Efficient Strategies for Epistasis Detection in Genome-Wide Data

By

Dominic Russ

A thesis submitted to
the University of Birmingham
for the degree of
DOCTOR OF PHILOSOPHY

Centre for Computational Biology
Institute of Cancer and Genomic Sciences
University of Birmingham
November 2022

ABSTRACT

Genome-Wide Association Studies have been carried out with SNP array technology since 2005, identifying thousands of loci for a great many traits and diseases. There are now large data sources, such as UK biobank, that provide medical and genetic data of hundreds-of-thousands of people. However, there is a shortfall in the heritability explained for the phenotypes that have been assessed. One of the explanations for this deficit is interactions between genes, called epistasis, that are not detected and so part of the causation missed. In this thesis, I carry out a comprehensive review of the large number of available epistasis detection tools in the literature. This is followed by a simulation benchmarking study to assess the ability of a representative group of these tools to detect epistatic interactions. From these tools, BOOST, MDR and MPI3SNP found the most interactions in this simulation study. Next, I set out three possible strategies for searching in biobank scale data in order to find a best practices workflow. These were exhaustive searching, an approach tailored to the tools' strengths and by splitting the data into linkage disequilibrium-based haplotype blocks and reducing the computational load. A simulation study was devised that found a mixed approach, using both BOOST and MDR for different types of interactions. The final pipeline initially uses the BOOST algorithm to find pure epistatic interactions and filter out insignificant pairs of SNPs. Those remaining variants with large single-locus effect sizes are assessed with MDR for impure interactions. Those interactions that are identified are assessed for significance, effect size and heritability explained. Finally, validation is carried out across each interacting pair, incorporating numerous sources of $apriori$ knowledge. This was applied to Atrial Fibrillation, Alzheimer's Disease and Parkinson's

Disease, three diseases that have previously been assessed for interactions. Although no statistically significant results were identified, this approach demonstrated an increased amount of heritability explained, showing that some of the missing heritability could be accounted for this way. A downstream analysis method was devised, finding genes in linkage with the interacting loci, applying a number of functional annotations and searching STRING-db for evidence of known interactions. Finally, the study was extended to examine rare variants in rare disease congenital hypothyroidism. As a systemic disorder, it could potentially have pathological interacting mutations. After variant calling, four $denovo$ variants were identified, potentially explaining the condition. Six related interactions were found, with one not present in the parents, so possibly explaining the condition. The mutations, present in TG and $PDIA4$ have evidence of an interaction in STRING-db and both being involved in thyroid hormone synthesis in the KEGG database. These contributions provide a novel, tested pipeline for identifying epistasis from GWAS data, as well as a corpus of simulated data for future researchers. A robust methodology is applied for testing resulting interactions statistically, as well as an approach for validating interactions by incorporating numerous data sources to find significant commonalities between variants.

DEDICATION

This thesis is dedicated to Christopher Robert John, my inspiration.

ACKNOWLEDGMENTS

Firstly, I would like to thank Georgios Gkoutos and John Williams for their enormous help, support, and encouragement. Without your assistance, this wouldn't be possible. The Gkoutos group has been incredibly helpful and supportive throughout this course, with some remarkable scientists and people to work with. So, a massive thank you to my fellow PhD students in Laura and Victor and more senior members of the group in Andreas, Animesh, and Luke. This also extends to the folks in the CCB, JB and staff, as well as Ben.

I would also like to thank Paul Schofield, Robert Hoehndorf and teams. Your help when working with the DDD data has been invaluable while improving my sequencing skills. I would also like to thank the King Abdullah fund, which part funded this PhD.

Finally, I need to thank Sophie, Mogwai and the Yardley family, who have been by my side supporting me through this course. My friends from Harpenden, Brighton and Southampton. Without you all, I don't know how I'd cope! Also, my family who have provided me with the advice and support to be able to do any of this.

During the course of this PhD I have been involved in the following publications:

John, CR. Watson, D. **Russ, D.** Goldmann, K. (2020) M3C: Monte Carlo reference-based consensus clustering. Scientific reports https://doi.org/10.1038/s41598-020-58766-1

Aziz, F. Cardoso, V. Bravo-Merodio, L. **Russ, D.** Pendleton, S. Williams, J. Acharjee, A. Gkoutos, G. (2020) Biomarker Prioritisation and Power Estimation Using Ensemble Gene Regulatory Network Inference. International Journal of Molecular Sciences https://doi.org/10.3390/ijms21217886

Aziz, F. Cardoso, V. Bravo-Merodio, L. **Russ, D.** Pendleton, S. Williams, J. Acharjee, A. Gkoutos, G. (2021) Multimorbidity prediction using link prediction. Scientific reports https://doi.org/10.1038/ 021-95802-0

