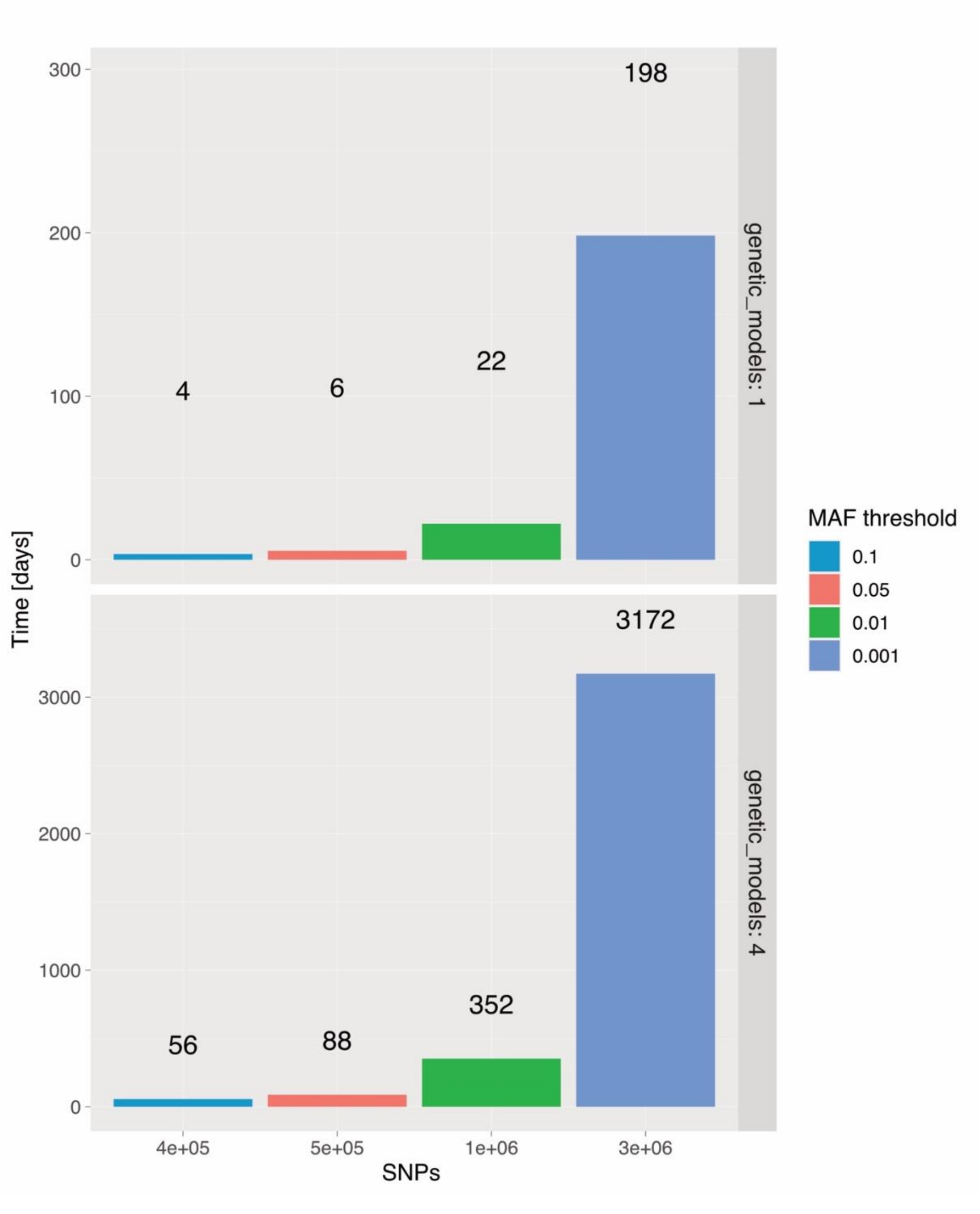


Figure 23: Estimation of the overall computation time in days for the chosen block size of 1000 SNPs for different total numbers of SNPs

The bar plot shows the estimated computation time in days for different number of SNPs roughly corresponding to different the different MAF threshold of 0.1, 0.05, 0.01, and 0.001 respectively for three different block sizes. Computation time was estimated based on the following hardware resources: 4 HPC nodes with 40 threads and 500 GB RAM each. The final estimated computation time for the chosen MAF threshold of 0.05 is roughly 3 months wall-clock time.

3.4 Results

3.4.1 Step 1: Episcan

We ran the first statistical screening step using the episcan R package on a QCed UKBB dataset with 314’728 controls and 17’228 cases. The pairwise combinations of the 537’930 SNP encoded in the dosage, dominant and recessive models amounted to a total of 1.446841×10^{12} epistasis pairs to be computed.

We report that 1’210’213 genetic-models-specific pairs passed the threshold of 1×10^{-6}, which we set as the minimum threshold for writing the pair’s detail to disk for storage reasons. We computed a Quantile-Quantile Plot (QQplot) using all these pairs to assess whether inflation in p-value could be observed. The QQ-plot does not show any signs of such inflation (Figure 24), with the quantile-quantile distribution following the slope 1 line until a very low p-value threshold (notice that the x and y axis do not start a 1 but at 1×10^{-6}). On the contrary, we start to see observe p-values rising above the line from the loose statistical threshold of 1×10^{-10} (vertical blue line) which we used for this first step. This argues that the signal retained with this threshold is likely to be true.

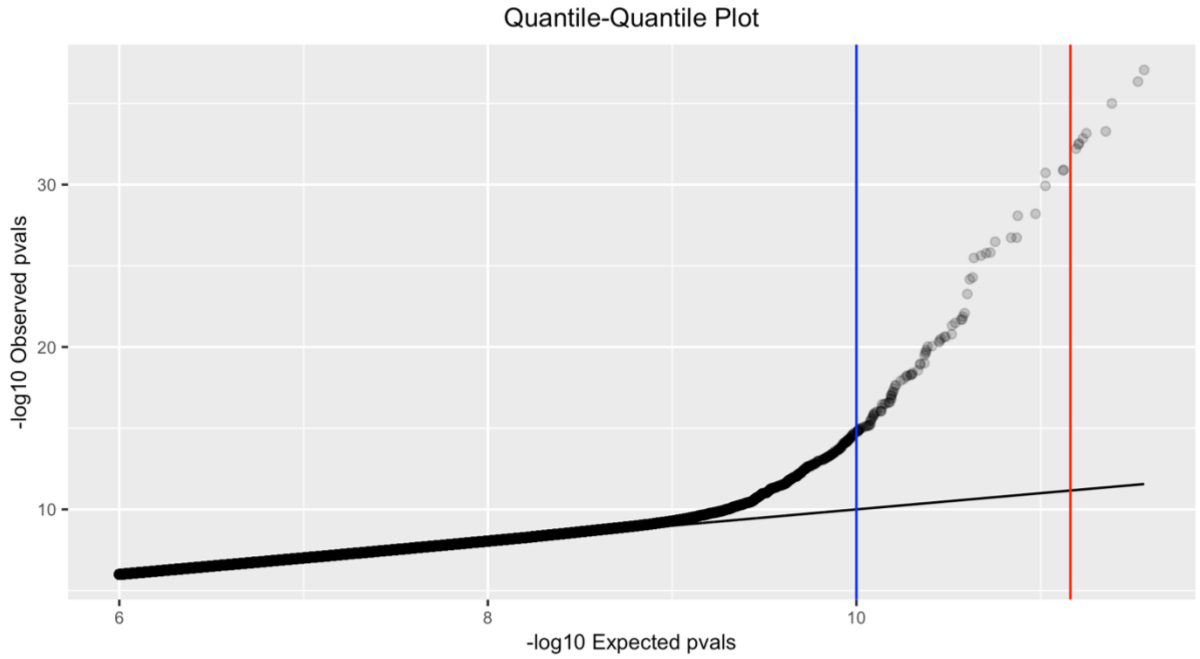


Figure 24: Quantile-Quantile plot for all pairs with a p-value < 1×10^{-6} in the episcan run
This QQ-plot shows the distribution of the observed p-value on the y-axis compared to the distribution of expected p-values under the null hypothesis (uniform distribution) on the x-axis. The black line represents a line of slope 1. The blue vertical line represents the loose statistical threshold of 1×10^{-10} and the red line represents the suggestive threshold of 6.9×10^{-12}. Both distributions start at 1×10^{-6} because lower p-values were not recorded because of storage reasons.

To get a better overview of the episcan results, we adapted the traditional Manhattan plot from GWAS studies to plot the epistasis results using a 3D scatter plot of the results (Figure 25). It strikingly appears that one association signal showing the characteristic "skyscraper" shape is much higher than any others. This signal between rs4708870 and rs575905913 is a cis-epistasis signal located on chromosome 6 between the LPAL2 and LPA genes. We observe a second very high association signal, between rs10757275-rs7857345 also in cis. These SNPs are both annotated to the CDKN2B-AS1 gene on chromosome 9. These two signals are composed of several epistasis pairs apparently in LD with each other with increasing levels of significance and therefore show good evidence for a true signal. In addition, these 3 loci are known risk factors for cardiovascular disease and we show in chapter 5.3.3 that interactions at the LPA locus are associated with CAD, thereby strengthening the claim that the episcan yielded plausible results.

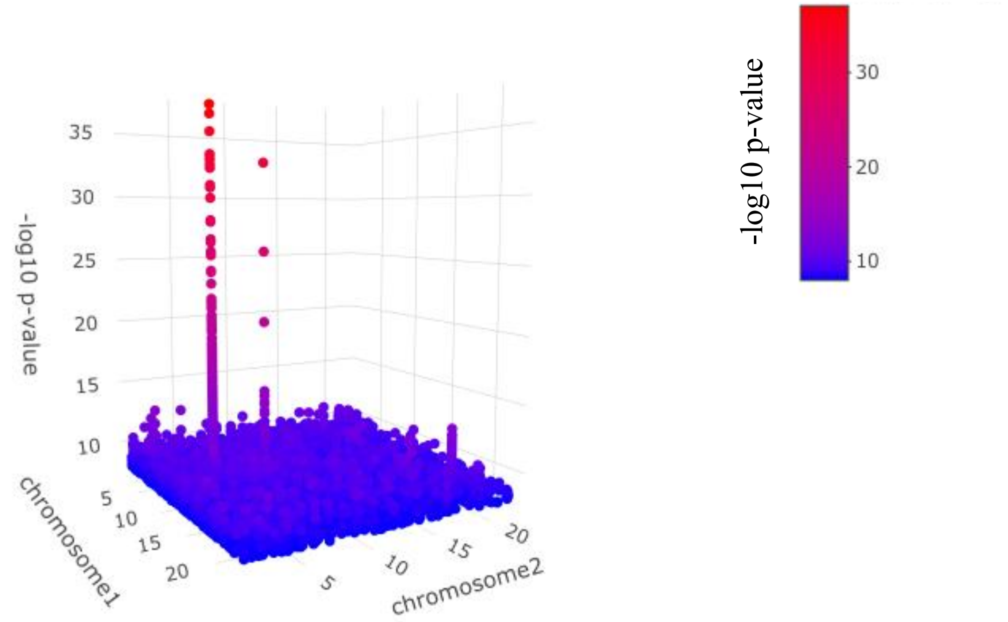


Figure 25:_3D Manhattan-like plot of the first filtering step using episcan
The x and y axis represent the interacting SNPs on the first and second chromosomes respectively. The z-axis and color represent the p-value of their interaction. Each SNP pair is represented only in the model combination which yielded the best p-value.

Generally, it appears that most of the signal above the threshold of 1×10^{-10} shows the common "towering" structure and are not singletons, as can be appreciated from the z-axis-truncated 3D Manhattan plot (Figure 26).

Together, with the QQ-plot, these results advocate that the episcan did pick real association signals and was not confounded even if it cannot consider any covariates.

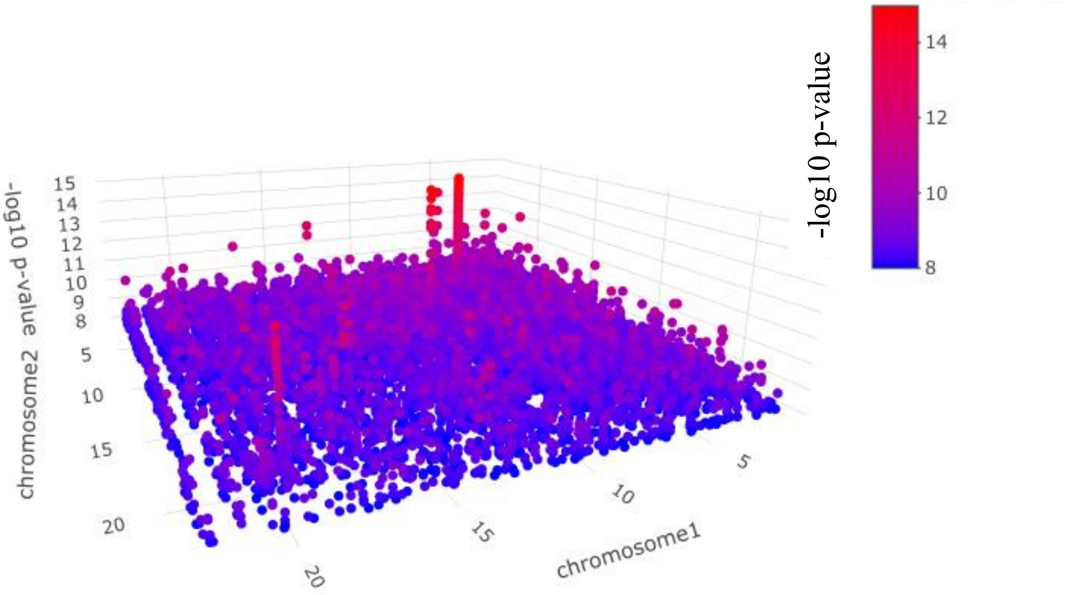


Figure 26: Truncated 3D Manhattan-like plot of the first filtering step using episcan
The x and y axes represent the interacting SNPs on the first and second chromosomes, respectively. The z-axis and color represent the p-value of their interaction. The plot was truncated on the z-axis at 1×10^{-14} to allow better visualization of the pairs with p-values in the range between 1×10^{-8} and 1×10^{-14}, which comprises all pairs except for the two “highest towers”. Each SNP pair is represented only in the model combination which yielded the best p-value.

In total, we report 192 epistasis pairs that did pass the relaxed threshold of 1×10^{-10} and will be forwarded to the next validation and fine-mapping step. Among them, 104 cis-epistasis pairs and 88 trans pairs. In comparison to the suggestive threshold of 6×10^{-12} or a more relaxed arbitrary threshold of 1×10^{-8}, it appears that the chosen threshold of 1×10^{-10} yields a manageable number of hypotheses to test while allowing for possible significance decrease due to imperfect LD between tags and causal variants (Figure 27).

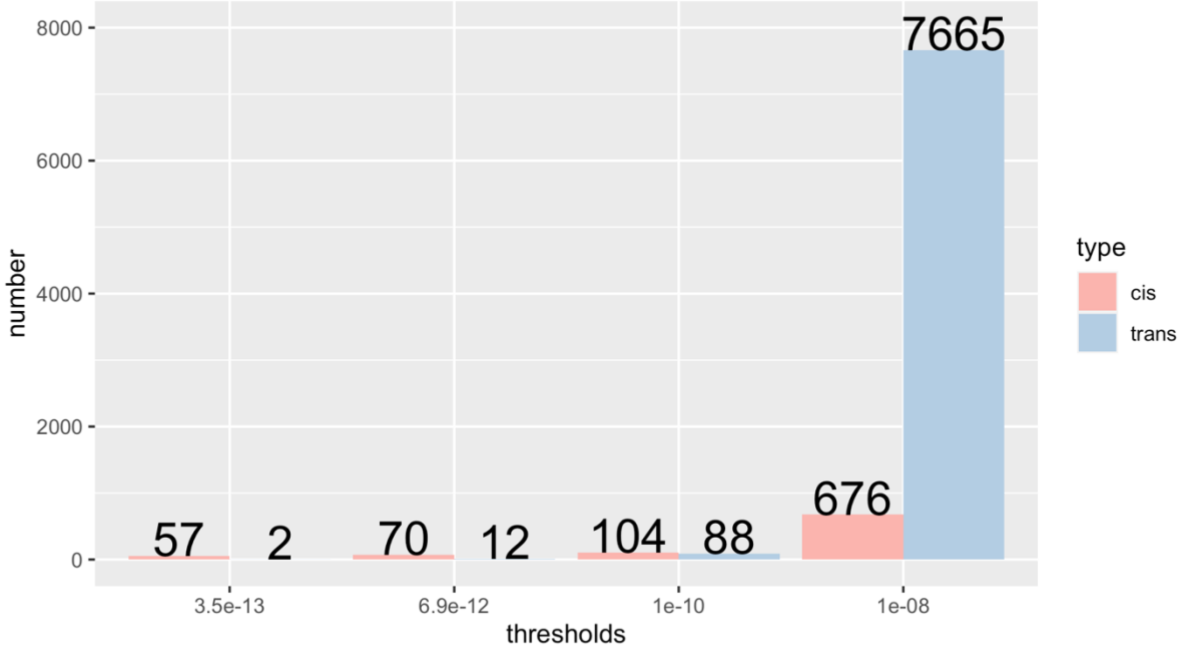


Figure 27: Summary of the significant results in the first filtering step using episcan
This barplot summarises the number of cis (red) and trans (blue) epistasis pairs at different significance threshold corresponding to, respectively and in ascending order, Bonferonni correction (0.05/number of LD independent pairs), suggestive threshold (1/number of LD independent pairs), relaxed threshold to capture incomplete LD between tags and causal variants, more relaxed threshold.

3.4.2 Step 2: Fine mapping and validation using logistic regression

In order to fine-map and validate the 192 epistasis pairs from step 1, we first searched for all tag-SNPs in LD with the SNPs forming the selected pairs (referred to hereafter as "step 1-pairs"). We did this search in the exact inverse way as compared to the LD pruning, using an r^2 threshold of 0.5 within a 10'000kb distance. For each of the step 1-pairs, we then listed all combinations of their tagged, resulting in 36'913 pairs SNP (referred to as step 2-pairs hereafter). The number of interaction pairs per step 1-pair follows a broad distribution as shown in Figure 28. We can see on this histogram that most of the step-1 pairs generated less than 1000 pairs to test. If we zoom in on the range 0–1000 (Figure 28) we can see that in fact most of the step 1-pairs generate less than 50 pairs.

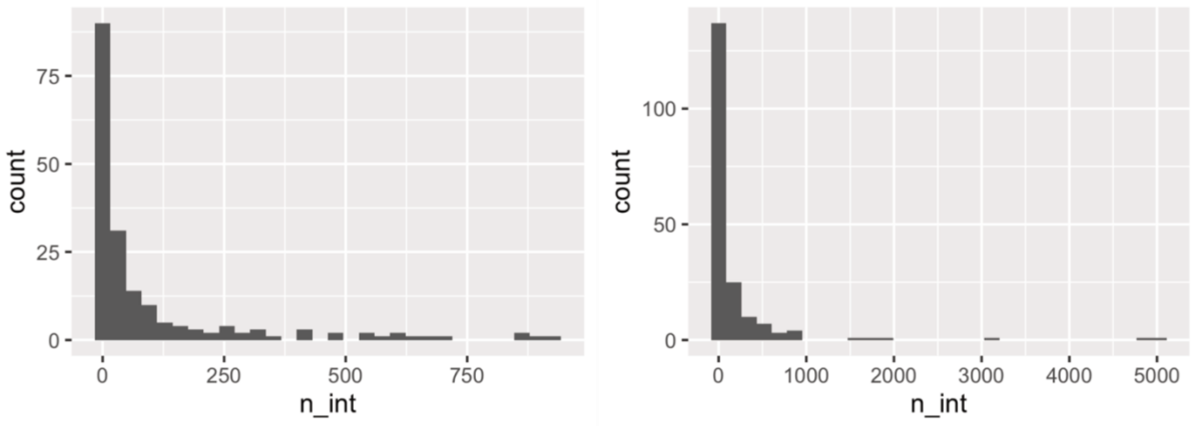


Figure 28: Number of "step 2-pairs" per "step 1-pair"
This histogram shows the distribution of the number of pairs created by the LD-expansion (see Section 3.2.2.2) of each of the significant pairs in the first step using episcan. The left panel shows a zoom-in on the range 0 to 1000.

We then tested each of the 36'913 tag pairs using a logistic regression model with the case status as outcome, the SNPs in the respective genetic encoding as the predictor, and sex, age as well as the 10 first MDS components of the genetic variance as covariates. At this stage, we decided to use the suggestive threshold of 6.9×10^{-12} computed as 1/number of LD independent SNP pair combinations which entered step 1. We wanted here to apply a more stringent threshold than at the first step, as the variants used in this fine-mapping should now be the causal variants or at least the best tag variant to the causal ones. We decided to use this suggestive threshold instead of the usual Bonferroni threshold (in this case 3.5×10^{-13}) because it yielded substantially more trans-epistasis pairs which we could try to replicate (Figure 29). Indeed, trans pairs are less likely to be tagging rarer variants or haplotypes and are therefore of great interest. The replication step being computationally very cheap compared to the whole two-step association testing, we decided to use the suggestive threshold and try replicating the 79 pairs significant at this threshold.

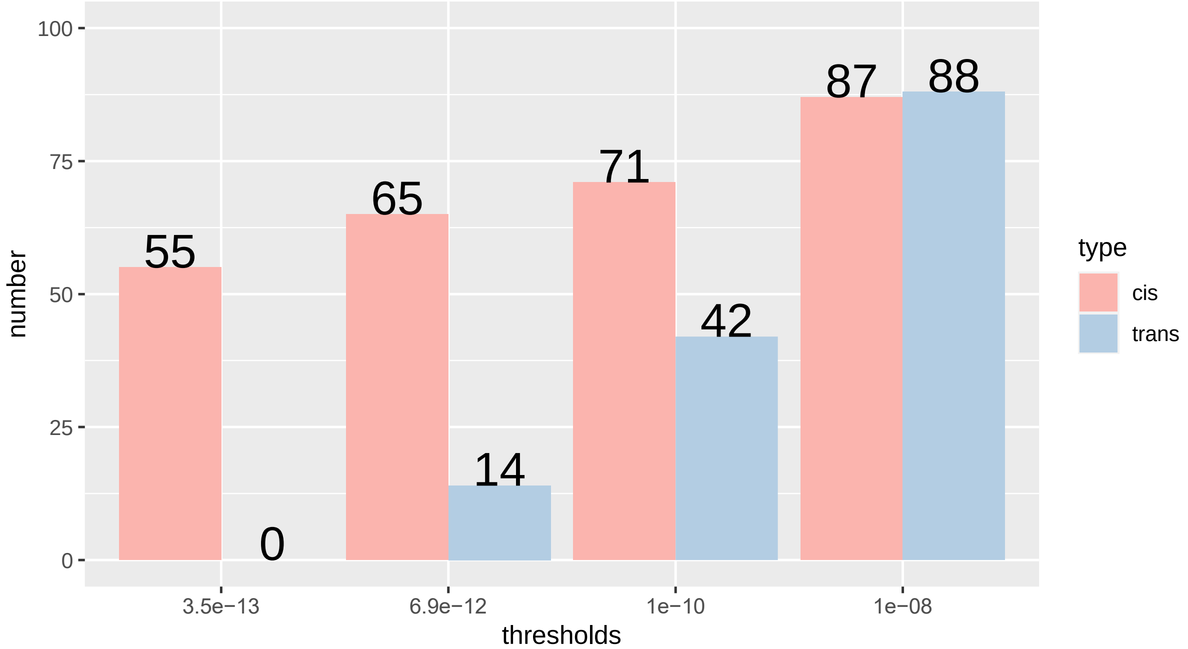


Figure 29: Summary of the significant results in the fine-mapping step using logistic regression
This barplot summarizes the number of cis (red) and trans (blue) epistasis pairs at different significance thresholds corresponding to, respectively and in ascending order, Bonferroni correction (0.05/number of LD independent pairs), suggestive threshold (1/number of LD independent pairs), relaxed threshold to capture incomplete LD between tags and causal variants, more relaxed threshold.

3.4.3 Episcan and fine-mapping for recessive models

We repeated the episcan and fine-mapping step for all combinations of SNPs for models containing at least one recessive encoding (i.e., for the models recessive-recessive, recessive-dominant, recessive-dosage, and recessive-heterozygous). For the recessive encoding, we did consider only SNPs with a MAF > 0.25 as we showed that under this threshold the episcan results were not reliable in this encoding. The Quantile-Quantile plot for these models indeed does not show any signs of inflation, on the contrary, there seems to be a p-value deflation (Figure 30). This is certainly because the power to detect association when one SNP is encoded in the recessive model is lower than for others. Raising the MAF threshold did avoid the aberrant behavior of the episcan algorithm but did not increase the power sufficiently. We, therefore, decided to keep these results but acknowledge that the power of this study in these models' space is low.

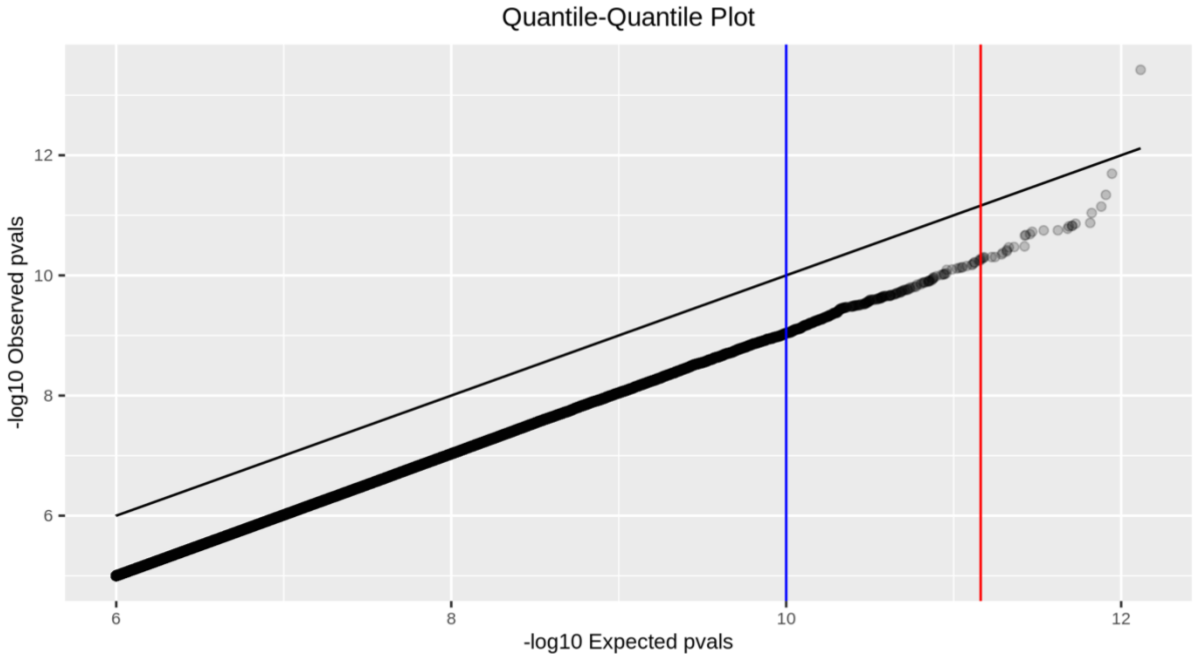


Figure 30: Quantile-Quantile plot for all pairs with a p-value < 1×10^{-6} in the episcan run with the recessive genetic models
This QQ-plot shows the distribution of the observed p-value on the y-axis compared to the distribution of expected p-values under the null hypothesis (uniform distribution) on the x-axis. The black line represents a line of slope 1. The blue vertical line represents the loose statistical threshold of 1×10^{-10} and the red line represents the suggestive threshold of 6.9×10^{-12}. Both distributions start at 1×10^{-6} because higher p-values were not recorded because of storage reasons.

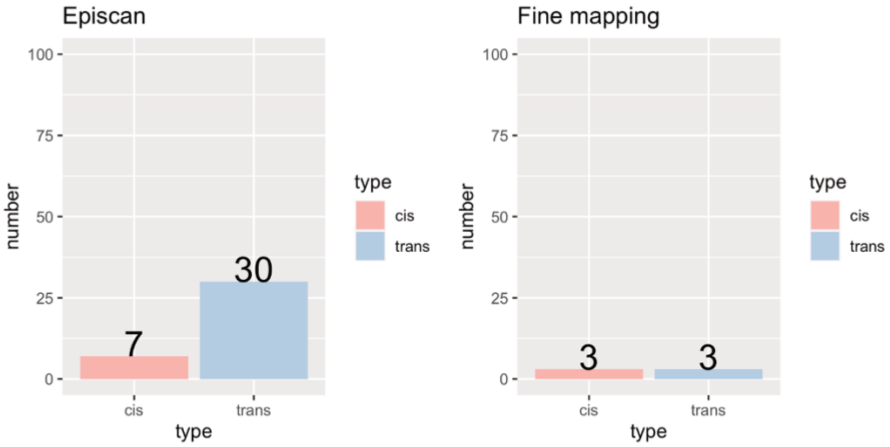


Figure 31:_Summary of the significant results for the recessive models
Number of significant SNPs at the suggestive threshold of 6.9×10^{-12} in the first (episcan) and second (fine-mapping) step for the recessive models.

We report 37 pairs that passed the episcan at the threshold of 1×10^{-10} (Figure 31 left). Among them, are 7 cis and 30 trans pairs. Interestingly, there are many more significant trans pairs than cis pairs in this case, whereas it was the contrary with the remaining model conditions (Figure 27). We do not have a very good hypothesis as to why this is the case, but these associations could be due to some hidden confounders. Indeed, only 3 of the 30 trans-pairs reached significance in the second step (Figure 31 right) when using a logistic model correction for known confounders, as compared to 3 out of 7 for the cis pairs. Out of these 6 pairs, 3 had not been identified or had a lower p-value in the other model's combinations. Therefore, the recessive model's combinations brought only 3 new pairs which were not discovered using the other models.

3.4.4 Replication

Taking the results of non-recessive and recessive model combinations, we report a total of 82 SNP-pairs, 65 cis, and 17 trans, to be replicated in the replication dataset made of 10 CAD studies. We ran the replication as a fixed effect inverse-variance weighted meta-analysis of the 10 studies, and considered a pair replicated if the interaction coefficient was of the same sign as in the discovery sample, and its p-value significant at the nominal threshold of 0.05 (Figure 32). Unfortunately, 32 of the pairs were impossible to replicate because neither the two SNPs nor any of their LD-tags (down to r^2=0.5) were present in the replication dataset.

Among the pairs that succeeded replication, were 36 cis-pairs, and only one trans-pair. Replication was impossible due to missing data for 7 of the 17 trans pairs and the other 9 failed replication. Among pairs which could technically be replicated, the replication rate is therefore very high in the cis-pairs and very low in the trans pairs. The replicated trans-pair between rs3895825 and rs4947818 is extremely unlikely to tag a rare variant or haplotype as the two SNPs are located on different chromosomes and belongs to our final list of epistasis (Table 2).

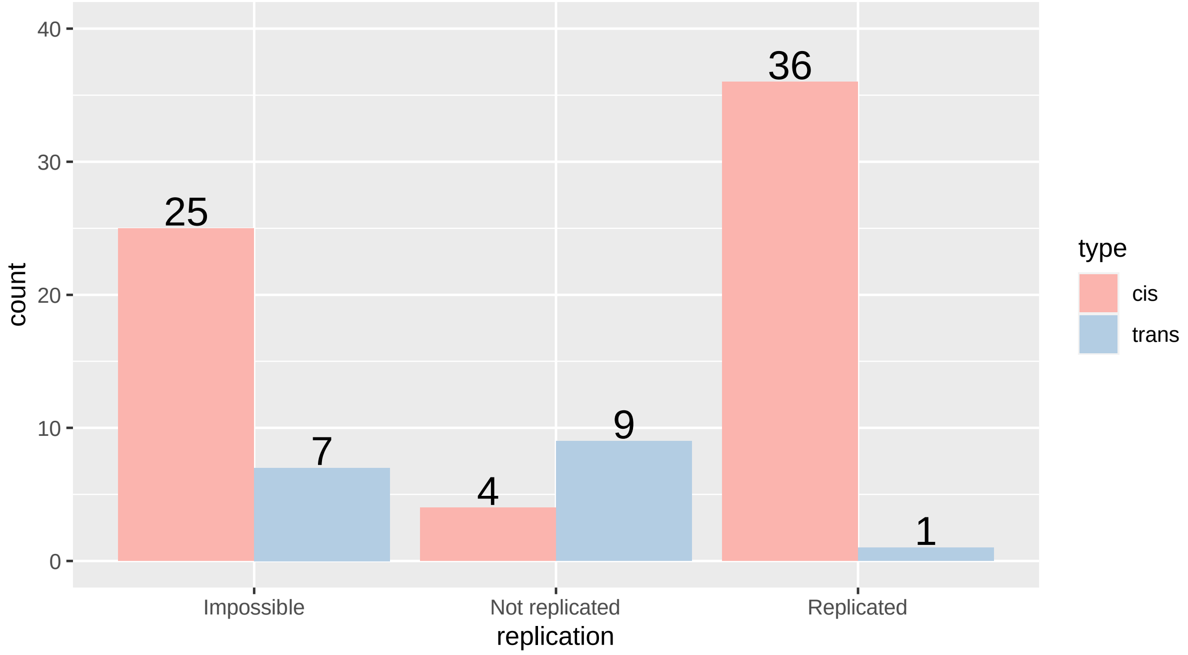


Figure 32: Replication of the significant pairs in a meta-analysis of 10 CAD studies
This barplot shows the results of the replication of the 90 significant pairs after the fine-mapping in a replication dataset composed of 10 CAD studies. Some of the pairs were impossible to replicate in this dataset because there were no LD tags (LD < 0.5) available in the replication dataset which was imputed using a smaller imputation panel.

3.4.5 Conditional analysis for cis-pairs

The final step for the validation of the 36 cis-pairs which were replicated in the replication dataset is to make sure that they are not merely tagging a rare variant or haplotype. Indeed, the combination of the two SNPs could be just co-inherited with a causal rarer variant for example. We, therefore, ran the conditional analysis (as described in section 3.2.2.4) for every SNP in the pre-QC dataset (i.e., in every imputed SNP regardless of the MAF, imputation quality, or other QC criteria) within a ±200kb distance from the SNPs of the pair. Indeed, it is important to include as many SNPs as possible in this analysis because the tagged variant does not have to fulfill any particular QC criteria. We chose this space window because the risk of co-inheritance decreases with the distance and 200kb yielded a manageable number of tests to run.

We report that 22 SNPs pairs did survive this conditional analysis (Table 2) and are therefore independent of any additive single variant effect, in the limit of the SNPs which were available to us for this analysis.

SNP1	SNP2	Chr1	Chr2	MAF1	MAF2	Gene1	Gene2	Model1	Model2	Pval step1	coef. step2	Pval step2	coef. repl.	Pval repl	Indep. LPA	Indep 6q26–q27	Indep CDKN2A/B
rs9632884	rs581876	9	9	0.48	0.39	het	dom	CDKN2B-AS1	CDKN2B-AS1	4.30E-12	0.38	1.80E-28	0.362	4.60E-10	-	-	YES
rs11751605	rs1084651	6	6	0.17	0.17	dom	dom	LPA	Intergenic (LPA)	3.20E-20	-0.38	1.30E-24	-0.322	1.80E-06	YES	NO	-
rs2661839	rs2872819	6	6	0.35	0.31	dos	dos	SLC22A3	Intergenic (LPA)	7.00E-16	-0.19	2.00E-24	-0.116	3.10E-04	YES	YES	-
rs7742018	rs4252051	6	6	0.49	0.16	dos	dos	Intergenic (LPA)	Intergenic (PLG)	9.80E-14	-0.23	2.00E-24	-0.094	1.40E-02	YES	YES	-
rs12210186	rs1084651	6	6	0.19	0.17	dom	dom	LPA	Intergenic (LPA)	8.30E-18	-0.37	8.90E-24	-0.318	1.60E-06	YES	YES	-
rs13211753	rs9457931	6	6	0.07	0.06	dom	dom	SLC22A2	LPAL2	1.90E-31	0.54	1.30E-23	0.559	2.30E-08	YES	YES	-
rs2457571	rs2314851	6	6	0.47	0.48	dos	dos	SLC22A3	PLG	1.40E-14	-0.16	2.10E-22	-0.057	4.10E-02	YES	YES	-
rs7857118	rs62555370	9	9	0.49	0.13	het	dom	Intergenic (CDKN2B-AS1)	CDKN2B-AS1	3.90E-14	0.39	5.90E-22	0.432	3.90E-09	-	-	YES
rs3125056	rs13198987	6	6	0.14	0.17	dos	dom	intergenic (SLC22A1)	LPA	5.90E-15	-0.32	6.30E-21	-0.31	6.60E-07	YES	YES	-
rs12660365	rs13211753	6	6	0.15	0.07	dom	dos	SLC22A3	SLC22A2	4.30E-14	0.38	8.80E-19	0.191	1.60E-02	YES	YES	-
rs3106162	rs9364559	6	6	0.40	0.20	het	het	LPAL2	LPA	2.50E-11	-0.33	3.40E-18	-0.325	3.50E-06	YES	NO	-
rs9457931	rs662421	6	6	0.06	0.48	dom	dos	LPAL2	SLC22A2	7.70E-14	-0.3	4.90E-18	-0.152	1.40E-02	YES	YES	-
rs4708870	rs2619268	6	6	0.06	0.23	dom	dos	LPAL2	SLC22A2	1.50E-17	0.33	5.80E-18	0.347	4.70E-06	YES	YES	-
rs4708870	rs79570361	6	6	0.06	0.08	dom	dom	LPAL2	PLG	1.00E-35	0.46	1.20E-17	0.403	9.80E-05	NO	YES	-
rs1397168	rs56393506	6	6	0.14	0.17	dom	het	SLC22A3	intergenic (LPA)	8.30E-12	-0.33	1.60E-17	-0.249	3.90E-04	YES	NO	-
rs1998045	rs512077	6	6	0.16	0.15	dom	dos	intergenic (LPA)	SLC22A3	2.50E-13	0.27	3.40E-16	0.282	9.30E-07	YES	YES	-
rs1937475	rs1652507	6	6	0.32	0.17	dom	dos	ENSG00000224371	LPA	9.50E-12	-0.26	8.80E-16	-0.1	7.80E-02	YES	YES	-
rs3731239	rs3814960	9	9	0.37	0.38	het	dom	CDKN2A	CDKN2A	1.10E-14	-0.27	1.90E-15	-0.327	2.10E-06	-	-	YES
rs5117	rs429358	19	19	0.24	0.16	het	dos	APOC1	APOE	4.30E-12	0.29	3.70E-15	0.251	1.90E-03	-	-	-
rs4708870	rs2315065	6	6	0.06	0.10	dos	dos	LPAL2	Intergenic (PLG)	7.70E-12	-0.39	9.60E-13	-0.308	4.50E-03	YES	YES	-
rs3895825	rs4947818	13	7	0.21	0.20	dos	het	ENSG00000284196	ENSG00000228627	3.70E-11	0.21	2.70E-12	0.102	5.90E-02	-	-	-
rs4708870	rs1084651	6	6	0.06	0.17	dom	dom	LPAL2	Intergenic (LPA)	1.60E-17	0.37	3.30E-12	0.401	1.80E-05	NO		-

Table 2: Summary of the final epistastic SNP-pairs

This table shows summary results for the 22 cis-pairs which successfully passed the conditional analysis, and whose interaction doesn't tag rare variants or haplotypes. Each of the SNPs was annotated to the nearest gene. The indep.LPA column indicates whether the interactions were also independent of the complex pattern of 3 interacting variants previously reported at the LPA locus. The indep 6q26–q27 columns indicate whether the interactions are also independent of the marginal effect of 3 SNPs and one haplotype previously reported in the SLC22A3-LPAL2-LPA gene cluster. The indep CDKN2A/B columns indicate whether the interactions are also independent of the marginal effect of 2 SNPs previously reported at the CDKN2A/B locus. "–" indicates that the interacting pairs are located on another chromosome and were therefore not tested for independence towards this signal.