Acharjee, A. Hazeldine, J. Bazarova, A. Deenadayalu, L. Zhang, J. Bentley, C. **Russ, D.** Lord, J. Gkoutos, G. Young, S. Foster, M. (2021) Integration of Metabolomic and Clinical Data Improves the Prediction of Intensive Care Unit Length of Stay Following Major Traumatic Injury. Metabolites https://doi.org/10.3390/metabo12010029

Williams, J. **Russ, D.** Cardoso, V. Bravo-Merodio, L. Pendleton, S. Aziz, F. Acharjee, A. Gkoutos, G. (2021) A Causal Web between Chronotype and Metabolic Health Traits. Genes https://doi.org/10.3390/genes12071029

Bravo-Merodio, L. Acharjee, A. **Russ, D.** Bishta, V. Williams, J. Tsaprounie, L. Gkoutos, G. (2021) Chapter Four - Translational biomarkers in the era of precision medicine. Advances in Clinical Chemistry https://doi.org/10.1016/bs.acc.2020.08.002

Russ, D. Williams, J. Cardoso, V. Bravo-Merodio, L. Pendleton, S. Aziz, F. Acharjee, A. Gkoutos, G. (2022) Evaluating the detection ability of a range of epistasis detection methods on simulated data for pure and impure epistatic models. PLOSONE DOI: https://10.1371/journal.pone.026339

Contents

List of Figures

List of Tables

Chapter One

Background

1.1 Thesis Overview

This thesis concerns the discovery of genetic interactions, known as epistasis, a concept that has evolved since its initial observation just after the turn of the 20th Century. The purpose of this research is primarily to assist in the search for additional understanding of common inheritable diseases, with a view to providing targets for potential therapeutic remedies or biological markers that can be used to forewarn of future complications. Through this introduction, I aim to provide the general background information required to be the basis for future chapters and the basic methodologies used.

Within the subject of genetics, epistasis was discovered early in the embryonic stages of the founding of what is now a cornerstone of biological study. It is a concept that formed a part of the development of statistics and population genetics, as they became key driving forces in our understanding of evolution, the aetiology of traits and the complex networks of molecular actors that interact within the cell and beyond. Arriving at the 21st Century, with the completion of the Human Genome Project (HGP), the future of genetics seemed set on translating the genetic sequence directly into medical advances and the solution of genetic traits.

It is inarguable that HGP has changed genetic studies monumentally, sequencing is now commonplace and relatively inexpensive. This has been automated, opening enormous avenues of research and moving researchers into roles analysing genetic data rather than laboriously capturing genetic sequences (Gibbs, 2020). However, numerous complicating factors have been recognized, such as those posed by epigenetics, when a gene's expression can be repressed by blocking transcription factors from binding, so that the protein it codes for is not synthesized. The dynamics of RNA in the nucleus, from competing non-coding RNA to transcription factors. Another, is the complex web of genetic and protein-protein interactions within the cell giving rise to a trait. It is here that the search for genetic interactions can help elucidate the functions within the cell, metabolic pathways that are involved in a trait and combinations of mutations that give rise to clinical conditions.

This thesis begins with a detailed background of the area of study, including the key discoveries that defined epistasis and how it was viewed through the 20th Century. It then introduces modern studies of genetics and how they came to be developed, with an overview of the procedures carried out in order to ensure the quality of results. Finally, the methods of interpretation of these results are discussed.

The second chapter focuses on the classification of different types of epistasis. There is a review of the current field of epistasis detection tools. These are then compared using a small-scale simulation study to find which is most accurate and an assessment of the resources required. The third chapter builds upon this by examining strategies to search for epistasis in large datasets, of the type that are currently being studied for genetic associations. The best strategy is applied to the conditions of atrial fibrillation, Parkinson's Disease and Alzheimer's Disease. Finally, in the fourth chapter the focus is on looking at sequencing data for the rare disease of congenital hypothyroidism, as appears in a cohort of children with intellectual disabilities and uses *a priori* data to find interacting genes related to their condition.

The aims of this thesis are as follows:

- Assess the field of epistasis detection methods available.
- Develop strategies of applying the most effective tools to the latest, large-scale datasets.
- Apply these methods to real examples, with a view to establishing the prevalence of genetic interactions within the formation of traits.
- Quantify the heritability attributable to epistatic interactions.
- Demonstrate the utility of epistasis detection for diseases, including rare ones.