3.4.6 Conditional analysis for the LPA locus

Because a substantial part of these pairs is located in the vicinity of the LPA regions (the LPA, LPAL2, PLG, and SLCA22 family are all located one after another on chromosome 6) and we previously have reported a complex pattern of epistasis at the LPA locus involving three SNPs (see Section 5.3.3 and (Zeng et al., 2022)), we wanted to assess whether the signals found here were independent. To do so, we conducted the same kind of conditional analysis described above for the 17 cis-pairs located on chromosome 6, but conditioning on the three SNPs rs140570886, rs1652507, rs9458001, and all their 2- and 3-way interactions. We report that only 2 pairs, namely rs4708870: rs79570361 (annotated to LPAL2 and PLG respectively) and rs4708870: rs1084651 (annotated to LPAL2 and LPA respectively) did not pass this conditional analysis at the nominal threshold of 0.05. All others did (Table 3) and therefore represent, at least partially, independent signals from the previously described epistasis signal at the LPA locus.

snp1	snp2	pval_LRT
rs11751605	rs1084651	1.03E-11
rs2661839	rs2872819	2.71E-15
rs7742018	rs4252051	1.51E-12
rs12210186	rs1084651	1.38E-10
rs13211753	rs9457931	0.0061
rs2457571	rs2314851	1.08E-14
rs3125056	rs13198987	4.72E-12
rs12660365	rs13211753	0.0159
rs3106162	rs9364559	6.00E-14
rs9457931	rs662421	0.0124
rs4708870	rs2619268	0.0420
rs4708870	rs79570361	0.0906
rs1397168	rs56393506	7.36E-09
rs1998045	rs512077	8.78E-05
rs1937475	rs1652507	1.86E-06
rs4708870	rs2315065	2.41E-07
rs4708870	rs1084651	0.09330

Table 3: Results of the conditional analysis on the three previously published interacting LPA SNPs

The p-value was computed using a likelihood ratio test between a null model with the marginal effect of SNP1 and SNP2 as well as the marginal effect of rs140570886 , rs1652507, rs9458001, and all their 2- and 3-way interactions and a full model including the SNP1:SNP2 interaction as well. A significant p-value indicates that the SNP1:SNP2 interactions provided new information on the phenotype that was not carried by the 3 other SNPs and their interaction.

3.4.7 Conditional analysis for the SLC22A3-LPAL2-LPA gene cluster

Previous associations with CAD have been reported for a 4 SNP haplotype and for 3 single SNPs (rs3088442, rs3798220, and rs10455872) in the SLC22A3-LPAL2-LPA gene cluster We therefore also ran two conditioning analyses, one on these 3 SNPs and one on this haplotype for all of the cis-pairs located on the chromosome 6. Three of them, rs11751605:rs1084651, rs3106162:rs9364559, and rs1397168:rs56393506 showed dependence on the 3 previously reported SNPs and were not significant after conditioning on them (Table 4).

snp1	snp2	P-value LRT 3 SNPs	P-value LRT haplotype

rs11751605	rs1084651	0.0965	8.14E-19
rs2661839	rs2872819	4.64E-07	2.75E-17
rs7742018	rs4252051	2.22E-09	5.08E-13
rs12210186	rs1084651	0.0327	1.26E-21
rs13211753	rs9457931	0.0002	8.51E-15
rs2457571	rs2314851	9.03E-06	4.63E-16
rs3125056	rs13198987	0.0369	6.23E-15
rs12660365	rs13211753	0.0058	5.67E-11
rs3106162	rs9364559	0.6927	0.1461
rs9457931	rs662421	0.0308	1.75E-10
rs4708870	rs2619268	0.0070	1.47E-10
rs4708870	rs79570361	0.0002	1.67E-11
rs1397168	rs56393506	0.1383	3.16E-07
rs1998045	rs512077	3.07E-06	9.80E-11
rs1937475	rs1652507	1.99E-06	6.64E-17
rs4708870	rs2315065	1.68E-07	1.92E-12
rs4708870	rs1084651	1.67E-05	3.98E-15

Table 4: Results of the conditional analysis on the three previously associated SNPs and one haplotype at the SLC22A3-LPAL2-LPA gene cluster
The p-value was computed using a likelihood ratio test between a null model with the marginal effect of SNP1 and SNP2 as well as the marginal effect of either rs3088442,rs3798220, and rs10455872 or the four SNPs haplotype rs2048327- rs3127599- rs7767084-rs10755578 and a full model containing the SNP1:SNP2 interaction in addition. A significant p-value indicates that the SNP1:SNP2 interactions provided new information on the phenotype that was not carried by the 3 other SNPs or haplotypes.

3.4.8 Conditional analysis for the CDKN2A-CDKN2B region

Previous associations with CAD have been reported for 2 SNPs in the CDKN2A-CDKN2B region where three of our top hit pairs are located. We therefore also ran a conditioning analysis on these 2 SNPs for the three cis-pairs located on chromosome 9. All their interaction terms remained significant when conditioning on the 2 previously reported SNPs, and these interactions are therefore independent of the previously reported signal (Table 5).

snp1	snp2	pval_LRT
rs9632884	rs581876	0.002123201
rs7857118	rs62555370	0.014112126
rs3731239	rs3814960	1.16E-05

Table 5: Results of the conditional analysis on 2 previously associated at the CDKN2A/B locus
The p-value was computed using a likelihood ratio test between a null model with the marginal effect of SNP1 and SNP2 as well as the marginal effect of rs1333049 and rs4977574 and a full model containing the SNP1:SNP2 interaction in addition. A significant p-value indicates that the SNP1:SNP2 interactions provided new information on the phenotype that was not carried by the 2 other SNPs.

Finally, after all these conditional analyses we report 17 interacting SNP pairs (Table 3), which are independent of any single locus association and present the strongest argument that statistical epistasis testing can bring for biological epistasis. All these pairs except 5 are cis-pairs located on chromosome 6 with both SNPs associated (by position) to the following 4 loci: LPA, LPAL2, PLG, and the SLC22A family (A1, A2, A3). The three cis-pairs on chromosome 9 are respectively annotated to CDKN2B-AS1 (both SNPs) and CDKN2A (both SNPs). One more cis pair is located on chrosome 19 between the APO1C and APOE loci. The only trans pair is unfortunately located between two lncRNAs without known names or functions.

3.4.9 Characterization of the significant epistasis pairs

To better understand the biological basis of the epistatic pairs that we report through the search for statistical epistasis we attempted to further characterize the type of epistasis represented. Indeed, a significant interaction term in the logistic model is the sign of a deviation of the additivity of the two alleles on the log-linear scale and can be caused by different types of epistasis. For each pair, we, therefore, compared the marginal effects of the two SNPs with the effect of the interaction to classify the interaction between antagonist/synergist and negative/positive epistasis (Table 6). Antagonist interactions arise when the effect of the interaction is opposite as compared to the marginal effects, and synergist interactions are the opposite. Epistasis interactions are classified as negative if the carriers of the minor alleles at the two loci have higher disease risk as would be expected from the sum of the marginal effect of the two alleles and as positive if they have a lower risk.

SNP1	SNP2	OR 1	P-val1	OR 2	P-val2	OR int	Pval Int	Epi.factor	antagonist/synergistic	positive/negative
rs9632884	rs581876	-0.186	4.08E-12	-0.34	3.91E-49	0.38	1.75E-28	-0.49	antagonist	negative
rs2661839	rs2872819	0.206	6.22E-34	0.08	8.53E-05	-0.19	1.99E-24	-1.11	antagonist	positive
rs7742018	rs4252051	0.092	1.19E-11	0.33	3.54E-31	-0.23	2.03E-24	-0.41	antagonist	positive
rs12210186	rs1084651	0.268	1.98E-35	0.12	1.43E-06	-0.37	8.87E-24	-0.72	antagonist	positive
rs13211753	rs9457931	-0.049	0.063	0.01	0.857	0.54	1.32E-23	-0.09	NA	negative
rs2457571	rs2314851	0.134	1.34E-11	0.20	5.38E-25	-0.16	2.06E-22	-0.85	antagonist	positive
rs7857118	rs62555370	-0.093	4.35E-07	-0.38	8.77E-36	0.39	5.89E-22	-0.24	antagonist	negative
rs3125056	rs13198987	0.236	1.21E-29	0.08	9.36E-05	-0.32	6.33E-21	-0.75	antagonist	positive
rs12660365	rs13211753	0.004	0.834	-0.05	0.085	0.38	8.79E-19	0.01	NA	negative
rs9457931	rs662421	0.489	4.12E-31	0.04	0.002	-0.30	4.88E-18	-1.62	antagonist	positive
rs4708870	rs2619268	-0.004	0.898	-0.02	0.269	0.33	5.76E-18	-0.01	NA	negative
rs1998045	rs512077	-0.004	0.848	-0.02	0.403	0.27	3.44E-16	-0.02	NA	negative
rs1937475	rs1652507	0.099	2.25E-07	0.15	2.03E-09	-0.26	8.76E-16	-0.38	antagonist	positive
rs3731239	rs3814960	0.111	2.40E-05	0.15	8.18E-11	-0.27	1.92E-15	-0.42	antagonist	positive
rs5117	rs429358	-0.210	7.31E-15	0.05	2.34E-02	0.29	3.72E-15	-0.71	antagonist	negative
rs4708870	rs2315065	0.259	3.08E-24	0.34	6.72E-73	-0.39	9.65E-13	-0.66	antagonist	positive
rs3895825	rs4947818	-0.059	5.86E-04	-0.12	3.01E-08	0.21	2.74E-12	-0.29	antagonist	negative

Table 6: Characterisation of the significant epistasis pair

This table shows the marginal effects (log Odds Ratios) of the two SNPs and their p-value along with their interaction effect of these p-value for each of the 17 significant cis-pairs. The epistatis factor can be computed as OR1/OR_int or OR2/OR_int respectively and indicates whether the epistasis is considered antagonist or synergic. A positive epistasis factor indicates synergic epistasis and a negative epistasis factor indicates antagonist epistasis. This factor was not computed for pairs whose SNPs both do not have significant marginal effect. The positive/negative column indicates whether the effect of the interaction is more positive or negative (with respect to the fitness) than would be expected from the sum of the two alleles. The effect sizes are reports as log(Odds Ratio) with respect to the minor allele. A negative log (Odds Ratio) shows protective effect (positive effect) of the minor allele and a positive log(Odds Ratio) a higher disease risk (negative effect).

Each of the two SNPs can have a marginal protective or deleterious effect or no effect and all interactions have a significant effect (this is what we looked for). There are therefore 10 possible situations as described in Table 7 which can be further classified as negative/positive and synergetic/antagonist. We report that 13 of the 17 interactions we found involve SNPs which both have significant ($p < 0.05$) marginal effects. All of them show a sign change between the two marginal effects and the interaction effect and are therefore antagonists. Nine are positive epistasis interactions and four are negative. In addition, we report 4 interactions between variants that do not have a significant marginal effect. All these pairs are classified as negative epistasis as the effect of the interaction is deleterious (OR > 1). Finally, one pair involves one variant with marginal effect and one without, with a deleterious interaction.

SNP1	SNP2	interaction	N	antagonist/synergistic	positive/negative
NULL	NULL	protective	0	NA	positive
NULL	NULL	deleterious	4	NA	negative
NULL	protective	deleterious	0	antagonist	negative
NULL	deleterious	deleterious	0	synergistic	negative
NULL	protective	protective	0	synergistic	positive
NULL	deleterious	protective	0	synergistic	positive
protective	protective	deleterious	4	antagnoist	negative
deleterious	deleterious	protective	9	antagonist	positive
deleterious	deleterious	deleterious	0	synergistic	negative
protective	protective	protective	0	synergistic	positive

Table 7: Summary of the different types of epistasis pairs detected
This table summarizes the 10 possible combinations of protective/deleterious or absent marginal effect of the two SNPs with the protective/deleterious effect of their interaction. The N column reports how many interactions of this type were present among the 17 final interactions that we report. The antagonist/synergic column indicates whether the interaction effect contributes to disease risk in the same “direction” as the main effects of the two SNPs. The positive/negative column indicates whether the effect of the interaction is more positive or negative (with respect to the fitness) than would be expected from the sum of the two alleles.

3.5 Discussion

In this study, we performed the first, to our knowledge, genome-wide SNP-SNP interaction study for coronary artery diseases. We report 22 SNP pairs whose interactions are significantly associated with CAD in the UKBB dataset and replicated in a meta-analysis of 10 independent CAD studies. These 22 pairs were moreover subjected to conditional analysis on every variant in the non-QCed UKBB dataset, one of the imputed datasets with the largest number of SNPs up to date, to ensure that none of the rarer SNPs included in this dataset could explain alone the interaction effect. From these 22 pairs, 17 were independent of any single marginal effect locus in their direct vicinity. With this, we argue to these 17 SNP pairs represent the strongest level of statistical evidence for epistasis available at this time.

The maybe more important contribution of this study is to show that such a genome-wide hypothesis-free epistasis study is possible using an exhaustive enumeration of all possible SNP pairs. Indeed, most of the interaction association studies, all traits together, were conducted using some sort of a-priori SNP selection. And most of the few truly genome-wide interactions studies were conducted almost a decade ago using sample sizes smaller than 10’000 (Combarros et al., 2009; Gyenesei, Moody, Semple, Haley, & Wei, 2012; Y. Liu et al., 2011; H. Wang et al., 2020). Showing that the two-stages approach that we propose allows to conduct an exhaustive search in a reasonable time using computational resources nowadays common, is therefore important to incite researchers to undertake more of these studies. Very interestingly, another GWIAs using a very similar two-stage approach on a big dataset of 445’221 people, yet with a different algorithm for the first screening, was published earlier this year and uncovered new genetic loci for lung cancer (R. Zhang et al., 2022). This strengthens the evidence that two-stages approaches are promising methods for such epistasis scans.

Very interestingly all of the SNPs involved in the 17 epistasis top hits (and annotated to a known gene) are annotated to a total of 10 genes (LPAL2, PLG, SLC22A1, SLC22A2, SCL22A3, LPA, CDKN2B-AS1, CDKN2A, APO1C, and APOE), which are all known risk factors for CAD.

3.5.1 Interactions in the 6q26–q27 region

In particular, the SLC22A3-LPAL2-LPA gene cluster has been associated with cardiovascular disease in a study by Trégouët and colleagues (Trégouët et al., 2009). In this study, they show that one haplotype derived by the genotype at the four SNPs rs2048327 in the SLC22A3 gene, rs3127599 in the LPAL2 gene, and rs7767084 and rs10755578 in the LPA gene is associated with elevated risks of CAD. They moreover show that this haplotype is exerting its effect on CAD risk by elevating the levels of Lp(a), the apolipoprotein-a encoded in the LPA gene. As Trégouët and colleagues say it is uncertain "whether the detected haplotype effects are the reflection of interactions between SNPs at the haplotypic level or whether they are tagging ungenotyped functional variants." Pursuing the latter hypothesis, two studies suggested that the apparent haplotype effect could be explained by single causal variants, rs3088442 for Wang (L. Wang et al., 2016) and colleagues and rs3798220 and rs10455872 for Koch and colleagues (Koch et al., 2013). We, therefore, tried to condition all the interactions on these 3 SNPs to check whether they are independent signals. It appeared that for three of the pairs that we report, the significance of their interaction term is lost when conditioning on these 3 SNPs. For all other pairs, the conditioning did not decrease the statistical significance of the interaction terms under the nominal threshold. We, therefore, argue that these 17 interactions are at least partially independent of this haplotype effect.

In a previous interaction association study with an a-priori statistical filter that we published recently (Zeng et al., 2022) and which we describe in Section 5, we reported another complex interaction pattern between 3 SNPs at the LPA locus associated as well with CAD and Lp(a) levels. In the original 23 SNP-pairs that passed the region-based conditional analysis, only two were indeed not independent of this 3-SNPs interaction at the LPA locus. This 3 SNP-interaction pattern was also subjected to a similar region-based conditional analysis in the original study (see Section 5.3.3) and a dependence on the four variants describing the haplotype reported by Tregouet or the three possible causal variants reported by Wang and Koch was not detected. This suggests that our 17 interacting SNP pairs, as well as the 3 interacting LPA SNPs and the haplotype and variants reported by other studies might all represent different signals in the 6q26–q27 region.

However, more advanced analyses are needed to understand if and how many distinct genetic interactions are happening or if marginal effects of rare variants are at play. This is a very difficult task with the micro-array data currently at hand because there are a lot of "blind" spots, namely ungenotyped variants in this area. Moreover, most of the published variants of interest might only be tags for causal variants that were not genotyped, absent from the imputation panel, or excluded in the QC or in the pruning. This makes the task of comparing results from these different studies (which underwent different QC and SNP pruning procedures) and estimating the number of independent signals very hard. Resolving the actual genetic basis of these association signals would certainly require deep sequencing of this candidate region. Only in this way can the whole complexity be modeled using all the variants in these loci. Our previous results on the LPA region showed that three-way interaction might also be responsible for part of the association signal, and we, therefore, suggest that further analysis should also look at higher-order interactions in this key region.

Indeed, a better understanding of these maybe complex genetic interactions is of great clinical interest for the treatment of CAD and risk prevention. The Lp(a) concentration is a routinely used measurement that is taken into account by clinicians to infer the risk of CAD (Kronenberg, 2019) and individuals with abnormally high Lp(a) levels can have up to 2.5 times higher risk to develop CAD (Kamstrup, Tybjærg-hansen, Steffensen, & Nordestgaard, 2009). Consequently, pharmacological control of Lp(a) levels is currently being thought as a promising therapy for CAD (Tsimikas et al., 2020). In this context, a better understanding of the genetic basis and molecular mechanisms governing the Lp(a) levels and its association with CAD could pave the way for the development of new therapeutical and diagnostic tools.

In addition to this three gene locus, we also report interactions between SNPs annotated to LPA2, SLC22A2, SLC22A3, and SNPs associated to PLG, another gene in this chromosomal region. This, therefore, suggests, that the PLG gene should also be included in the more detailed analysis of this region.

3.5.2 Interactions at the CDKN2A-CDKN2B region

The three of the 22 hits pairs that are located on chromosome 9 are all in the CDKN2A-CDKN2B region which has been previously associated with CAD. Indeed, two pairs are between SNPs associated with the CDNK2B antisense RNA 1 which was shown to be associated with CAD in multiple GWAS and candidate SNPs studies (McPherson et al., 2007; Nilesh et al., 2007). CDNK2B-AS1 is thought to influence the risk of CAD by modifying endothelial cell migration and the transit of monocytes through these endothelial cells (Cho et al., 2019). The third interaction takes place between two SNPs annotated to CDKN2A, which is located right next to the CDNK2B-AS1 gene. Two SNPs, rs1333049 and rs4977574 have been reported in several studies to mediate the association of CDNK2B-AS1 with CAD in different population (K. Huang et al., 2019; Tibaut, Naji, & Petrovič, 2022; Yuan et al., 2020). To assess whether the interactions signal that we report here are independent of those marginal single-locus effects, we tried to condition the three interactions on the effect of these two SNPs in the logistic regression model. We report that the three interaction terms passed the conditional analysis and represent independent signals from the already known variants. As for the interactions on chromosome 6 however, we cannot exclude that other non-accessible variants might be responsible for the interaction signal or the signal represented by the previously published SNPs or that there are more complex patterns of interaction at play at this locus.

3.5.3 Interactions at other loci

In addition to the multiple interactions on chromosome 6 and 9 we report one cis-interacting pair on chromosome 19 and a single trans-epistasis pair between chromosomes 13 and 7.

The two interacting SNPs on chromosome 19 are respectively annotated to APO1C and APOE. The APOE gene has 3 different alleles which produce 3 isoforms of the Apolipoprotein E (apo E). ApoE2 has been associated with hyperlipoproteinemia and ApoE4 with atherosclerosis (Mahley, 2016). The APOC1 locus is located right next to the APOE locus on chromosome 19 and encode the apolipoprotein C1, an apolipoprotein involved in lipid transport. Very interestingly, ApoC1 inhibits the ApoE mediated binding of lipoproteins on their receptor (Fuior & Gafencu, 2019). The deleterious interaction that we report between these two loci could therefore be indicative of a perturbation of this molecular mechanism leading to increased plasma lipid levels. Further validation of this hypothesis could be attempted in vitro and potentially bring more understanding of the precise role of ApoC1 which is currently still underexplored.

The two SNPs involved in the only significant trans-pair are unfortunately annotated to two unknown lncRNA. At this stage it is therefore impossible to make any hypothesis about their possible association with CAD or their putative function.

3.5.4 Validation of cis-epistasis pairs

In an often cited article, Fish and colleagues claimed that most of the reported associations between cis-epistatic interactions and gene expression were statistical artifacts (Fish, Capra, & Bush, 2016). They reported that most of the interaction could be explained more parsimoniously by confounders (such as population structure) or by highly complex LD patterns with other causal variants. To avoid producing such artifacts we paid special attention to their and others' recommendation. Concerning the confounding due to cryptic population structure, we firstly conducted our analysis on a dataset formed only of individuals with European white ancestry. While this greatly limits the generalizability of our findings, this limits the possibility of confounding through population stratification. We believe that this is especially needed, as the frequencies of some lead genetic markers for CAD have been shown to be different between populations (Enkhmaa et al., 2016). In addition, following one of the suggested methods in (Gusareva & Van Steen, 2014) we included the first ten MDS components of the genetic variance in our fine-mapping and validation logistic regression model to account for any remaining large-scale genetic structure in the population. Finally, we included age and sex as covariates in our models, as these are two important risk factors for CAD. It is worth noting that there is debate about the potential negative effect of the inclusion of non-confounding covariates in logistic models (Xing & Xing, 2010). We, therefore, refrained to include more covariates.

In order to answer Fish's second concern about cis-epistasis, we conducted several large conditional analyses to make sure that the interactions that we report are not tagging other additively causal variants or haplotypes. As further proof that this step is indeed needed, 14 of the 36 replicated cis-pairs lost their statistical significance when conditioned on all the variants within a ±200kb region. For three other regions for which association with either single effect locus, interactions, or haplotype had been previously reported, we decided to conduct a more complex conditional analysis by conditioning on the joint signal of these known loci. By doing this, we discovered that 5 more of our SNP-SNP interactions were actually not independent from these known signals. We therefore cannot stress enough the importance of validating cis-epistasis pair by ruling out that they tag known additive effects. We however have to acknowledge that this proof of the independence of the interaction signal is only as good as the data allows. Indeed, one would need to conduct such analysis for every base pair in the region to ensure that no rare or untyped variant is at the origin of the signal. Hopefully, the advent of NGS might allow these kinds of analyses in the near future.

3.5.5 Implications for the Genetic Architecture of Epistasis in Common Diseases

Very interestingly, although we conducted a hypothesis-free and exhaustive search for epistasis we do report only interactions between loci which had been previously associated with CAD in single-locus associations analysis. This result could support several different hypotheses about the genetic architecture of common diseases but also has practical implications for future epistasis studies in CAD and maybe other diseases.

If, indeed, epistatic interactions arise only between SNPs which also have a marginal effect associated with the disease, then epistasis scans based on previous evidence from GWAS should be more powerful than hypothesis-free GWIAs. This would, in addition, be the case independently of whether these marginal effects are indeed truly stand-alone effects of the variant or if the apparent additive effect is due to some kind of non-additive and perhaps actually epistatic variance. In contradiction with this argument, a previous study of this kind, which investigated interactions between SNPs in the direct vicinity of all known CAD GWAS hits (Zeng et al., 2022) and which we present in Section 5, did not find the SNPs pairs that we found here. This study had however a much smaller sample size and, correspondingly, lower power. We suggest that conducting this marginal-effect-based filtered study on the UKBB dataset could certainly help draw stronger conclusion on this apparent discrepancy.

Although these results seem to speak in favor of the fact that genetic interactions would arise only or principally between genes that are associated with the disease in an additive way, we should refrain from making a too strong conclusion. Indeed, we should remember that also our study had quite low power to detect interaction effects with an effect size smaller than 1.2. Therefore, our results would be totally compatible with small-to-medium effect interactions happening between genes that are not associated with CAD with their marginal effect and which we did not have the power to discover here.

In addition, we noticed that although the interactions happened between genes that had been previously associated with the disease, they did not happen between the exact genetic variants which had been associated. This could be due to differences in the SNPs available in the different datasets as well as QC and imputation procedures. Alternatively, different SNPs, with possibly different biological consequences, in the same gene could have different modes of action. For example, a variant in the protein-coding region of a gene might have an effect on the disease on its own by changing the structure of the protein, whereas the interaction of two other variants located in the promoter and enhancer of this same gene could impact the expression of the protein. We hypothesize that a better mapping of the exact causal variants with the advent of bigger imputation panels and next-generation sequencing as well as a better understanding and prediction ability for the functional consequences of single nucleotide substitutions will help provide part of the answer to this question in the next future.

It is striking that all of the final interaction pairs that we report, except one, are situated on the same chromosome (cis-epistasis). Indeed, only one pair situated on different chromosomes (trans-pairs) did replicate in our replication dataset. It could therefore be tempting to conclude that epistasis happens preferentially between SNPs on the same chromosome. As most of the SNPs in the genome are in non-coding regions and therefore exert their effect by modulating gene expression via their interaction with transcription factors, enhancers, and other players of the transcription machinery, two SNPs located on the same chromosome are possibly more likely to modulate these processes by their interaction. However, a recent large-scale study aiming at investigating the properties of epistatic interactions across more than 200 diseases reported that the overwhelming majority of epistasis was located in trans (Chatelain et al., 2021). In the light of this finding, it is possible that we found predominantly cis-epistasis in this particular study because of the genetic architecture of CAD. Indeed, several of the strongest risk factors for CAD, such as the LPA, LPAL2, and SLC22A3 genes happen to be located on the same chromosome, namely the sixth.

The multiple definitions of epistasis (see Section 1.2.1) cover different kinds of interaction which will appear as deviation from the additivity in the logistic model. In particular, one can distinguish between antagonist and synergist epistasis as well as between negative and positive epistasis. Some studies have reported only positive epistasis while some other have reported only negative epistasis or both and it is therefore unclear if one type is more important than the others. These different types of epistasis have also been proposed to be driven by different evolutionary forces and underlain by different molecular mechanisms (de Visser, Cooper, & Elena, 2011). In order to further characterize the 17 significant interactions that we report , we looked at the combinations formed by the two marginal effect of their SNPs (no marginal effect, protective, deleterious) and the effect of the interaction. Very interestingly, out of the 10 possible combinations, all pairs fall within three categories.

More precisely, in 13 pairs the two marginal effects are significant but the interaction has the opposite effect. These interactions belong to the antagonist class. This result is aligned with the findings of Chatelain and colleagues (Chatelain et al., 2021) who described a very strong over-representation of antagonist epistasis in their UKBB atlas of epistasis. Mouse studies have also shown antagonist interaction to be dominant in interaction networks influencing bone and body composition (Tyler, Donahue, Churchill, & Carter, 2016). In antagonist epistasis, the interaction effect tends to counter-balance the marginal effects. Such interactions tend to stabilize the phenotypes towards the mean and reduce extreme phenotypes.

Among these 13 antagonist pairs, 9 are examples of positive epistasis where the effect of the interaction is beneficial even if the marginal effects are deleterious. From a mechanistic point of view, the two SNPs could have opposite but detrimental effects on the same biological molecule, for example, a protein or enzyme. In this example, SNP1 would increase the protein function with a deleterious effect in a neutral genetic background. SNP2 would decrease the function of this same protein, also with a deleterious effect in a neutral background. However, when present together, they compensate each other leading to the appropriate protein function and therefore having a positive effect. Rauscher and colleagues have recently discovered an example of such positive epistasis impacting the function of CFTR in human. In their study, they show a synonymous SNP which can alter translation speed, thereby increasing domain-domain interaction time and rescuing the impaired protein function caused by another missense mutation causing SNP (Rauscher et al., 2021). Similar mechanisms could also be happening between different proteins of the same pathway, by up-regulating one step of the pathway and down-regulating the second one.

The other 4 antagonist pairs are classified as negative epistasis, as the marginal effect of each SNP is protective against CAD but their interaction is detrimental. Although we could not find well-documented examples of molecular mechanisms of such interactions, negative epistasis has been proposed to be associated with genetic robustness but also with evolvability (de Visser et al., 2011) and to be required to explain the origin and maintenance of sexual reproduction under the mutational deterministic hypothesis (Azevedo, Lohaus, Srinivasan, Dang, & Burch, 2006).

The third category of epistasis represented in this study, by 4 pairs out of the 17 significant ones, shows no significant marginal effect of the two SNPs but a detrimental effect of their interaction. While this is also an example of negative epistasis, it seems to imply different mechanistic underpinnings. These two SNPs only have a detrimental effect when present together in the genome. Such kind of interactions could actually indicate a variation of what is called compound heterozygosity (CH). CH is defined as the presence of two different mutant alleles at a particular locus, each one on one of the two homologous chromosomes. Generalized CH, is a relaxed form of CH in which the genetic variants are not necessarily coding, rare, and deleterious and is likely involved in a wide range of human polygenic traits (Zhong et al., 2017). The epistasis described above could be explained by two different single nucleotide substitutions in the same gene located each one on one chromosome, and which jointly have a deleterious effect but not separately. Mechanistically, each of the two SNPs could lead to a protein loss of function. (via a different mechanism), which can be compensated when one copy of the gene is functional.

In an even more generalized form of CH, the two SNPs could be in two different genes involved in a same pathway and located on the same chromosome. Some level of robustness or redundancy in the pathway may allow for one mutation and only when the two mutated alleles are present , each on a different exemplar of the chromosome, is the pathway finally affected. Such mechanism could be at play in the SNPs pairs of this category. Indeed, they are all annotated to the LPAL2, PLG genes, and SLC22A1/2/3 gene family which have all been associated with Lp(a) levels (Ronald et al., 2011; L. Wang et al., 2016) and therefore influence the same pathway in the context of CAD.

Very interestingly, Penman and colleagues have reported that such epistatic interactions happen between mutations in the alpha and beta globulin (which assemble together to form the hemoglobulin) to create compound thalassemic genotypes which are associated with thalassemia risk and malaria protection (Penman, Pybus, Weatherall, & Gupta, 2009).

These types of interactions would be compatible with two different mutated alleles being on a single chromosome or on two homologous chromosomes. Indeed, in the absence of phased data, we cannot know if the interacting minor alleles are present on the same or different chromosomes.

In conclusion, we found that three types of epistasis interactions were over-represented in the statistically significant epistatic pairs that we report. These three types of interactions seem to indicate the presence of different types of epistasis possibly underlain by different molecular mechanisms and evolutionary driving forces. Deeper functional and molecular characterization of the involved variants would be needed to test these hypotheses.

3.5.6 Limitations

Despite our effort to use robust methods, we need to acknowledge several technical limitations that need to be kept in mind when interpreting the results.

Firstly, despite being one of the first genome-wide interaction association studies using one of the new "modern" genetic datasets with several hundred thousand individuals and therefore having substantially more power than most of such previous studies, the power of our study was not optimal for all effect sizes. Interestingly, for an effect size such as the one of the 17 significant pairs, which approximately lie in the range between exp(0.2)=1.22 and exp(0.4)=1.4 our power was close to 25% for a SNPs with MAF of 5% and close to 100% for SNPs with a MAF of 1%. However, our power was close to null for OR lower than 1.15. Generally, if we can conclude that our study brings evidence towards the fact that there are no epistatic interactions with OR greater than 2 or smaller than 0.5 between SNPs more frequent than 5%, as we did not find any and had a power of 100% in this regime, we can say very little about the presence or absence of interaction with lower effect or MAFs.

One methodological point that contributed to decreasing the power of the present study is the probably sub-optimal multiple testing correction method that we used. Indeed, we used thresholds based on Bonferroni correction computed on the number of combinations between LD-independent SNPs that we were testing. We relaxed these thresholds in the case of the episcan in the hope of capturing causal SNP pairs in moderate LD with the tag pairs included in the analysis. And we also used a relaxed version of this correction in the validation-fine-mapping step (that we call suggestive threshold, computed as 1/number of tests) to somewhat counterbalance the over-conservative nature of Bonferroni correction in this setup. Indeed, although the SNPs were LD-pruned to an r^2 threshold of 0.5, some collinearities remain between them. The number of effective tests was, therefore, smaller than adjusted for in the Bonferroni correction. Moreover, we computed the Bonferroni correction on the number of LD-independent SNPs entering the first step of our analysis. A better correction should take into account the two-step nature of our analysis strategy. Although different multiple testing correction strategies have been proposed that could be better suited for two-step interaction association analysis, we do believe that their properties have still not been validated enough and decided to use the well-established Bonferroni correction. In the future, we should, however, consider alternatives that might offer gain in power and more importantly in specific power, or the probability to discover causal pairs but no others.

Replication in an independent dataset is a crucial step to validate the results of genetic interaction studies. Unfortunately, we were unable to attempt validation of 32 pairs, among them 7 trans-pairs, because they were not present in our replication dataset and did not have any good proxy in this dataset either. This is a strong limitation of our study, as these pairs are potentially true epistatic signals but will certainly never be validated. Indeed, finding replication data is often a long and complicated process, and will probably not be attempted to validate these SNPs. The separate imputation with a new imputation panel and subsequent merging of the genetic data of 10 different studies is also a very effort- and time-consuming process which we could not undertake.

Finally, as already mentioned previously, data availability for rarer and ultimately all variants in the vicinity of the interacting pairs is crucial to rule out tagging of single locus additive effects. In this study, we did this conditional analysis as rigorously as possible, by including all SNPs available, also the ones which would fail QC. We moreover, ran special analyses with the variants located in the vicinity of the interacting SNPs which had been previously reported to be associated with CAD. However, we ultimately cannot exclude that some of the effects of these interactions are better explained by another variant to which we did not have access.

4 Prior knowledge filtered interaction study

4.1 Introduction

In parallel to the hypothesis-free genome-wide interaction study, we also conducted several "knowledge-based" approaches in which prior biological knowledge is used to filter the SNPs before the association testing to reduce the search space.

In this study in particular we used two different approaches relying on different sources of knowledge. The first approach aimed a testing only SNPs that have a regulatory function, or in other words which influence the transcription of some gene. Such approaches have already been used in GWAS for example (Arloth et al., 2020), restricting the search space to variants that were associated with predicted chromatin features using DeepSea (Zhou & Troyanskaya, 2015). Taking advantage of the progress of deep learning models the authors of Deepsea recently released a new tool called ExPecto which can directly predict if a SNP is likely to influence the expression of any genes (Zhou et al., 2018). Our idea was therefore to restrict the search space to any combination of two SNPs that are predicted by ExPecto to influence the expression of at least one gene.

In the second approach, we wanted to restrict the search space to SNPs that are associated with genes for which there is molecular evidence of interactions. This would include for example genes that are involved in the same biological process or pathways but also genes whose proteins have been shown to interact in protein-protein interaction databases. To generate this list of SNPs we make use of the Biofilter software developed by the Ritchie lab (Bush et al., 2008), which scrutinize several sources of information to produce a list of candidates SNPs pairs that are likely to interact given this biological knowledge.

4.2 Regulatory SNPs filter

4.2.1 Methods

This "filter-based" analysis is a companion study to the genome-wide interactions scan presented in Section 3. It was conducted with the same dataset, using the same QC and the same methodology. The only difference is that the search space was narrowed down before step 1 using one of the filters that we present in this chapter.

4.2.1.1 definition of different regulatory SNPs filters

The available ExPecto framework allows to predict the effect of DNA sequence, and hence SNPs, on the expression of any genes in multiple tissues. In addition to the trained deep learning model allowing to predict the effect of new sequences, the authors made available a list of all single nucleotide substitutions within 1kb to the representative TSS of a gene and all 1000 Genomes variants that passed a minimum predicted effect threshold (>0.3 log fold-change in any tissue). We, therefore, decided to use this list of SNPs to filter the SNPs before running episcan.

Because it appeared that the overlap between this list and our discovery dataset (in its QCed but not-pruned version) was extremely low, we used the trained ExPecto deep learning model to predict the gene expression effect of all the 6'388'394 bi-allelic QCed SNPs of our dataset.

As a sensitivity test for these filters, we generated for both approaches the following filters:

- A cardiac filter: with SNPs which had a predicted effect in cardiac tissues
- A random filter: with SNPs which had a predicted effect in randomly selected non-cardiac tissues but not in cardiac tissues.
- An all-tissue filter: with SNPs which had a predicted either in cardiac or in randomly selected non-cardiac tissues.

In addition, using the SNPs list created by predicting de novo on our dataset we also designed the following filter:

- A cardiac-only filter: with SNPs which had a different predicted effect in the cardiac tissues as compared to the other tissues. For this filter, I computed the mean of the predicted effect in the cardiac and non-cardiac tissues. I then tested for their difference using a t-test. SNPs whose effect was significantly different in cardiac tissues after Bonferroni correction with a means difference greater than 0.1 were kept in this filter.

The list of cardiac tissues was as follows:

Artery Aorta, Heart Atrial Appendage, Right Atrium, Aortic Smooth Muscle Cell, Cardiac Ventricle Fibroblast, Endometrial Microvascular Endothelial Cell, Heart, Regular Cardiac Myocyte, Smooth Muscle Cell of The Pulmonary Artery, Thoracic Aorta Endothelial Cell, Vein Endothelial Cell, Artery Coronary, Heart Left Ventricle, Aorta, Left Ventricle, Cardiac Atrium Fibroblast, Endothelial Cell of Coronary Artery, Fibroblast of the Aortic Adventitia, Pulmonary Artery Endothelial Cell, Smooth Muscle Cell of The Coronary Artery.

4.2.1.2 Testing the filters using the data generated in the hypothesis-free episcan

To be able to compare these different filters without having to run a full epistasis analysis for each of them, we wanted to make use of the results computed for the genome-wide interaction analysis described in Section 3. More precisely we aimed at skipping the first statistical filtering step with episcan by looking up the episcan p-value for the pairs selected in our filters. However, the episcan step in Section 3 was performed on the LD-pruned dataset. The SNPs selected in the filter here were therefore possibly not directly included in the episcan as they might have been pruned. We, therefore, used plink to identify all LD-tags (at a threshold of 0.5 and with a window of 500 kb) of all SNPs included in each of the filters. For each combination of SNPs within the filter, we then looked up if any combination of their respective LD tags was present in the genome-wide episcan step and recorded their p-value.

4.2.1.3 Testing the original ExPecto Cardiac filter by running a new scan

To confirm the inexact results obtained by the method explained above, which is plagued by incomplete LD between SNPs in the filter and their tags present in the genome-wide episcan results, we ran the association analysis for the original ExPecto cardiac filter from scratch. We used the exact same methodology for step 1, step 2, and replication as described in Section 3 for the genome-wide scan, with the difference that the dataset was limited to the pairs between the SNPs from this filter.

4.2.2 Results

As the first step in this epistasis study between predicted regulatory SNPs, we aimed to identify SNPs with a gene regulatory function using the ExPecto framework (Zhou et al., 2018). The ExPecto framework uses a deep learning model which can predict changes in gene expression caused by changes in a DNA sequence. It is, therefore, able to predict the effect on gene expression of the two alleles of a SNP. The easier approach to identifying regulatory SNPs for our filter is to make use of a list of regulatory SNPs published with the ExPecto framework. This list was generated by the authors of ExPecto by predicting gene expression for all SNPs in the vicinity of a gene and all SNPs for the 1000 Genome Project. However, unfortunately, our dataset was imputed using the HRC + UKBiobank imputation panel and had very little overlap with this list. Indeed, only 5877 SNPs with a regulatory effect were present in our dataset. If we were considering only SNPs for which the regulatory effect was present in a list of cardiac tissues (see Methods) this number dropped to 2510 (Figure 33). The filters derived from this list are referred to hereafter as the “original” filters. Although this number of SNPs would allow for high power in the study, it would greatly impair our ability to detect new biology. Indeed, there would not even be one SNP per gene in the genome.