Contributions to the field are:

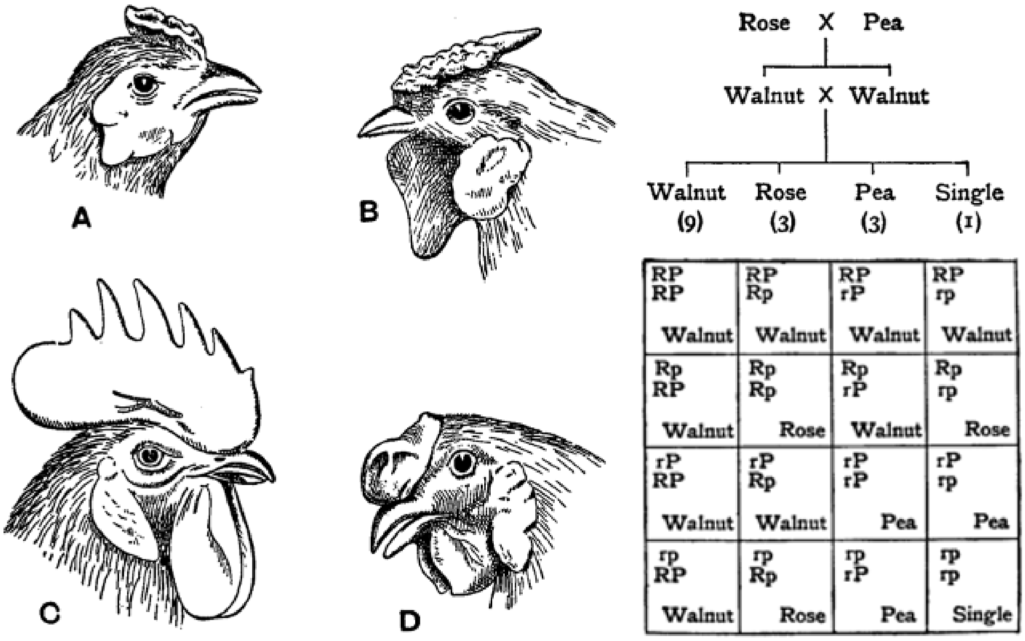
- Chapter 2 provides the most comprehensive review and benchmarking assessment of epistasis detection methods to date, with a varied simulated test dataset.
- Chapter 3 is the first simulated study for an end-to-end workflow study for epistasis detection, outside of simple exhaustive approaches. Also carried out are the largest genome-wide epistasis detection studies for Atrial Fibrillation, Alzheimer's Disease and Parkinson's disease.
- Chapter 4 isolates an unstudied group with a rare disease and proposes not previously found potential, explanatory mutations for their condition.

1.2 Origins of Epistasis

Epistasis was one of a number of terms defined by William Bateson when genetics was in its infancy, including 'genetics' itself. Derived from the Greek, with prefixes *epi* meaning 'upon' or *hypo* meaning 'under' and the suffix *stasis* meaning stopping, his definition was the suppression of a genetic element upon or by another element. This discovery built upon his previous work cataloguing discontinuous traits and his recognition of the importance of Mendelian inheritance (W. Bateson, 1894; W. Bateson et al., 1909).

His research, with Punnett and Saunders examined a variety of plants and animals, focusing on traits' adherence to the Mendelian model. An early example of interest was the comb of poultry. In the case of rose and pea combed fowl, the resulting progeny only exhibited walnut combs. However, when these walnut combed fowl are bred, the resulting progeny have one of the four phenotypes in a ratio of 9:3:3:1 of walnut:rose:pea:single (Figure 1.1). The Punnett Square is used to demonstrate the Mendelian relationships of dominance and recessivity, in order to understand how those ratios are produced. There are two alleles described as R/r and P/p, with a demonstrative interaction between the two factors. Bateson explains this with the 'Presence and Absence' hypothesis, in which the alleles represent rose/no rose and pea/no pea. When the fowl has both rose and pea, they exhibit a walnut comb. Conversely, in the absence of the rose and pea alleles, a single comb is produced (Punnett, 2009; W. Bateson, 1905).