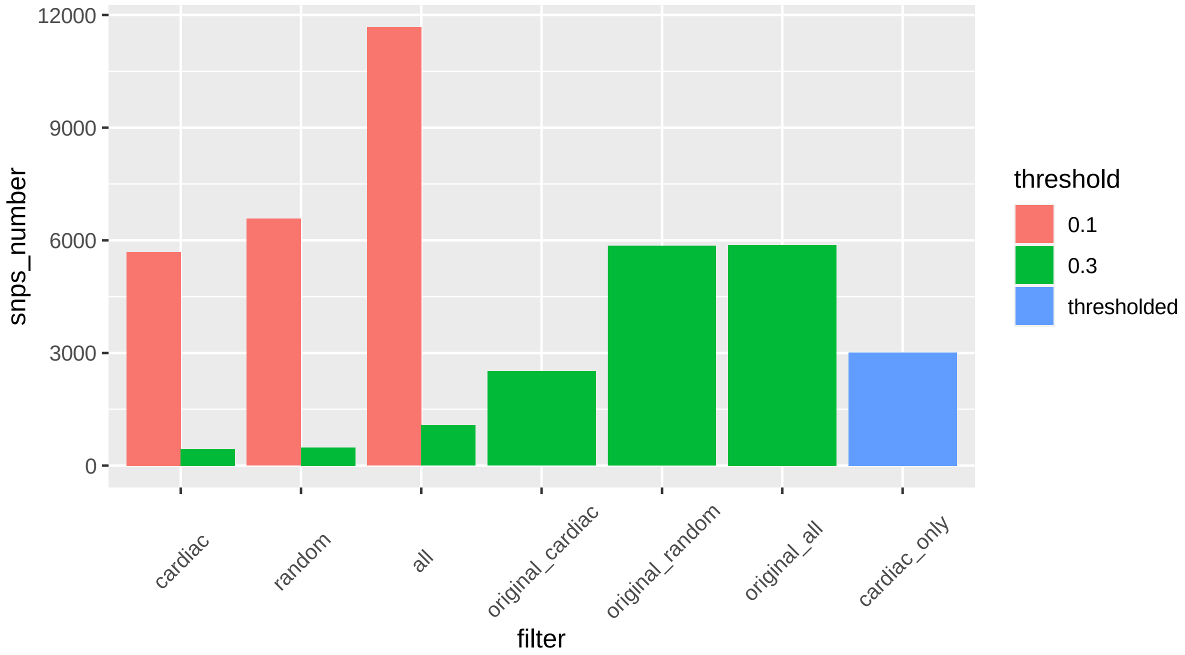


Figure 33: Number of SNPs in the different filters created with the ExPecto model
The barplot shows the number of SNPs included in each of the filters that we generated using the ExPecto Results. The „original“ filters were composed using pre-computed SNPs effect on gene expression published by the ExPecto team using a log fold-change threshold of 0.3. The other filters were computed by using the trained ExPecto model to predict the effect on gene expression of every SNP in our UKBB dataset. The color represents the two different thresholds applied on the log fold-change of gene expression when creating the filters.

We, therefore, decided to use the trained ExPecto deep learning model to predict *ab initio* the effect of each of the SNPs in our dataset. Very disappointingly, this approach yielded very few SNPs with a predicted effect on gene expression. Indeed, setting the threshold at 0.3 of the log fold change (as was done originally in the list the ExPecto team generated) yielded 1074 SNPs with an effect in any tissue and 444 with an effect in cardiac tissues (Figure 33). Lowering the threshold to 0.1 increased the number of SNPs to 11'678 and 5689 in all tissues and cardiac tissues, respectively.

Using these two lists of predicted regulatory SNPs, we also generated filters with SNPs having an effect in a list of randomly selected non-cardiac tissues, as a sensitivity test. Indeed, one would expect these filters to produce worse results than the cardiac one, if the filtering is meaningful. We finally generated a last filter with SNPs whose effect on cardiac tissues was significantly different in cardiac tissues as compared to non-cardiac tissues, as identified with t-tests. This filter contained 3012 SNPs (Figure 33).

4.2.2.1 Testing the filters using the data generated in the hypothesis-free episcan

The first aim of this analysis was to identify which of the generated filters had the best potential to discover interacting pairs. Because running a full analysis, with the first episcan step and the fine mapping on each of the filters would have been too time-consuming, we decided to use the results computed in the genome-wide episcan step (Section 3) for a first estimation. Basically, we used the LD structure to identify tag pairs in the genome-wide episcan results for each of the pairs generated by the filter. Looking at the p-value distribution of these tag pairs for each of the filters, however, unfortunately, reveals that none of the filters shows enrichment in low p-values (Figure 34). On the contrary, except for the two filters that we generated by re-predicting the SNPs effect and using the low threshold of 0.1 (“cardiac 0.1” and “random 0.1”) the number of pairs on the plot is very low. This means that their LD-tags had a p-value higher than 1×10^{-6} which was therefore not even recorded during the genome-wide episcan. Even for the “cardiac 0.1” and “random 0.1”, we do observe very few SNPs with p-value lower than 1×10^{-7}, and none lower than 6.13×10^{-8} (1/(5711*5710)/2) or 4.58×10^{-8} (1/(6602*6601)/2) which would be the suggestive significance threshold for these two filters respectively.

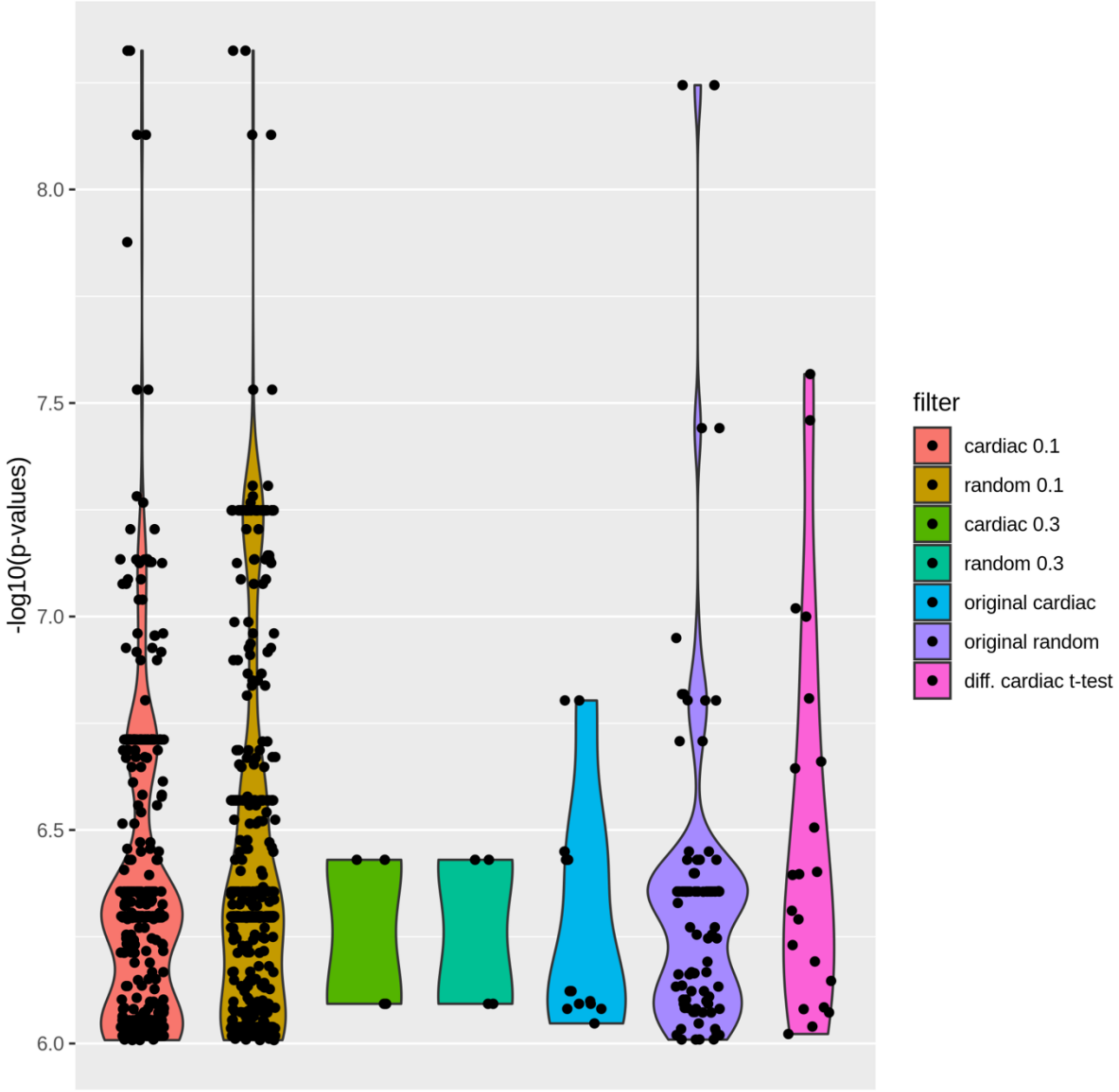


Figure 34: P-value distribution for the SNPs pairs in the different ExPecto filters as estimated based on LD from the genome-wide association scan.

There are two competing explanations for these results, which have different consequences for our approach. The first one is that there are indeed no strong interactions between any of the SNPs in the filter. Either because epistasis does not happen preferentially between SNPs with a regulatory role, or because the ExPecto framework was not able to reliably predict these functions for our dataset.

The second one is that there were no good LD tags for most of the SNPs in the filter and that the incomplete LD with the LD tags was not strong enough to allow for the interactions to reach significant p-values. If this is true, there might still be true epistatic signals between these pairs, and they would be picked up by running the full analysis as described in section 4.2.1.3.

To distinguish between these two explanations we decided to run the complete analysis, with a full episcan run and validation through logistic regression, on one of the filters. We chose the "cardiac original" filter, although it was one of the filters with the worse p-values (Figure 34). Our reasoning goes as follows: None of the filters yielded p-values that would be significant at their respective suggestive threshold and all did poorly. We hypothesized that this is because of poor LD with the tag-pairs from the genome-wide scan and therefore believe that the results in Figure 34 are not indicative of the actual potential of the filter. We, thus, choose the filter which we expect to have the highest biological validity, and we argue that this is the "cardiac original" one. Indeed, it is specific for cardiac tissues and uses predictions made on the dataset for which the ExPecto model has been trained. On the other hand, the very few SNPs reaching the selection threshold in the "cardiac" filter, made by predicting expression on our dataset led us to think that the ExPecto model has poor generalizability on did not make accurate predictions on our dataset (see discussion).

4.2.2.2 Testing the original Cardiac filter by running a new scan

We ran the fast filtering step using episcan on the 3148795 SNPs pair between the 2510 SNPs included in the "original cardiac" ExPecto filter. Although we do not have to account for incomplete LD this time, as the SNPs entering the episcan step are the exact SNPs included in the filter and not LD proxy, we still used a relaxed threshold of 1×10^{-6} instead of the suggestive threshold of 1.58×10^{-8} to account for the fact that we cannot include covariates in the episcan algorithm and the fact that p-values derived from the episcan are still an approximation of p-value obtained in the logistic regression. Also, there would have been no significant results to forward to the validation step with this threshold. On the contrary, we report 7 SNPs pairs that pass the relaxed threshold. Unfortunately, the distribution of episcan p-values does not show any enrichment in low p-values (Figure 35). Moreover, there is no overlap between the list of significant pairs at this step and the SNPs identified by the LD-tagging method described in the previous paragraph and shown in Figure 34. These pairs with predicted low p-values in the LD-based estimation did not even reach the threshold of 1×10^{-2} in this episcan run. This shows that indeed, the LD-based estimation of the episcan p-value is not reliable.

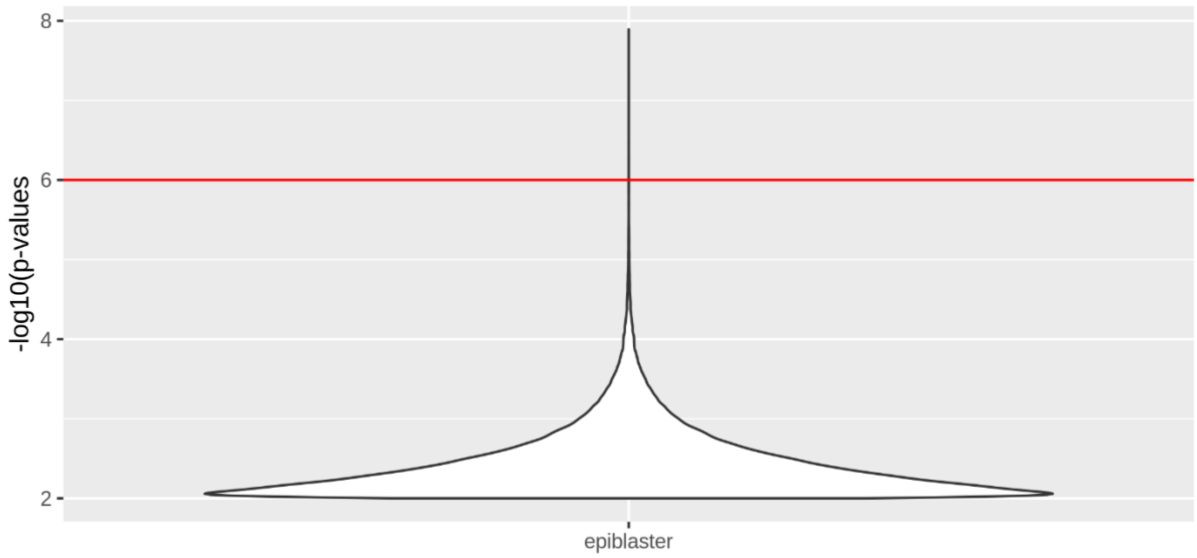


Figure 35: P-value distribution from the episcan run with the original cardiac filter
The original cardiac filter was built using a list of SNPs leading to a predicted log gene expression change greater than 0.3 for at least one gene in at least one cardiac tissue.

We then used the logistic regression model described in 3.2.2.2 to validate the selected pairs from the episcan step. Unfortunately, none of them did reach the suggestive threshold of 1.58×10^{-8} (Table 8). However, as there were only 7 of them we decided to try to replicate them anyways in the replication dataset. The replication was possible only for 3 of the 7 pairs as no LD proxy was found for the others in the replication dataset. And every one of these 3 pairs failed replication with p-values higher than 0.3 and even opposite effects for two of them (Table 8).

SNP1	SNP2	model SNP1	model SNP2	P-value episcan	P-value log.reg.	Coef. log reg.	Coef. repl.	P-value repl.
15:78556580	rs764097820	dom	rec	6.00E-07	2.41E-06	0.013	NA	NA
rs192386882	rs34472985	dos	het	7.24E-07	1.57E-07	-0.011	NA	NA
rs3211714	rs71460609	het	dos	9.70E-07	1.38E-06	-0.008	NA	NA
rs7164139	rs3734208	dos	rec	8.51E-07	1.83E-07	0.010	-0.056	0.391
rs6503807	rs3750524	dom	rec	5.50E-07	2.21E-07	-0.014	NA	NA
rs4782899	rs3755724	dos	rec	1.28E-08	1.72E-07	0.009	0.053	0.418
rs6741762	rs9782	dos	dom	9.93E-07	5.72E-07	-0.010	0.054	0.511

Table 8: Summary results for the 9 SNP-pairs which passed the episcan filtering step
The “Coef log.reg” and “Coef repl” columns respectively correspond to the log Odds for the interaction term in the validation step using logistic regression and in the replication. Opposite signs indicating opposite effect in these two models are considered a criteria to fail replication.

4.3 Biofilter SNPs filter

4.3.1 Methods

This "filter-based" analysis is a companion study to the genome-wide interactions scan presented in Section 3. It was conducted with the same dataset, using the same QC and the same methodology. The only difference is that the search space was narrowed down before step 1 using the filter that we present in this chapter.

4.3.1.1 definition of the Biofilter SNPs filter

We used the Biofilter software (Bush et al., 2008) to generate a list of SNP pairs to test based on evidence of their interaction at different biological levels. The Biofilter software brings together different sources of biological knowledge to predict possible SNPs interaction. More precisely, it integrates information about gene and protein interactions (BioGRID, MINT, and PharmGKB database), and information about biological pathways (Gene Ontology, KEGG, Reactome, Netpath) and other sources of information to predict possible interactions.

In this case, we fed the list of all SNPs after QC into biofilter and asked it to produce a list of pairs of genomic regions containing any of these SNPs which had evidence for interactions. Internally, the SNP's positions are mapped to known regions, genes, proteins, and other biological entities, and the whole database is scrutinized for interaction at the level of these entities. Biofilter then translates this list of putative biologically interacting entities back to genomic regions using annotations databases. Practically, we requested biofilter to output the 3000 pairs of regions with the highest level of evidence for interaction for each of the pairwise chromosome combinations. This threshold was set to allow a good coverage of the whole genome but still be computationally tractable and offer enough power. We then pruned each of these regions separately with an LD r^2 threshold of 0.5. For each suggested pair of genetic regions, we then listed all pairwise combinations of LD-pruned SNPs in the two regions.

4.3.1.2 Testing the Biofilter filter by running a new scan

We ran the episcan algorithm for all of the SNPs pairs in the Biofilter filter, in all combinations between the dominant, dosage, and heterozygous genetic models. Based on the realization that very few pairs were discovered specifically using combination including the recessive model and to save computation time we decided to not include the recessive dosage which needs a higher MAF threshold (see Section 3.3.1.2).

The analysis including the first episcan step and subsequent logistic regression, replication, and conditional analysis step was conducted as described for the genome-wide scan (see Section 3.2).

4.3.2 Results

We used the Biofilter software to integrate many available databases spanning different levels of biological entities and generate a list of SNPs pairs that are likely to interact. We generated a list of 84’353 pairs of regions to test. After region-wise LD pruning of each region and pairwise combination of LD-independent SNPs in every two regions of each pair, we obtained a list of 84'431'814 pairs to test. For each of these pairs, all combinations of genetic encoding, with the exception of the recessive one, were tested. For the first epistasis step, we used a relaxed threshold of 1.18×10^{-6} (computed as (1/number of LD independent SNP pairs) * 100) instead of the suggestive threshold of (1.18×10^{-8}) to account for incomplete LD between the SNPs included in the analysis and the possibly causal pruned SNPs that they might tag. Overall, the episcan results do not show a strong enrichment towards low p-value (Figure 36), similarly to what was observed using the ExPecto-based filter even if more than 500 pairs reach the relaxed threshold. More precisely, we report 521 SNPs pairs that passed this threshold and were forwarded to the validation step using logistic regression.

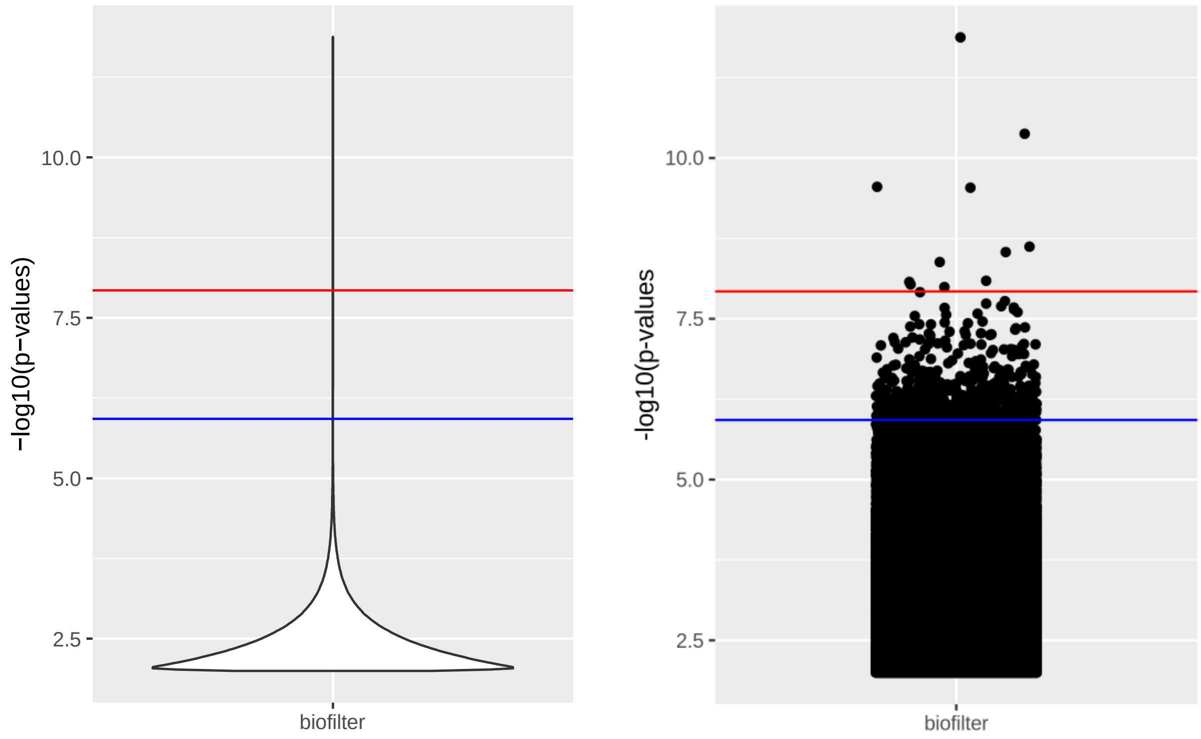


Figure 36: P-value distribution from the episcan run with the Biofilter-based filter
The overall shape of the distribution is shown on the violon plot on the left whereas the plot on the right displays the individual p-values.The blue line represents the relaxed threhsold of 1.18e-6 and the red line the suggestive threshold of 1.18e-8.