It was in his 1909 book that Bateson expanded his definition of genetic interactions beyond this hypothesis. Examining the case of the colour of rabbits' fur. Albinism is observed as a recessive trait to grey fur, so a cross of coloured rabbits by albino rabbits produces first all grey rabbits, with the following generation normally yielding a ratio of 3 grey to 1 albino, as expected under a model of Mendelian Inheritance. However, there were some rabbits bred in which an exception was observed. Here, the second generation also resulted in black rabbits as well, in the ratio 9 grey to 3 black and 4 albino. Since the black phenotype only existed in some experiments, this was taken to be a separate genetic element to grey. This means there are three genetic elements; absence or presence of colour; greyness; and blackness. However, in the presence of greyness, the black colour was suppressed (Figure 1.2. He described grey fur as being epistatic to black fur, with the reverse situation being called hypostatic (W. Bateson et al.,



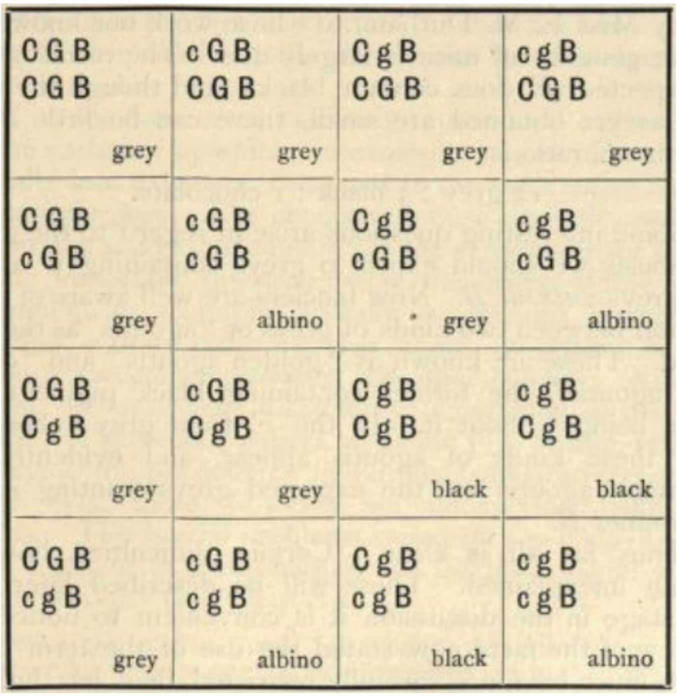
RP RP Walnut	RP Rp Walnut	RP rP Walnut	RP rp Walnut
Rp RP Walnut	Rp Rp Rose	Rp rP Walnut	Rp rp Rose
rP RP Walnut	rP Rp Walnut	rP rP Pea	rP rp Pea
rp RP Walnut	rp Rp Rose	rp rP Pea	rp rp Single

Figure 1.1: Punnett's illustrations of the experiments regarding poultry comb classification. Of the four fowl - A, pea; B, rose; C, single; D, walnut. The tree diagram shows the experimental procedure. Fowl with the rose and pea comb type were bred, resulting in fowl with a walnut comb. When the walnut comb fowl were bred, the progeny displayed four different each of the four comb phenotypes, with ratios in brackets underneath. The table shows a Punnett Square, illustrating the allelic combinations found (Punnett, 2009).

1909).

1.3 Development of Epistasis in Population Genetics

The early decades of genetics include contributions that are still in use today and deserve some discussion. R.A. Fisher, was responsible for many statistical innovations, such as his Exact test, variance, and maximum likelihood estimation (Efron, 1998). He was immersed in biological research throughout his career. In his early days at Cambridge, he addressed an issue of interest



CGB CGB grey	cGB CGB grey	CgB CGB grey	cgB CGB grey
CGB cGB grey	cGB cGB albino	CgB cGB grey	cgB cGB albino
CGB CgB grey	cGB CgB grey	CgB CgB black	cgB CgB black
CGB cgB grey	cGB cgB albino	CgB cgB black	cgB cgB albino

Figure 1.2: Punnett Square showing the different genotypes of rabbit fur colour. C/c refers to colouration, G/g greyness and B/b blackness in which the uppercase allele indicates presence and the lowercase represents absence. The phenotype present is written in the lower, right corner of each square (W. Bateson et al., 1909).

to Bateson, reconciling Mendelian discontinuous traits with quantitative continuous traits in terms of evolution. His solution was that additivity of many discrete Mendelian factors could lead to a continuous outcome (Piegorsch, 1990). It is from this contribution that we see his interpretation of epistasis. Here he presents ‘epistacy’ as a deviation from additivity for a quantitative phenotype (Fisher, 1919).

Another influential figure was Sewall Wright. His Shifting Balance Theory attempted to explain evolution by observing generations of livestock. He had been observing many traits through experiments and creating tree diagrams to show how genotypes were mapped (Wright, 1920). Taking the lower estimate at the time, of 1000 genes per higher organism and assuming 10 ’allelomorphs’ or alleles of each gene, he calculated 10^{1000} possible genetic configurations.

Thus, even at the lower boundary, enormously complex networks of genetic combinations were possible. In this very highly dimensional space, he foresaw different combinations of genotypes lending more or less beneficial adaptions. He imagined this in the form of a landscape with peaks and troughs of fitness. Within this landscape, he envisioned that alleles of a gene that may be helpful in combination with alleles of another gene, could be detrimental when combined with certain alleles of a different gene (Wright, 1932).

Wright and Fisher were in correspondence during this time, and their disagreements culminated in what is known as the Fisher-Wright Controversy. Leaving aside other parts of this discussion, essentially Fisher viewed evolution as being driven by many mutations with small effects over a large population. However, Wright saw genetic combinations as pervasive, with evolution driven towards mutations that favoured a systemic improvement in fitness of the organism via complex biochemical interactions (Skipper, 2009). Assessing their insights from the present day, it is remarkable how their thoughts now complement current discoveries, with studies often finding many mutations with small effects leading to a trait and also, the complex webs of protein-protein interactions that exist within biological systems.

As a result of these debates and innovations over time, a shifting definition of epistasis has occurred. There are three factors that need to be considered. Firstly, the outcome variable makes some difference in how we conceptualize the interaction. The main categories are continuous phenotypes, such as height, and another for a binary trait, often presence or absence of a clinical condition. Secondly, there is the model of epistasis. As we saw in Figure 1.2, Mendelian relationships can occur between genotypes in epistatic relationships. However, we must consider all possible genotypes of a pairwise interaction - AA, Aa, aa and BB, Bb, bb. Combining these gives nine possible genotypes. Analysis of all possible conformations of these genotypes using a combinatorial model of point configurations using the package TOPCOM reveals there to be 387 possible models, made up of 69 symmetrical classes. This includes combinations of Mendelian inheritance as well as less obvious relationships. Finally, effects can be dominant, additive, and multiplicative in their influence over the dependent variable (Hallgrímsdóttir et al., 2008).

1.4 Complex Disease and Genetic Modelling

1.4.1 Complex Disease

Complex diseases are polygenic and involve multiple factors, including environmental drivers. Complexity arises from the systemic biochemical networks and physiological systems coordinating in such a way to bring forth a given phenotype. The human population has a relatively long history, adapting to many environments and having undergone numerous striations. This has led to a diverse range of genetic and environmental influences and evolutionary pressures. The result is heterogeneity, with multiple causes for the same trait. From an additive perspective, there is a view that an individual can inherit a number of genes or have mutations which, above a threshold, will lead to a condition. Different mutations can collectively be risk factors, leading to pervasive heterogeneity. Then there are interaction effects, such as gene x gene or gene x environment interactions, with some genes potentially leaving an individual vulnerable to certain environmental triggers. Another factor is the life cycle impacting or triggering a trait, such as developmental stages like puberty or ageing with wear and tear, predisposing a person to a particular phenotype (Schork, 1997).

However, there are further layers of entanglement due to factors which affect genetic expression and regulation. From an environmental perspective, epigenetics can play an important role in gene expression as a result of cellular activities or external drivers (Lehnen et al., 2013). DNA methylation, histone modification and chromatin structure can cause changes in gene expression. An example is how gestational malnutrition and overnutrition can lead to di-

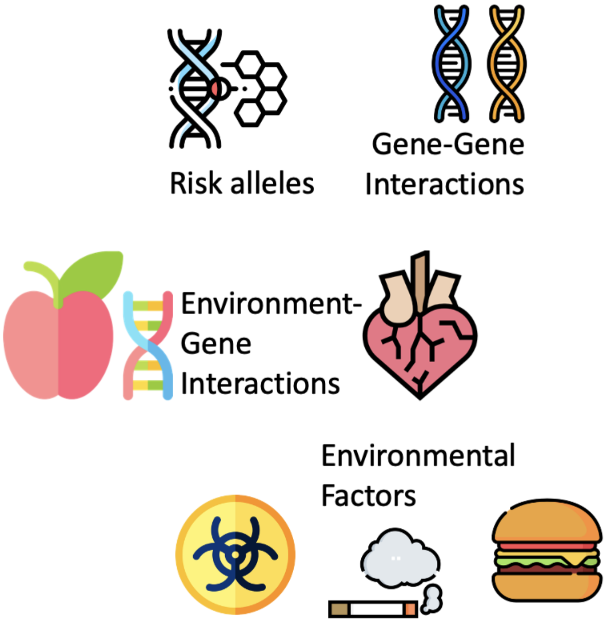


Figure 1.3: Graphic to show the various factors affecting complex diseases

abetes later in life. Within the considerations of transcription, RNA can be spliced in different ways and short non-coding RNA can affect the translational rates of other RNA and is implicated in a wide variety of disorders (Cooper et al., 2009). RNA splice sites are coded within the DNA sequence. Additional to other regions such as transcription factors, they form a network of gene regulatory structures. Thus expanding genetic complexity beyond just gene coding regions (Szathmáry et al., 2001). Altogether, this provides many lines of investigation into the drivers and networks of factors contributing towards heritable disease.