From all the LD-expanded pairs resulting from these 521 pairs, only 154 , with a large majority of trans-pairs, were significant at the suggestive threshold of 1.18×10^{-8} in the logistic regression and were forwarded to replication. Unfortunately, none of the trans-pairs did actually pass the replication, 88 of them failed and 29 were impossible to replicate due to missing proxies in the replication dataset. Out of the 37 cis-pairs, 19 did replicate and where forwarded to the conditional analysis, which they all failed. Indeed, for each of them there was at least one SNP in their vicinity which made their interaction effect drop below nominal significance when added to the model. In the end, we must report that this analysis based on the Biofilter framework to generate putative interacting SNPs did not reveal any significant and true epistasis pairs.

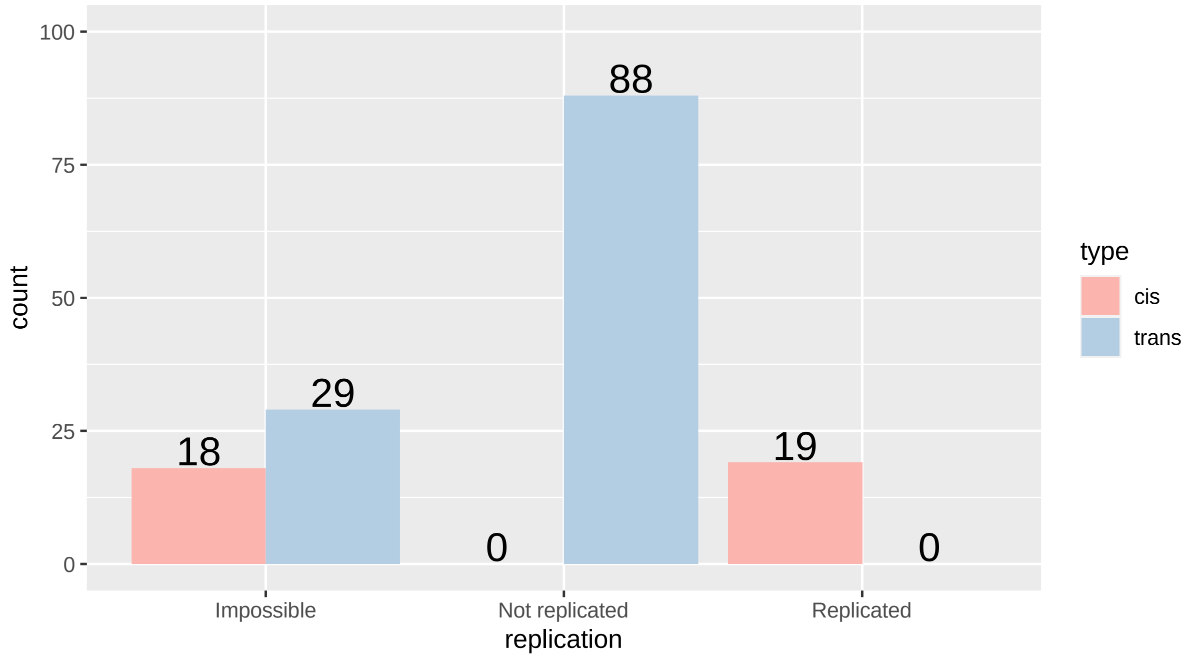


Figure 37: Replication of the significant pairs of the Biofilter-based analysis in a meta-analysis of 10 CAD studies

This barplot shows the results of the replication of the 154 significant pairs after the fine-mapping in a replication dataset composed of 10 CAD studies. Some of the pairs were impossible to replicate in this dataset because there were no LD tags (LD < 0.5) available in the replication dataset which was imputed using a smaller imputation panel.

4.4 Discussion

In this chapter, we presented two different filtering approaches to reduce the search space using different kinds of prior biological knowledge. Using such filters allows to mitigate the computation time and multiple testing burden problem created by the sheer number of SNPs combination to test in the human genome.

4.4.1 Regulatory SNPs filter

The first approach was aiming at screening for epistasis between SNPs with a predicted marginal effect on gene expression. Disappointingly this filter did not allow to detect any significant interactions in the fine-mapping and validation step. Looking at the distribution of the p-values at this step also indicates that there is absolutely no enrichment in low or close-to-significant p-values among the SNPs in this filter.

There could be two main explanations for these results. Firstly, while the authors of the ExPecto package published a very large list of 2‘443‘754 SNPs with a predicted in any kind of tissues, only 5877 of these SNPs, were present in our dataset. This could indicate that very few regulatory SNPs were indeed included in the imputation panels used on our dataset. Alternatively, most of the SNPs in our dataset might not have been included in the analysis done by the ExPecto authors but could still have a regulatory effect. Following this hypothesis we predicted the effect on gene expression for all the SNPs in our dataset using the pre-trained expecto model. But here as well, we found that very few of our SNPs had a predicted effect on gene expression. This could of course indicate that there is indeed an absence of regulatory SNPs in our imputation panel. Alternatively, the ExPecto model might not generalize well on our dataset.

The second possible explanation is that epistatic interactions might not happen between regulatory SNPs in CAD in particular or more generally. This hypothesis finds some support in the results of the hypothesis-free interaction association analysis that we conducted and which reported that a majority of the interactions happened between variants situated in coding regions (see Section 3.4.5) and not in the intergenic regions which are supposed to play a major regulatory role. In opposition to these results, Chatelain and colleagues report that the vast majority of epistasis interactions found in different diseases involve SNPs situated in intergenic regions. Whether this apparent discrepancy reflects different genetic architecture between CAD and other diseases or some possible differences between the two studies and their limitations remains unclear.

ExPecto is not the only tool that can be used to produce a list of regulatory SNPs. Indeed, big eQTLs databases such as GTEX or ENCODE could be scrutinized to produce a list of SNPs that affect gene expression. Conducting the same analysis as reported here using a filter based on these databases, whose results might generalize better than complex deep learning models, could potentially help determine if there is indeed no epistasis interaction between regulatory variants in CAD or if the methods and filter used in our analysis were suboptimal.

4.4.2 Biofilter SNPs filter

This second approach was aiming at screening for interactions between SNPs which are likely to interact based on prior biological knowledge at different biological levels, but agnostically to the disease's genetic architecture. While the first step of the analysis yielded more than 500 pairs of interest and a fair share of them were still significant in the logistic regression step, in the end we do not report any significant results.

Firstly, 88 trans pairs representing the majority of the 154 significant pairs after the second step did fail the replication. Whereas this number seems very big, a closer look at these pairs revealed that actually arose from the LD-expansion of only 10 of the 521 significant episcan pairs. They were therefore actually likely to represent only 10 different signals, which could not be replicated. There could be two reasons, for this failure of replication. Either these pairs do not have any significant effect in the replication dataset, or they have an opposite effect. It turns out, that none of them reached significance in the replication, although most of the effect size were in the same direction. We hypothesized that this low replication rate could by explained by the fact that the SNPs involved in these pairs might have had bad, or worse than average, proxy in the replication data. However, for 85 pairs over the 88, no proxy was needed as the original SNPs were present in the replication. We can therefore exclude that the replication failure was due to poor LD-tags.

Secondly, the 19 replicated cis-pairs did all fail the conditional analysis. Here as well, a closer look at these pairs reveals that both of their SNPs all came from the same genomic region near the CDKN2A gene. Actually, all of these 19 pairs involved the same SNP: rs3731239. And for all of these pairs, the decrease in the significance of the interaction term was the strongest when conditioned on the same SNP, namely rs4007642. rs4007642 is a known risk factor for CAD and for abdominal aortic aneurysm (J. Z. Liu, Erlich, & Pickrell, 2017; Singh, Field, Bown, Jones, & Golledge, 2021). Very interestingly, we found this same SNP to interact with rs3814960 (also located at the CDKN2A locus) in the hypothesis-free analysis that we conducted on the same data (see Section 3.4). And this interaction was independent of any testable variant located at this locus, including rs4007642.

This represents one more example of complex patterns of epistasis interactions and marginal effects at risk loci for CAD, similar to the situation which we describe at the LPA locus (see Section 5.3.3). It also underlines one more time, the importance of the conditional analysis to disentangle epistasis and the tagging of marginal effect to fine-map genuine interactions.

These results would tend to demonstrate the superiority of the hypothesis-free approach as compared to filtering approach. Indeed, because the hypothesis-free approach is unconstrained it was able to successfully find the true interaction signal at the CDKN2A locus. On the other hand, although previous biological knowledge apparently pointed towards this region, as it was included in the filter, the filter-based approach failed to identify the right pair because it was too constrained. This was arguably due to our implementation of the filter, and one could imagine using Biofilter to define regions of interest, and then based on those generate a filter in a different way that could encompass larger regions, or define regions around them using more biological knowledge. However, this clearly demonstrates that the filter-based approaches are extremely sensitive to the precise filter definition and implementation. And given our limited knowledge of epistasis mechanisms, it is very difficult to make prior hypotheses on how to generate a good filter.

Moreover, the results from the hypothesis-free interaction analysis (see Section 3) tend to suggest that interactions might happen preferentially between known risk factors of CAD. The two filters which we tested here are both agnostic to the genetic basis or etiology CAD, which could explain why they were unsuccessful. Whether this is trait-dependent or a general principle of epistasis is unknown. We can therefore not exclude that these two filters could perform better on other traits or diseases.

5 Cis-epistasis at the LPA locus and risk of cardiovascular diseases

5.1 Introduction

In this chapter, we present a filter-based interaction association study conducted on 56 GWAS CAD loci identified at the time of the study. The analysis presented in this chapter was conducted before the analysis discussed in Sections 3 and 4 but complements the set of different GWIAS approaches together with the hypothesis-free one and biological knowledge filter ones. Indeed, we use another popular filtering strategy which relies on the assumption that genetic variants involved in epistatic interaction are likely to have an additive effect as well. Importantly, we use here a broader version of this hypothesis as we included all SNPs in a ±300kb range from the GWAS hits. The actual assumption here is therefore that interactions happen between SNPs that are positionally, and therefore likely functionally, related to variants that have a marginal effect.

From a methodological point of view, this study was the predecessor of the two-step strategy presented in Section 3, with the difference that the fast statistical filtering step was conducted using GLIDE, an algorithm that approximates linear regression models, and not with episcan.

In this chapter, we will focus not focus on the first part of the epistasis scan, namely the actual identification of lead pairs, but on their validation and conditional analysis. Indeed, the first was conducted by colleagues and can be read in the open access manuscript that we published. I want to focus here on the part of the analysis that I carried out myself and that illustrate the difficulty to entangle epistasis signal for the tagging of rare variants and some methodology that we designed to tackle this important question. The whole analysis was published in Cardiovascular Research in 2022, (Zeng et al., 2022) and this chapter is therefore an adaptation of this open access manuscript that I drafted.

5.2 Methods

An overview of the two-step strategy for the identification of epistasis pairs using a meta-analysis of 10 CAD studies is depicted in Figure 38. Details about the cohorts and methods are described in (Zeng et al., 2022).

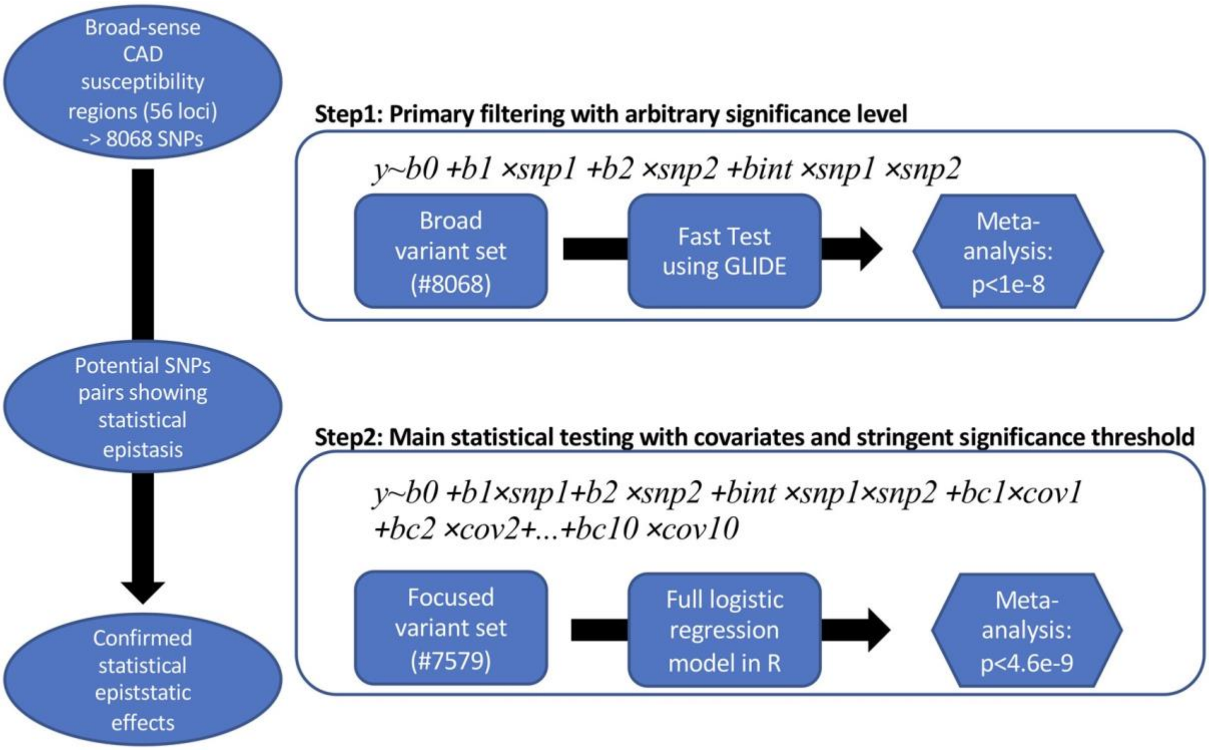

Broad-sense CAD susceptibility regions (56 loci) -> 8068 SNPs
Potential SNPs pairs showing statistical epistasis
Confirmed statistical epiststatic effects
Step1: Primary filtering with arbitrary significance level
y~b0 +b1 ×snp1 +b2 ×snp2 +bint ×snp1 ×snp2
Broad variant set (#8068)
Fast Test using GLIDE
Meta-analysis: p<1e-8
Step2: Main statistical testing with covariates and stringent significance threshold
y~b0 +b1×snp1+b2 ×snp2 +bint ×snp1×snp2 +bc1×cov1 +bc2 ×cov2+...+bc10 ×cov10
Focused variant set (#7579)
Full logistic regression model in R
Meta-analysis: p<4.6e-9

Figure 38: Scheme of the two-stage statistical interaction scanning procedure
“Step 1 aimed at the fast identification of potential significant interaction terms using the GLIDE GPU computation tool. For each pair of LD-independent SNPs in the susceptibility regions (n = 8068 SNPs), we fitted a linear model with the additive and interaction effect of the two SNPs in each of the 10 CAD studies separately. The 10 p-values were then meta-analyzed. A loose and arbitrary defined significance level ($p < 1\times10^{-8}$) was applied with the assumption that if there exists true epistasis between two lead SNPs, loose signals should be detectable between the SNPs within the corresponding LD block. Step 2 aimed at validating the results of the first step using a logistic regression model including the first 10 multi-dimensional scaling (MDS) components of the genetic relationship matrix to correct for population structure. Step 2 also allowed the fine-mapping of candidate SNP pairs by screening for the strongest signal among all the SNPs within the LD blocks forwarded from step 1. In this second step, we applied a stringent significance threshold of $4.6\times10^{-9,}$ calculated as a Bonferroni correction ($0.05/(n_{SNP\_indep} \times (n_{SNP\_indep} - 1)/2) = 4.6178\times10^{-9}$) on the number of LD-independent SNPs resulting from step 1 ($n_{SNP\_indep} = 4654$).”. Figure and Legend from (Zeng et al., 2022)

5.3 Results

5.3.1 Initial discovery of epistatic pair

The two-step epistasis scan conducted on 8068 LD-independent SNPs located at a ±500kb distance from one of the 56 GWAS hits yielded 4 significant SNPs pairs at the threshold of 4.618×10^{-9} in the fine-mapping step. Subsequent replication in the UKBB dataset was successful for two SNP pairs and the most significant one was prioritized for further analysis. This cis-epistasis pair was located on chromosome 6 between two SNPs annotated to the LPA gene, rs1800769, and rs9458001. Interestingly the interaction term between these two SNPs was also associated with the circulating LPA levels in the KORA F3/F4 dataset and one of the 10 CAD studies used for the discovery (the LURIC study). Finally, the interaction between the two SNPs was also found to be associated with LPA gene expression in liver tissue in the STARNET study. Altogether, these results provided strong evidence that the interaction between rs1800769 and rs9458001 was associated with CAD and even provided a mechanism of action, namely the up-regulation of circulating LPA levels.

5.3.2 Effect of rs140570886 at the LPA locus

However, during the review process, one reviewer argued that the interaction effect could be more parsimoniously explained by the additive effect of one rare variant at the LPA locus, rs140570886. Although we had conducted a conditional analysis (as described in section 3.2.2.4), we had missed rs140570886 because we conducted this analysis on the QCed dataset from which it was excluded because of a MAF of 0.015. When conditioning on this rarer variant, the significance of the rs1800769:rs9458001 interaction term dropped drastically from $p = 8.95\times10^{-14}$ to $p = 0.022$, thereby confirming that it might not independent of the effect of rs140570886.

We next confirmed that rs140570886 was associated with CAD in an additive way in our dataset (OR = 1.98, $p = 1.14\times10^{-21}$) and also with the LPA levels in the KORA studies (β=1.54, $p = 9.52\times10^{-82}$).

5.3.3 Complex interactions pattern at the LPA locus between rs140570886, rs1800769, and rs9458001

Although some of the effects that we originally attributed to the rs1800769:rs9458001 interaction seemed to be due to the additive effect of rs140570886, the still nominally significant p-value for the interaction term when conditioned on rs140570886 (p=0.022) suggested that a weak epistasis signal could be present in addition. We, therefore, set out to investigate more in detail the interplay of these 3 SNPs at the LPA locus, and to more generally develop a strategy to untangle epistasis from the effect of additive variants in its neighborhood

In a first step, we started building nested logistic models of increasing complexity and assessed the gain in model fit using the likelihood ratio test (Table 9). To increase the power of this analysis, we conducted it on the discovery (10 CAD studies) and replication dataset (UKBB) jointly. We started with the most parsimonious model containing the additive effect of rs140570886 only. Adding the marginal effect of rs9458001 and rs1652507 (rs1652507 is a proxy for rs1800769 (r^2=0.965) available in the UKBB dataset) did significantly increase model fit ($P=5.1\times10^{-8}$). Given that neither rs1800769 nor rs9458001were associated with CAD in their marginal effect (p=0.59, odds ratio (OR) = 0.99 for rs1800769[T]; p = 0.08, OR = 1.04 for rs9458001[A]) this already suggested that some non-additive effect might be at play. Surprisingly, the interaction between rs9458001 and rs1652507 to this previous model did not increase model fit significantly. However, adding all three-way interactions between the 3 SNPs did significantly increase model fit (p=0.004) and yielded the best model of our model-building strategy. The addition of marginal and epistasis effect of a fourth SNPs, which decreased the p-value of the rs9458001:rs1652507 interaction in the original conditional analysis, did not improve model fit further.