1.4.2 Genetic Variation in Humans

The human genome is stored in standardized builds, with the latest release from the Genome Reference Consortium (GRC) being GRCh38 (submitted 2013/12/17) (*GRCh38* n.d.). Whilst this has been surpassed in coverage by the Telomere-to-Telomere Consortium (T2T), it is still the reference build in most use, so we will focus on the GRC build (Nurk et al., 2022). It spans 3,099,734,149 base pairs across all chromosomes and mitochondrial DNA (Yan Guo et al., 2017). Genetic annotations have been documented by the GENCODE consortium, with the latest version being GENCODE 41. Genetic features are annotated manually and with aid of computational tools or experimental data. From this work, they have identified 19,804 protein-coding genes, 25,134 non-coding genes and 15,240 pseudogenes, with the latter two coding for RNA, that can serve a function and inactive, archaic genes respectively (Frankish et al., 2021).

Variation in the human genome comes in a range of different forms. The simplest is the point mutation, known as a single nucleotide polymorphism or variant (SNP/SNV). The database dbSNP contains all known SNPs across all populations. Its most recent version, build 155, has recorded 660,773,127 SNPs, known as a RefSNP and designated a number starting with 'rs' (Sherry et al., 2001). There are additional structural variants, such as insertions and deletions (InDels) of base pairs to the genome. It is worth mentioning that shorter examples of these have been included in the dbSNP count of SNPs, with verified larger variations being stored in dbVar. There are many forms of structural variation, a more common type are copy number variations (CNVs). These are short repeating sequences that can extend for thousands of base pairs (Lappalainen et al., 2013). A more extreme type of structural variation is the loss or gain of an entire chromosome, known as aneuploidy. This kind of variation is the leading genetic cause of miscarriage and congenital birth defects (Hassold et al., 2007).

For any one mutation, irrespective of the type, its presence in the population relies upon a few factors. The mutation can be a germline or somatic mutation. Indicating that it was inherited from parental gametes or occurred within the individual's life, respectively. In common diseases, the mutation is likely to be a germline mutation that has been transmitted through generations, as such, the mutation cannot be so severe that it inhibits the chances of that happening. This relates to the penetrance of the mutation, the amount of negative effects imbued from the genetic defect (Donaldson et al., 2015). Mutation rates in germline cells, have been

	Rare Disease	**Common Disease**
Genetic Architecture	Monogenic	Polygenic
Penetrance	High	Normally low
Location	Coding regions	Mixed coding and non-coding
Discovery	Family studies	Large population and association studies

Table 1.1: Comparison of common and rare diseases (Frayling, 2014)

estimated to happen at a rate of approximately 64 per generation, with more found in the male gamete (Drake et al., 1998). Mechanically, mutations also occur at particular hotspots, given that more point mutations and indels are clustered within repetitive DNA sequences. This is theorized to happen as these regions can stall the enzyme DNA polymerase (McDonald et al., 2011). So, we can see certain regions are more prone to mutation. Since those that are more damaging are less likely to remain in the population, there exists the relationship seen in Table 1.1. Monogenic diseases have single, high penetrance risk alleles with rare mutations and the opposite being true for common, complex diseases (Frayling, 2014).

1.4.3 Quantifying Heritability

Since complex diseases have many different drivers, it is useful to estimate how much of the cause is genetic and how much environmental. For this, we use the metric known as **Heritability**. This measure is defined as the degree of the phenotype's variance that can be explained by variance of genetic factors. Here, the variance of the phenotype is equal to the sum of the variance caused by genotype, environment, and interactions between these. The genetic variance can be further broken down to the sum of the variance of additivity, dominance and genetic interactions. There are two types of heritability. Broad-sense heritability, H^2, is the proportion of phenotypic variance caused by genetic variance, given as:

$$H^2 = \frac{V_G}{V_P} \tag{1.1}$$

with V_G being genetic variance and V_P as phenotypic variance. Narrow-sense heritability, h^2 is the proportion caused by additive effects:

$$h^2 = \frac{V_A}{V_P} \tag{1.2}$$

With V_A as variance attributable to genetic additive effects. It is important to note that heritability varies from population to population, so is not absolute and cannot be transplanted. It also does not quantify the proportion of the phenotype that is caused by genes, since it is based on variance. There can also be some confusion regarding familial traits. Family's share genes but often share the same environment and that can play a role in related individuals having certain outcomes (Russell, 1998).

1.4.4 Genetic Linkage

The theory that inheritance happened through the chromosome was developed in 1903, establishing that genetic material could be transmitted in groups (Sutton, 1903; Baltzer, 1964). It was Bateson et al. who first observed linkage between inherited traits. In their experiments with sweet pea between flower colour and pollen grain length, they observed that the phenotypes were transmitted together (W. Bateson et al., 1905). Later, in experiments with thousands of *D. Melanogaster*, Morgan was able to observe that eye pigmentation and vestigial wing inheritance were for the most part inherited together. There were cases in which this linkage did not occur, however. He theorized this was due to crossing over of chromosomes in meiosis and the loss of linkage, potentially indicating the distance between genes (Morgan et al., 1915). Now known primarily as the recombination rate, this process also bears his name as its unit of measurement, normally in centimorgans (cM). His student, Sturtevant, first demonstrated that genetic elements were linked in a linear fashion, in terms of percentage of 'cross-overs' between different elements (Sturtevant, 1913). He was able to create maps of these linked traits, with the recombination rate used to demonstrate the distance between the genetic elements.

Haldane devised a mapping function, since named for him, to better estimate distances from data on recombinations. His formula uses the natural logarithm to correct for unobserved recombination sites based on a Poisson distribution, allowing for more accurate genetic maps. It is also here that he introduces the measurement of distance as cM (Haldane, 1919).

It was Morton who developed a method to better understand linkage in *Homo Sapiens.* He recognized the limited sample sizes available in human experimentation, when compared to the possibilities allowed by growing sweet pea or breading *D. Melanogaster.* He was able to develop the logarithm of the odds (LoD) score as a more powerful metric for estimating linkage within families. This uses pedigree information within a family and takes the logarithm of the probability of the sequence with linkage, divided by the probability without linkage (Morton, 1955). By these techniques, researchers have been able to identify genetic variants leading to Huntingdon's disease and the mutations in the BRCA genes involved in breast cancer (MacDonald et al., 1992; Hall et al., 1990).

1.4.5 Linkage Disequilibrium

The principle of **linkage disequilibrium** (LD) was introduced by Lewontin and Kojima (1960). This describes how linkage appears across multilocus polymorphisms. In the absence of linkage and given random mating, the alleles of two loci should appear as probabilistic functions of their allele frequencies. However, in the presence of linkage between the two sites, this breaks down into disequilibrium. They provide a metric denoted D to demonstrate LD at two loci, A and B with genotypes of g. The minor allele frequency (MAF) of A is p and for B is r.

$$\hat{D} = g_{AB} \cdot g_{ab} - g_{Ab} \cdot g_{aB} \tag{1.3}$$

$$= p \cdot r \times (1-p)(1-r) - p(1-r) \times r(1-p) \tag{1.4}$$

It is expected that in the case of linkage equilibrium, this number will be 0. As a metric, it is somewhat limited when it comes to making comparisons between different pairs of loci, given that the range of values is dependent on the allele frequencies present. In order to standardize it, the resulting value is divided by the maximum possible absolute value with the given allelic frequencies. This is known as D' (D prime) and has ranges between -1 and 1 (Lewontin, 1964). However, the D' value performs poorly in cases of low allele frequency, as the outcome remains small. It also is not a good measure to predict from one locus to another. To overcome both of these issues, the squared correlation coefficient is used, denoted r^2 that extends D and its parameters:

$$r^2 = \frac{D^2}{p\left(1-p\right) r\left(1-r\right)} \tag{1.5}$$

In this case, the correlation is calculated, yielding values ranging from 0 to 1. A score of 1 indicates that both loci provide the same information (W. G. Hill et al., 1968; Laird et al., 2011).

1.4.6 Genome-Wide Association Studies

Following the elucidation of the structure of DNA and efforts to sequence parts of the chromosome, a greater granularity in the study of genetic material was realizable. It is at this point that it is possible to identify a mutation as a change in nucleic acid present on the DNA strand. This gave rise to the association study, in which a small part of the DNA was sequenced for a cohort of unrelated individuals. This allowed researchers to break free of the limitations of pedigree based, linkage studies and use more samples in a more generalizable, large population. This was typically done by sequencing a part of a gene and then carrying out a simple test, such as a χ^2 test (Braude et al., 1986; Bell et al., 1984; Svejgaard et al., 1994).