Model	Residuals. Df	Residuals Deviance	Df	Deviance	P-value	P-value LRT model 2
1) CAD ~ covariates	342046	222021				NA
2) CAD ~ rs140570886 + covariates	342045	221804	1	217.04	4×10^{-49}	NA
3) CAD ~ rs140570886 + rs9458005 + rs1652507 + covariates	342043	221770	2	33.57	5.1×10^{-08}	5.1×10^{-08}
4) CAD ~ rs140570886 + rs9458005 * rs1652507 + covariates	342042	221768	1	2.32	0.13	7.9×10^{-08}
5) CAD ~ rs140570886 * rs9458005 * rs1652507 + covariates	342039	221755	3	13.42	0.004	6.5×10^{-09}
6) CAD ~ rs140570886 * rs9458005 * rs1652507 +rs3798220+ covariates	342038	221753	1	1.93	0.16	8.2×10^{-09}
7) CAD ~ rs140570886 * rs9458005 * rs1652507 * rs3798220+ covariates	342031	221746	7	6.22	0.51	8.2×10^{-09}

Table 9: ANOVA table reporting likelihood ratio test results for nested models in the model selection procedure at the LPA locus

"The table displays the result of a series of successive likelihood ratio tests between a nested model of increasing complexity performed on the merged dataset including the 10 CAD studies and the UK Biobank dataset. The first and second columns report the Residual Deviance and degrees of freedom of each row's model. The 'Df' and 'Deviance' columns respectively report the difference in degrees of freedom and deviance between each row's model and the model from the previous row. The 'p-value' column reports the P-value of the likelihood ratio test between each row's model and the previous one. The 'p-value LRT model 2' column reports the p-value of the likelihood ratio test between each model and the model containing only the additive effect of rs140570886. The * operator denotes factor crossing: a*b is interpreted as a + b + a × b and a*b*c as a + b + c + a × b + a × c + b × c. The 10 multi-dimensional scaling components of the genetic variance and the study were included as covariates in every model." Table and legend reproduced from (Zeng et al., 2022).

In order to further examine the seemingly complex pattern of interaction between rs140570886, rs9458001, and rs1652507 we evaluated the strength and significance of every coefficient in the winning model using a type III Sum of Squares ANOVA (Table 10).

Variable	Estimate	Std. Error	P-value
rs140570886	0.08	0.14	0.55
rs9458001	– 0.02	0.01	0.07
rs1652507	-0.07	0.01	3.75E-08
rs140570886:rs9458001	0.29	0.13	0.02
rs140570886:rs1652507	0.25	0.12	0.04
rs9458001:rs1652507	0.00	0.03	0.92
rs140570886:rs9458001:rs1652507	-0.13	0.11	0.23

Table 10: Summary statistics for the rs140570886, rs1652507 and rs9458001 and their interactions in the final 3-way interaction model
Table and legend reproduced from (Zeng et al., 2022).

Interestingly, the additive effect of rs140570886, which was thought to carry all of the association signal after the conditional analysis, is not significant anymore when conditioned on the 2 other SNPs and all 2 and 3 ways interactions. On the contrary, the most significant effect is now carried out by rs1652507. The two interactions of rs140570886 with rs1652507 and rs9458001 respectively are also significant and display the biggest effect size of all variables.

To get a less statistical and more biological view of the data, we identified in every individual the 3-SNP-haplotypes defined by these three variants and computed the effect size associated with each of the eight haplotypes. It appears in Figure 39, which represents the effect size of each haplotype, that all haplotypes containing the T allele at rs140570886 have similar ORs as the most common one. Interestingly, substituting the rs140570886 T allele for the C increases the CAD risk in presence of the rs1652507 C allele (blue vs red lines) but tends to decrease it in presence of the rs1652507 T allele. This is the correlate of the significant interaction term between rs140570886 and rs1652507 in Table 10. Similarly, substituting the T allele at rs140570886 for the C decreases the risk of CAD in presence of the rs9458001 A but not in the presence of the G allele (difference between blue square and blue triangle). This is coherent with the significant interaction between rs140570886 and rs9458001 in Table 10. The increase in CAD risk when substituting the rs9458001 G allele for the A allele in the rs140570886 T background (difference between red square and red triangle) is also likely to contribute to this significant interaction term.

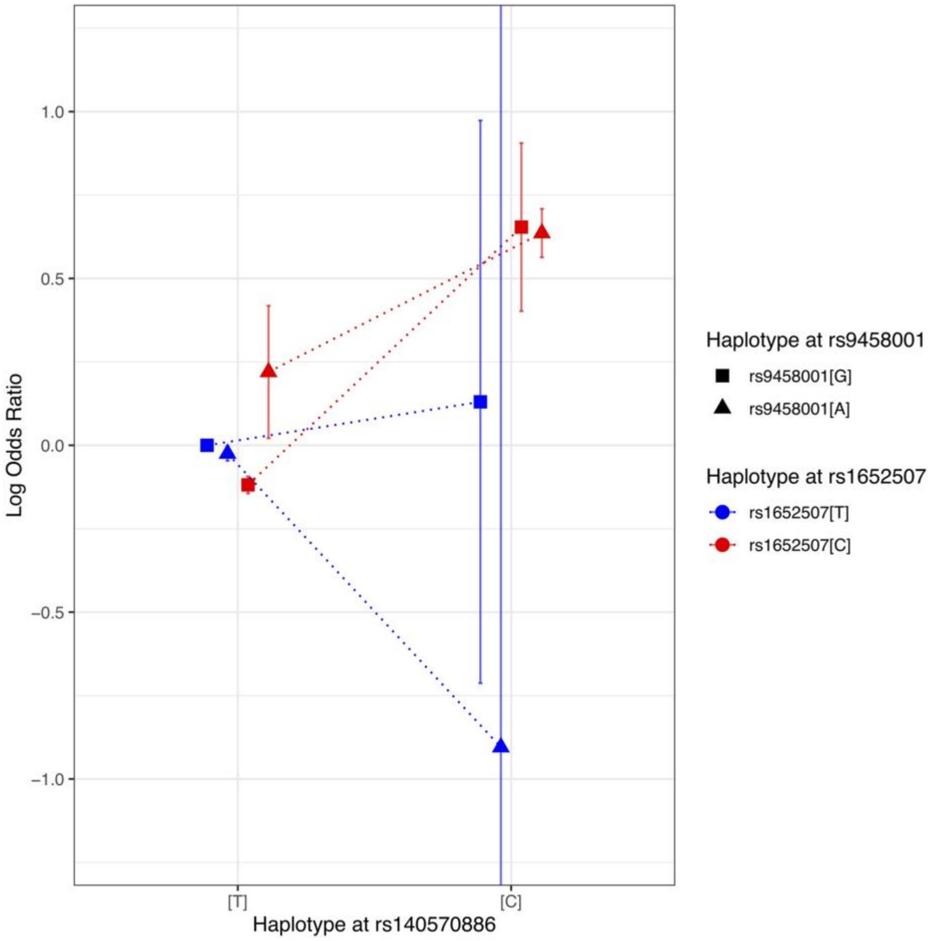


Figure 39: Relative effect of the rs140570886-rs1652507-rs9458001 3-SNP haplotypes on CAD risk "Relative odds ratio (OR; with reference to the most frequent TTG haplotypes) for the eight possible 3-SNP haplotypes on CAD risk. The red and blue colors represent the base at the rs1652507 SNP, the square and triangle shapes represent the base at the rs9458001 SNP and the position on the x-axis represents the base at the rs140570886 SNP. Together they indicate the eight possible 3-SNP haplotypes. The putative haplotypes were computed using the happasoc R package on a merged dataset of the 10 CAD studies ($N = 29\ 755$). Error bars represent the standard error of the log OR." Figure and legend reproduced from (Zeng et al., 2022).

Taken together, these results show that the significant interaction terms in the model that best fit the data are reflecting the different effects of the two alleles at rs140570886 in different genetic backgrounds defined by the two other SNPs. As this is one of the definitions of epistasis, we conclude that there is indeed complex epistatic interactions between these three SNPs that modulate the risk of CAD.

Although the previous results provide abundant evidence for epistatic interactions at the LPA locus, one could still argue that the 3-SNP-combinations defining the 8 haplotypes described in Figure 39 could actually just tag one more, rarer or ungenotyped causal variant. To formally test this hypothesis, we ran a last analysis to compare models based on haplotypes and models based on interactions. Namely, we used the Akaike Information Criterion (AIC) to compare non-nested models including either the 3 SNPs encoded in the dosage model, 3-SNPs haplotypes or interactions, or both. Table 11 shows that the lowest AIC is achieved when including the marginal effect of the SNPs and the interactions in the model. Furthermore, adding the haplotype to this model does not increase model fit (p=0.262), as tested using the likelihood ratio between these nested models. This analysis, therefore, brings one more piece of evidence for a complex pattern of epistatic interactions between rs140570886, rs1652507, and rs9458001 at the LPA locus.

Model	Model name	AIC	Comparison M_SNPs	Comparison M_interact
no genetics	M_null	222063.0	NA	NA
haplotypes	M_haplo	221813.6	NA	NA
SNPs	M_SNPs	221818.3	NA	NA
SNPs + interactions	M_interact	221810.6	0.0034	NA
haplotypes + SNPs + interactions	M_full	221814.4	0.0268	0.262

Table 11: Model selection using AIC and Likelihood ratio test confirms epistatic interactions at the LPA locus

"This table displays the Akaike Information Criterion (AIC) and results of likelihood ratio test for nested models of increasing complexity performed on the merged dataset including the 10 CAD studies and the UK Biobank dataset. The "Comparison M_SNPs" and "Comparison M_interact" columns respectively report the p-values of the likelihood ratio tests with the M_SNPs and M_interact models as null model. The 10 multidimensional scaling components of the genetic variance and were included as covariates in every model. NA=non-applicable." Table and legend reproduced from (Zeng et al., 2022).

5.4 Discussion

In this study, we conducted an interaction association analysis for CAD based on variants previously associated in two GWAS and reported four SNP pairs whose interactions were significantly associated with the disease. Further investigation on the most significant association between rs1800769 and rs9458001 both located at the LPA locus on chromosome 6 led to the discovery of a complex interaction pattern between these two SNPs and a rare variant rs140570886.

The small number of interactions discovered in this study shows once again the difficulty of detecting epistasis in complex traits. Several reasons could explain this low number of results. Firstly, it is possible that epistasis does not happen between genes that are associated with CAD in an additive manner, which was the major assumption of this filter-based approach. Secondly, although already pulling together data from 10 different studies, the power of this study to detect epistasis with low to moderate effect size might have been very low. The fact that our hypothesis-free genome-wide association study using the much bigger UKBB dataset (see Section 3) discovered 15 interacting pairs located in the LPAL2-SLC22A2/3-LPA gene cluster, which was included in this analysis because it falls within the ±300 kb distance of one of the GWAS 56 GWAS hits, tends to confirm the hypothesis of a lack of power. Indeed, ad-hoc power computation using the epilogitpower package (which had not been developed at the time when this study was planned) indicates a power of ca. 10% to detect effect sizes of OR 2 between SNPs with a MAF of 5% in this study. This highlight one more time, the extremely big sample size needed to detect epistatic interactions and the need to estimate them properly while planning a GWIAS.

This study also demonstrates the importance of conditional analysis to disentangle epistasis effects from the tagging of a haplotype or rare variant. Moreover, it proposes a method to determine the independence of epistasis signal from neighboring marginally associated variants and to resolve complex cases of interactions that could be mistaken with marginal effects. We show that by building models of increasing complexity including the main effect of the interaction SNPs and the variants supposedly tagged by their interaction first, followed by their interactions one by one, one can actually resolve the actual nature of the association. By computing ORs for each of the eight haplotypes produced by the combination of these 3 SNPs, we moreover demonstrate that our statistical method indeed identified changes in the effect of allele substitution at one locus depending on the genotype at other loci, which is one of the definitions of biological epistasis.

Finally, taken together these results provide new insights into the genetic architecture of the LPA locus, one important risk locus for CAD, and a potential target for new drug development (Tsimikas et al., 2020).

6 Conclusion and Outlook

In this thesis, we aimed at tackling some of the challenges linked to the identification of epistasis using statistical genetics and providing methodological advances to overcome them. Moreover, using four complementary approaches based on different hypotheses regarding the architecture of genetic interactions, we aimed at uncovering the basic principles governing epistasis in CAD and discovering interactions modulating the risk to develop the disease.

6.1 Power computation

We developed the epilogitpower package which implements fast power computation for GWIAs. Using this package, we could show that very big sample sizes are needed to achieve sufficient power in the settings that are commonly used in epistasis studies and raised awareness that smaller datasets will result in underpowered studies. We hope that the availability of this user-friendly and well-documented R package, will allow other researchers to conduct power calculations when planning interaction association studies and help interpret the results of these studies. We also show how we used this package to make informed decisions about QC criteria such as the MAF in our hypothesis-free epistasis scan.

6.2 Haplotype effect and rare-variant tagging

Haplotype or rare variants tagging was a known pitfall of interaction association studies (Fish et al., 2016). In this thesis, we developed a simple conditional analysis method allowing to rule out this confounding. We advocate that this should become a step in its own right in any GWIAS, as we report that half of the initially significant interactions in the hypothesis-free GWIAS and all of them in the Biofilter-based approach were merely tagging rare associated variants. Moreover, our results seem to indicate that epistasis happens preferentially at locus which are associated to CAD with additive effects. This presence of additive effects in the vicinity of the putative interacting pairs makes it even more important to conduct such conditional analysis. We moreover show that the dependence of the interaction coefficient on a rare variant might be indicative of more complex interaction pattern and propose a method to resolve this complicated case. Applying this method, we report a complex pattern of epistasis at the LPA locus, an important risk factor for CAD.

6.3 Hypothesis-free interaction association study

Exhaustive association testing of all possible SNP pairs in the genome has been a prohibitively heavy computational task until now (Niel et al., 2015). In this thesis, we show that this is actually feasible using a two-step strategy with a fast statistical filter. We also show that adapting the testing strategy as well as its software implementation to the available hardware resources can result in substantial computation time gain, which can make the difference between the feasibility of the study and its impossibility.

Using this hypothesis-free approach we could uncover 17 new interacting pairs associated with CAD, and potentially more pairs that still need to be replicated. Due to our limited understanding of the genetic architecture of epistasis in complex diseases, we argue that the hypothesis-free approach represents the most promising method to detect novel gene interactions. Indeed, we show that the two filter-based methods could not identify any of 17 pairs that the hypothesis-free scan could. More dramatically, they could not identify any significant independent SNP interactions. Based on the fact that all of the 17 pairs that we identified are located within CAD risk loci, we hypothesize that a filter based on the regions surrounding known CAD risk loci might be more successful than the regulatory and Biofilter-based filter which we tried here. Indeed, a filter of this kind allowed us to discover a complex interaction pattern at the LPA locus with a much smaller dataset than the one used for the other filters. We, therefore, think that conducting an interaction search in the vicinity of the now 321 know risk loci for CAD (Chen & Schunkert, 2021) in a large sample size might discover new epistatic interactions for CAD.

6.4 Principle of epistasis in CAD

Assessing the presence of epistasis using the significance of the interaction term in a logistic model allows to discover different types of interactions, which might be underpinned by different biological mechanisms but lead to the same deviation in additivity in the logistic model (de Visser et al., 2011). Very interestingly, of all the possible combinations of the marginal effects of the interacting SNPs and interaction effect, we actually report that only three of them were represented. Namely, we found antagonist-positive and antagonist-negative epistasis between SNPs that have a marginal deleterious or protective effect but an opposite interaction effect. In addition, we also found negative epistasis between SNPs that do not have a marginal effect but whose interaction is deleterious. Moreover, all these interactions happened between SNPs located in known CAD risk loci and located on the same chromosome (cis-epistasis). Further hypothesis-free GWIAS will be needed to investigate whether these results represent fundamental disease-agnostic principles of the genetic architecture of epistasis or if they differ between traits.

6.5 Outlook

With always improving computational hardware and increasing sample sizes, well-powered genome-wide epistasis studies will certainly become more and more frequent. Their result should help confirm or inform the hypothesis that we draw here about the different types of epistasis and possible common genetic principles for genetic interactions across diseases. Increasing sample sizes will also allow to look for epistasis between rarer variants and possibly discover more interactions, which together might contribute substantially to the genetic variance of traits and help personalized risk prediction.

Continuous improvement in the functional characterization and effect prediction for SNPs (Cheng et al., 2019; Wagih et al., 2018; Yazar & Özbek, 2021) for example along with the democratization of NGS data should moreover allow researchers to increase the resolution of their analysis and detect more fine-grained and complex patterns of interaction and additive effect as well as strengthen our understanding of the molecular basis of interactions and their contribution to disease pathophysiology.

7 References

Anholt, R. R. H., Dilda, C. L., Chang, S., Fanara, J. J., Kulkarni, N. H., Ganguly, I., … Mackay, T. F. C. (2003). The genetic architecture of odor-guided behavior in Drosophila: Epistasis and the transcriptome. *Nature Genetics*, *35*(2), 180–184. https://doi.org/10.1038/ng1240

Arloth, J., Eraslan, G., Andlauer, T. F. M., Martins, J., Iurato, S., Kühnel, B., … Mueller, N. S. (2020). DeepWAS: Multivariate genotype-phenotype associations by directly integrating regulatory information using deep learning. *PLoS Computational Biology*, *16*(2), e1007616. https://doi.org/10.1371/journal.pcbi.1007616

Azevedo, R. B. R., Lohaus, R., Srinivasan, S., Dang, K. K., & Burch, C. L. (2006). Sexual reproduction selects for robustness and negative epistasis in artificial gene networks. *Nature*, *440*(7080), 87–90. https://doi.org/10.1038/nature04488

Bomba, L., Walter, K., & Soranzo, N. (2017). The impact of rare and low-frequency genetic variants in common disease. *Genome Biology*, *18*(1), 1–17. https://doi.org/10.1186/s13059-017-1212-4

Bush, W. S., Dudek, S. M., & Ritchie, M. D. (2008). Biofilter: A Knowledge-Integration System for the Multi-Locus Analysis of Genome-Wide Association Studies. *Biocomputing 2009*, 368–379. https://doi.org/10.1142/9789812836939_0035

Bush, W. S., & Moore, J. H. (2012). Chapter 11: Genome-Wide Association Studies. *PLoS Computational Biology*, *8*(12). https://doi.org/10.1371/journal.pcbi.1002822

Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L. T., Sharp, K., … Marchini, J. (2018). The UK Biobank resource with deep phenotyping and genomic data. *Nature*, *562*(7726), 203–209. https://doi.org/10.1038/s41586-018-0579-z

Camp, K. M., & Trujillo, E. (2014). Position of the academy of nutrition and dietetics: Nutritional genomics. *Journal of the Academy of Nutrition and Dietetics*, *114*(2), 299–312. https://doi.org/10.1016/j.jand.2013.12.001

Chatelain, C., Lessard, S., Thuillier, V., Carliez, C., Rajpal, D., & Augé, F. (2021). Atlas of epistasis. *MedRxiv*. Retrieved from https://doi.org/10.1101/2021.03.17.21253794

Chen, Z., & Schunkert, H. (2021). Genetics of coronary artery disease in the post-GWAS era. *Journal of Internal Medicine*, *290*(5), 980–992. https://doi.org/10.1111/joim.13362

Cheng, J., Nguyen, T. Y. D., Cygan, K. J., Çelik, M. H., Fairbrother, W. G., Avsec, Ž., & Gagneur, J. (2019). MMSplice: Modular modeling improves the predictions of genetic variant effects on splicing. *Genome Biology*, *20*(1), 1–15.

https://doi.org/10.1186/s13059-019-1653-z

Cheverud, J. M., & Routman, E. J. (1995). Epistasis and its contribution to genetic variance components. *Genetics*, *139*(3), 1455–1461. https://doi.org/10.1093/genetics/139.3.1455

Cho, H., Shen, G. Q., Wang, X., Wang, F., Archacki, S., Li, Y., … Wang, Q. K. (2019). Long noncoding RNA ANRIL regulates endothelial cell activities associated with coronary artery disease by up-regulating CLIP1, EZR, and LYVE1 genes. *Journal of Biological Chemistry*, *294*(11), 3881–3898. https://doi.org/10.1074/jbc.RA118.005050

Combarros, O., Cortina-Borja, M., Smith, A. D., & Lehmann, D. J. (2009). Epistasis in sporadic Alzheimer's disease. *Neurobiology of Aging*, *30*(9), 1333–1349. https://doi.org/10.1016/j.neurobiolaging.2007.11.027

Cordell, H. J. (2009). Detecting gene-gene interactions that underlie human diseases. *Nature Reviews Genetics*, *10*(6), 392–404. https://doi.org/10.1038/nrg2579

Costanzo, M., VanderSluis, B., Koch, E. N., Baryshnikova, A., Pons, C., Tan, G., … Boone, C. (2016). A global genetic interaction network maps a wiring diagram of cellular function. *Science*, *353*(6306), aaf1420. https://doi.org/10.1126/science.aaf1420

de Visser, J. A. G. M., Cooper, T. F., & Elena, S. F. (2011). The causes of epistasis. *Proceedings of the Royal Society B: Biological Sciences*, *278*(1725), 3617–3624. https://doi.org/10.1098/rspb.2011.1537

Dunham, I., Kundaje, A., Aldred, S. F., Collins, P. J., Davis, C. A., Doyle, F., … Lochovsky, L. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, *489*(7414), 57–74. https://doi.org/10.1038/nature11247

Enkhmaa, B., Anuurad, E., & Berglund, L. (2016). Lipoprotein (a): Impact by ethnicity and environmental and medical conditions. *Journal of Lipid Research*, *57*(7), 1111–1125. https://doi.org/10.1194/jlr.R051904

Fish, A. E., Capra, J. A., & Bush, W. S. (2016). Are Interactions between cis-Regulatory Variants Evidence for Biological Epistasis or Statistical Artifacts? *American Journal of Human Genetics*, *99*(4), 817–830. https://doi.org/10.1016/j.ajhg.2016.07.022

Fuior, E. V., & Gafencu, A. V. (2019). Apolipoprotein c1: Its pleiotropic effects in lipid metabolism and beyond. *International Journal of Molecular Sciences*, *20*(23), 1–25. https://doi.org/10.3390/ijms20235939

Gretton, A., Borgwardt, K. M., Rasch, M., Schölkopf, B., & Smola, A. J. (2007). A kernel method for the two-sample-problem. *Advances in Neural Information Processing Systems*, 513–520. https://doi.org/10.7551/mitpress/7503.003.0069

Gudicha, D. W., Schmittmann, V. D., & Vermunt, J. K. (2017). Statistical power of likelihood ratio and Wald tests in latent class models with covariates. *Behavior Research Methods*, *49*(5), 1824–1837. https://doi.org/10.3758/s13428-016-0825-y

Gusareva, E. S., & Van Steen, K. (2014). Practical aspects of genome-wide association interaction analysis. *Human Genetics*, *133*(11), 1343–1358. https://doi.org/10.1007/s00439-014-1480-y

Gyenesei, A., Moody, J., Semple, C. A. M., Haley, C. S., & Wei, W. H. (2012). HHigh-throughput analysis of epistasis in genome-wide association studies with BiForce. *Bioinformatics*, *28*(15), 1957–1964. https://doi.org/10.1093/bioinformatics/bts304

Hemani, G., Shakhbazov, K., Westra, H. J., Esko, T., Henders, A. K., McRae, A. F., … Powell, J. E. (2014). Detection and replication of epistasis influencing transcription in humans. *Nature*, *508*(7495), 249–253. https://doi.org/10.1038/nature13005

Hill, W. G., Goddard, M. E., & Visscher, P. M. (2008). Data and theory point to mainly additive genetic variance for complex traits. *PLoS Genetics*, *4*(2). https://doi.org/10.1371/journal.pgen.1000008

Huang, J., Howie, B., McCarthy, S., Memari, Y., Walter, K., Min, J. L., … Zhang, W. (2015). Improved imputation of low-frequency and rare variants using the UK10K haplotype reference panel. *Nature Communications*, *6*, 1–9. https://doi.org/10.1038/ncomms9111

Huang, K., Zhong, J., Li, Q., Zhang, W., Chen, Z., Zhou, Y., … Zhang, S. (2019). Effects of CDKN2B-AS1 polymorphisms on the susceptibility to coronary heart disease. *Molecular Genetics and Genomic Medicine*, *7*(11), 1–8. https://doi.org/10.1002/mgg3.955

Kam-Thong, T., Czamara, D., Tsuda, K., Borgwardt, K., Lewis, C. M., Erhardt-Lehmann, A., … Müller-Myhsok, B. (2011). EPIBLASTER-fast exhaustive two-locus epistasis detection strategy using graphical processing units. *European Journal of Human Genetics*, *19*(4), 465–471. https://doi.org/10.1038/ejhg.2010.196

Kamstrup, P. R., Tybjærg-hansen, A., Steffensen, R., & Nordestgaard, B. G. (2009). Genetically Elevated Lipoprotein (a). *The Journal of the American Medical Association*, *301*(22), 2331–2339.

Kerem, B., Rommens, J. M., Buchanan, J. A., Markiewicz, D., Cox, T. K., Chakravarti, A., … Tsui, L. (1989). Identification of the Cystic Fibrosis Gene: Genetic Analysis. *Science*, *245*(4922), 1073–1079.

Koch, W., Mueller, J. C., Schrempf, M., Wolferstetter, H., Kirchhofer, J., Schömig, A., & Kastrati, A. (2013). Two Rare Variants Explain Association with Acute Myocardial Infarction in an Extended Genomic Region Including the Apolipoprotein(A) Gene. *Annals of Human Genetics*, *77*(1), 47–55. https://doi.org/10.1111/j.1469-

1809.2012.00739.x

Kronenberg, F. (2019). Prediction of cardiovascular risk by Lp(a) concentrations or genetic variants within the LPA gene region. *Clinical Research in Cardiology Supplements*, *14*, 5–12. https://doi.org/10.1007/s11789-019-00093-5

Krzywinski, M., & Altman, N. (2013). Points of significance: Power and sample size. *Nature Methods*, *10*(12), 1139–1140. https://doi.org/10.1038/nmeth.2738

Kuzmin, E., VanderSluis, B., Wang, W., Tan, G., Deshpande, R., Chen, Y., … Myers, C. L. (2018). Systematic analysis of complex genetic interactions. *Science*, *360*(6386). https://doi.org/10.1126/science.aao1729

Lappalainen, T., Montgomery, S. B., Nica, A. C., & Dermitzakis, E. T. (2011). Epistatic selection between coding and regulatory variation in human evolution and disease. *American Journal of Human Genetics*, *89*(3), 459–463. https://doi.org/10.1016/j.ajhg.2011.08.004

Liu, J. Z., Erlich, Y., & Pickrell, J. K. (2017). Case-control association mapping by proxy using family history of disease. *Nature Genetics*, *49*(3), 325–331. https://doi.org/10.1038/ng.3766

Liu, Y., Xu, H., Chen, S., Chen, X., Zhang, Z., Zhu, Z., … Kong, X. (2011). Genome-wide interaction-based association analysis identified multiple new susceptibility loci for common diseases. *PLoS Genetics*, *7*(3). https://doi.org/10.1371/journal.pgen.1001338

Mackay, T. F. C. (2014). Epistasis and quantitative traits: Using model organisms to study gene-gene interactions. *Nature Reviews Genetics*, *15*(1), 22–33. https://doi.org/10.1038/nrg3627

Mackay, T. F. C., & Moore, J. H. (2014). Why epistasis is important for tackling complex human disease genetics. *Genome Medicine*, *6*(6), 6–8. https://doi.org/10.1186/gm561

Mahley, R. W. (2016). Apolipoprotein E: from cardiovascular disease to neurodegenerative disorders. *Journal of Molecular Medicine*, *94*(7), 739–746. https://doi.org/10.1007/s00109-016-1427-y

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., … Visscher, P. M. (2009). Finding the missing heritability of complex diseases. *Nature*, *461*(7265), 747–753. https://doi.org/10.1038/nature08494

McPherson, R., Pertsemlidis, A., Kavaslar, N., Stewart, A., Roberts, R., Cox, D. R., … Cohen, J. C. (2007). A common allele on chromosome 9 associated with coronary heart disease. *Science*, *316*(5830), 1488–1491. https://doi.org/10.1126/science.1142447

Moore, C. M., Jacobson, S. A., & Fingerlin, T. E. (2020). Power and Sample Size

Calculations for Genetic Association Studies in the Presence of Genetic Model Misspecification. *Human Heredity*, *84*(6), 256–271. https://doi.org/10.1159/000508558

Moore, J. H., & Williams, S. M. (2005). Traversing the conceptual divide between biological and statistical epistasis: Systems biology and a more modern synthesis. *BioEssays*, *27*(6), 637–646. https://doi.org/10.1002/bies.20236

Morton, N. (1982). *Outline of Genetic Epidemiology*. Karger.

Nelson, C. P., Goel, A., Butterworth, A. S., Kanoni, S., Webb, T. R., Marouli, E., … Deloukas, P. (2017). Association analyses based on false discovery rate implicate new loci for coronary artery disease. *Nature Genetics*, *49*(9), 1385–1391. https://doi.org/10.1038/ng.3913

Niel, C., Sinoquet, C., Dina, C., & Rocheleau, G. (2015). A survey about methods dedicated to epistasis detection. *Frontiers in Genetics*, *6*(285). https://doi.org/10.3389/fgene.2015.00285

Nilesh, S., Erdmann, J., Hall, A. S., Mangino, M., Mayer, B., Dixon, R. J., … Schunkert, H. (2007). Genomewide Association Analysis of Coronary Artery Disease. *New England Journal of Medicine*, *357*(5), 443–453. https://doi.org/10.1056/NEJMoa072366.Genomewide

Novembre, J., Johnson, T., Bryc, K., Kutalik, Z., Boyko, A. R., Auton, A., … Bustamante, C. D. (2008). Genes mirror geography within Europe. *Nature*, *456*(7218), 98–101. https://doi.org/10.1038/nature07331

Penman, B. S., Pybus, O. G., Weatherall, D. J., & Gupta, S. (2009). Epistatic interactions between genetic disorders of hemoglobin can explain why the sickle-cell gene is uncommon in the Mediterranean. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(50), 21242–21246. https://doi.org/10.1073/pnas.0910840106

Phillips, P. C. (2008). Epistasis - The essential role of gene interactions in the structure and evolution of genetic systems. *Nature Reviews Genetics*, *9*(11), 855–867. https://doi.org/10.1038/nrg2452

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., … Sham, P. C. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*, *81*(3), 559–575. https://doi.org/10.1086/519795

Ranganathan, P., Aggarwal, R., & Pramesh, C. (2015). Common pitfalls in statistical analysis: Odds versus risk. *Perspectives in Clinical Research*, *6*(4), 222.

https://doi.org/10.4103/2229-3485.167092

Rauscher, R., Bampi, G. B., Guevara-Ferrer, M., Santos, L. A., Joshi, D., Mark, D., … Ignatova, Z. (2021). Positive epistasis between disease-causing missense mutations and silent polymorphism with effect on mRNA translation velocity. *Proceedings of the National Academy of Sciences of the United States of America*, *118*(4). https://doi.org/10.1073/pnas.2010612118

Riordan, J. R., Rommens, J. M., Kerem, B. S., Alon, N. O. A., Rozmahel, R., Grzelczak, Z., … Tsui, L. C. (1989). Identification of the cystic fibrosis gene: Cloning and characterization of complementary DNA. *Science*, *245*(4922), 1066–1073. https://doi.org/10.1126/science.2475911

Ritchie, M. D., & Van Steen, K. (2018). The search for gene-gene interactions in genome-wide association studies: challenges in abundance of methods, practical considerations, and biological interpretation. *Annals of Translational Medicine*, *6*(8), 157–157. https://doi.org/10.21037/atm.2018.04.05

Rommens, J. M., Iannuzzi, M. C., Kerem, B., Drumm, M. L., Melmer, G., Dean, M., … Collins, F. S. (1989). *Identification of the Cystic Fibrosis Gene: Chromosome Walking and Jumping*. *245*(4922), 1059–1065.

Ronald, J., Rajagopalan, R., Cerrato, F., Nord, A. S., Hatsukami, T., Kohler, T., … Jarvik, G. P. (2011). Genetic variation in LPAL2, LPA, and PLG predicts plasma lipoprotein(a) level and carotid artery disease risk. *Stroke*, *42*(1), 2–9. https://doi.org/10.1161/STROKEAHA.110.591230

Sackton, T. B., & Hartl, D. L. (2016). Genotypic Context and Epistasis in Individuals and Populations. *Cell*, *166*(2), 279–287. https://doi.org/10.1016/j.cell.2016.06.047

Schork, N. J., Murray, S. S., Frazer, K. A., & Topol, E. J. (2009). Common vs. rare allele hypotheses for complex diseases. *Current Opinion in Genetics and Development*, *19*(3), 212–219. https://doi.org/10.1016/j.gde.2009.04.010

Shabalina, S. A., & Spiridonov, N. A. (2004). The mammalian transcriptome and the function of non-coding DNA sequences. *Genome Biology*, *5*(4). https://doi.org/10.1186/gb-2004-5-4-105

Singh, T. P., Field, M. A., Bown, M. J., Jones, G. T., & Golledge, J. (2021). Systematic review of genome-wide association studies of abdominal aortic aneurysm. *Atherosclerosis*, *327*(May), 39–48. https://doi.org/10.1016/j.atherosclerosis.2021.05.001

Spielmann, M., & Mundlos, S. (2016). Looking beyond the genes: The role of non-coding variants in human disease. *Human Molecular Genetics*, *25*(R2), R157–R165.

https://doi.org/10.1093/hmg/ddw205

Tibaut, M., Naji, F., & Petrovič, D. (2022). Association of Myocardial Infarction with CDKN2B Antisense RNA 1 (CDKN2B-AS1) rs1333049 Polymorphism in Slovenian Subjects with Type 2 Diabetes Mellitus. *Genes*, *13*(3). https://doi.org/10.3390/genes13030526

Trégouët, D. A., König, I. R., Erdmann, J., Munteanu, A., Braund, P. S., Hall, A. S., … Samani, N. J. (2009). Genome-wide haplotype association study identifies the SLC22A3-LPAL2-LPA gene cluster as a risk locus for coronary artery disease. *Nature Genetics*, *41*(3), 283–285. https://doi.org/10.1038/ng.314

Tsimikas, S., Karwatowska-Prokopczuk, E., Gouni-Berthold, I., Tardif, J. C., Baum, S. J., Steinhagen-Thiessen, E., … Witztum, J. L. (2020). Lipoprotein(a) reduction in persons with cardiovascular disease. *New England Journal of Medicine*, *382*(3), 244–255. https://doi.org/10.1056/NEJMoa1905239

Tyler, A. L., Donahue, L. R., Churchill, G. A., & Carter, G. W. (2016). Weak Epistasis Generally Stabilizes Phenotypes in a Mouse Intercross. *PLoS Genetics*, *12*(2), 1–22. https://doi.org/10.1371/journal.pgen.1005805

Van Steen, K., & Molenberghs, G. (2012). Multicollinearity. In *Encyclopedia of biopharmaceutical statistics* (3rd ed.).

Van Steen, K., & Moore, J. H. (2019). How to increase our belief in discovered statistical interactions via large-scale association studies? *Human Genetics*, *138*(4), 293–305. https://doi.org/10.1007/s00439-019-01987-w

Visscher, P. M., Brown, M. A., McCarthy, M. I., & Yang, J. (2012). Five years of GWAS discovery. *American Journal of Human Genetics*, *90*(1), 7–24. https://doi.org/10.1016/j.ajhg.2011.11.029

Visscher, P. M., Wray, N. R., Zhang, Q., Sklar, P., McCarthy, M. I., Brown, M. A., & Yang, J. (2017). 10 Years of GWAS Discovery: Biology, Function, and Translation. *American Journal of Human Genetics*, *101*(1), 5–22. https://doi.org/10.1016/j.ajhg.2017.06.005

Wagih, O., Galardini, M., Busby, B. P., Memon, D., Typas, A., & Beltrao, P. (2018). A resource of variant effect predictions of single nucleotide variants in model organisms. *Molecular Systems Biology*, *14*(12), 1–16. https://doi.org/10.15252/msb.20188430

Wald, A. (1942). Tests of Statistical Hypotheses Concerning Several Parameters When the Number of Observations is Large. *Transcations of the American Mathematical Society*, *54*(3), 426–482.

Wang, D. G., Fan, J. B., Siao, C. J., Berno, A., Young, P., Sapolsky, R., … Lander, E. S.

(1998). Large-scale identification, mapping, and genotyping of single- nucleotide polymorphisms in the human genome. *Science*, *280*(5366), 1077–1082. https://doi.org/10.1126/science.280.5366.1077

Wang, H., Yang, J., Schneider, J. A., De Jager, P. L., Bennett, D. A., & Zhang, H. Y. (2020). Genome-wide interaction analysis of pathological hallmarks in Alzheimer’s disease. *Neurobiology of Aging*, *93*, 61–68. https://doi.org/10.1016/j.neurobiolaging.2020.04.025

Wang, L., Chen, J., Zeng, Y., Wei, J., Jing, J., Li, G., … Zhou, L. (2016). Functional variant in the SLC22A3-LPAL2-LPA gene cluster contributes to the severity of coronary artery disease. *Arteriosclerosis, Thrombosis, and Vascular Biology*, *36*(9), 1989–1996. https://doi.org/10.1161/ATVBAHA.116.307311

Wellek, S., & Ziegler, A. (2009). A genotype-based approach to assessing the association between single nucleotide polymorphisms. *Human Heredity*, *67*(2), 128–139. https://doi.org/10.1159/000179560

Wood, A. R., Tuke, M. A., Nalls, M. A., Hernandez, D. G., Bandinelli, S., Singleton, A. B., … Weedon, M. N. (2014). Another explanation for apparent epistasis. *Nature*, *514*(7520), E3–E5. https://doi.org/10.1038/nature13691

Xing, G., Lin, C. Y., Wooding, S. P., & Xing, C. (2012). Blindly Using Wald’s Test Can Miss Rare Disease-Causal Variants in Case-Control Association Studies. *Annals of Human Genetics*, *76*(2), 168–177. https://doi.org/10.1111/j.1469-1809.2011.00700.x

Xing, G., & Xing, C. (2010). Adjusting for covariates in logistic regression models. *Genetic Epidemiology*, *34*(7), 769–771. https://doi.org/10.1002/gepi.20526

Yazar, M., & Özbek, P. (2021). In Silico Tools and Approaches for the Prediction of Functional and Structural Effects of Single-Nucleotide Polymorphisms on Proteins: An Expert Review. *OMICS A Journal of Integrative Biology*, *25*(1), 23–37. https://doi.org/10.1089/omi.2020.0141

Yuan, W., Zhang, W., Zhang, W., Ruan, Z. B., Zhu, L., Liu, Y., … Zhang, L. F. (2020). New findings in the roles of Cyclin-dependent Kinase inhibitors 2B Antisense RNA 1 (CDKN2B-AS1) rs1333049 G/C and rs4977574 A/G variants on the risk to coronary heart disease. *Bioengineered*, *11*(1), 1084–1098. https://doi.org/10.1080/21655979.2020.1827892

Zeng, L., Moser, S., Mirza-Schreiber, N., Lamina, C., Coassin, S., Nelson, C. P., … Schunkert, H. (2022). Cis-epistasis at the LPA locus and risk of cardiovascular diseases. *Cardiovascular Research*, *118*(4), 1088–1102. https://doi.org/10.1093/cvr/cvab136

Zhang, J., & Yu, K. . (1998). What’s the relative risk? A method of correcting the odds ratio in cohort studies of common outcomes. *Journal of the American Medical Association*, *280*(19), 1690–1691.

Zhang, R., Shen, S., Wei, Y., Zhu, Y., Li, Y., Chen, J., … Christiani, D. C. (2022). A Large-Scale Genome-Wide Gene-Gene Interaction Study of Lung Cancer Susceptibility in Europeans With a Trans-Ethnic Validation in Asians. *Journal of Thoracic Oncology*, *17*(8), 974–990. https://doi.org/10.1016/j.jtho.2022.04.011

Zhong, K., Zhu, G., Jing, X., Hendriks, A. E. J., Drop, S. L. S., Ikram, M. A., … Kayser, M. (2017). Genome-wide compound heterozygote analysis highlights alleles associated with adult height in Europeans. *Human Genetics*, *136*(11–12), 1407–1417. https://doi.org/10.1007/s00439-017-1842-3

Zhou, J., Theesfeld, C. L., Yao, K., Chen, K. M., Wong, A. K., & Troyanskaya, O. G. (2018). Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. *Nature Genetics*, *50*(8), 1171–1179. https://doi.org/10.1038/s41588-018-0160-6

Zhou, J., & Troyanskaya, O. G. (2015). Predicting effects of noncoding variants with deep learning-based sequence model. *Nature Methods*, *12*(10), 931–934. https://doi.org/10.1038/nmeth.3547