These tests often had problems with reproducibility and false positive results, with larger sample sizes and confounding factors having the potential to overcome these issues (Hirschhorn et al., 2002). Risch et al, in 1996, demonstrated that association analysis had more statistical power than linkage studies. They suggested that the future of genetic studies would be in much larger scale versions, with many more samples and across the whole genome. By using LD to infer some loci, it could be possible to carry out such tests at scale, given technological improvements. This was the birth of the concept of the genome-wide association study (GWAS). It is also here that we see the now commonplace, significance threshold for GWAS set at $5e-8$ as a Bonferroni corrected estimate for the number of genetic markers estimated to be needed for thorough coverage that is still used (Risch et al., 1996).

Data Collection

SNP Microarrays

DNA probes were developed with complementary sequences in 1979 to demonstrate the presence of specific genotypes via hybridization, the process of two complementary sequences of DNA binding (Wallace et al., 1979). The later development of the Polymerase Chain Reaction to amplify volumes of DNA was also a major milestone (Saiki et al., 1985). From these roots, SNP Microarrays were introduced, with the first from Affymetrix composing of 135,000 probes, complementary to sequences in the human Mitochondrial genome (Chee et al., 1996). This larger scale, parallel identification of specific genetic material started the process of reaching the technology required for GWAS (Landegren et al., 1998).

Capturing a genome in its entirety at this time was prohibitively expensive, with the Human Genome Project running for over a decade at a cost of around $3 billion (*Human Genome Project Fact Sheet* n.d.). Instead, SNPs had been identified as a major source of genetic variability and more straightforward to capture. The basic approach is to have hundreds-of-thousands of immobilized oligonucleotide sequences about 25 base pairs in length. The captured DNA is cut into small sequences by enzymatic action, the sequences that complement the probes hybridize with them. An additional nucleotide base is added to the sequence with a fluorescent tag, colour dependent on which base. The light emitted is recorded to indicate the genotype at that location (Orntoft et al., 2006; McGall et al., 2002; Kallioniemi, 2001). As of 2018, the cost per sample of conducting SNP capture for a human stood at $28-$95 (You et al., 2018).

Next-Generation Sequencing

Next-Generation Sequencing (NGS) allows for a greater resolution of genetic data capture and the collection of rare variants. It is a more expensive option compared to SNP Microarrays, with whole-genome sequencing (WGS) costing €1669 and whole-exome sequencing (WES) costing €792 as of 2016 (Nimwegen et al., 2016). The process is achieved using a number of methodologies. Illumina is the major producer of these machines. Initially, a process called bridge amplification is used, in which short reads of DNA are repeatedly copied whilst being held by a ligated adapter to a glass surface with short complementary oligonucleotides. Following this, a process called sequencing by synthesis creates complementary sequences to those captured on the glass using fluorescently labelled modified bases. As these are added, they release a coloured light that is detected, similar to how microarrays work (Slatko et al., 2018). At this stage, the data is a collection of short strings of base readings. These are aligned against the reference, with multiple overlapping reads giving depth of coverage at any one point. At this stage, algorithms can be used to assess the genotype along the genome in a process called variant calling, usually generating a variant call format file (VCF) (Muzzey et al., 2015).

Early GWAS

The early stages of GWAS were dependent upon the discovery of locations of SNPs as well as details around the LD structure of the genome. These needs were provided for by the Human Genome Project and the HapMap Project (Lander et al., 2001; Venter et al., 2001; Altshuler et al., 2005). The HapMap Project extended the work undertaken in the Human Genome Project by sequencing people of different ethnic backgrounds to better understand the variation present in different genomes, but also to map the LD patterns and how they appear as haplotypes. Using this information, SNP arrays could be better targetted to provide as much information on an individual's genome as possible, using inference by LD (Ikegawa, 2012).

As the technique was developed and popularized, there were mistakes made along the way that demonstrated the need for rigorous quality control. The most high profile was the Wellcome Trust Case Control Consortium (WTCCC) (Winzer et al., 2006). This was the largest GWAS to date, but the procedures that were used to carry out the sample collection and analysis led to bias. Different labs collected and processed case and control data at different geographical sites and with different array chips. This led to an enormous number of the loci generating tiny p-values when they were, in fact, false positives (Lambert et al., 2012).

As well as these problems inherited from new experimental procedures and equipment, there are sources of error that had been previously identified in association studies a decade earlier. One issue is contamination during the extraction and processing steps that can be detected at later stages (Jun et al., 2012). There are also the problems of population stratification and cryptic relatedness. The former refers to the difficulties arising from different ancestral haplotypes. If there is an ethnically admixed cohort and an imbalance between cases and controls, then this can give rise to significant signals that are not due to the trait (Hirschhorn et al., 2002). Cryptic relatedness refers to unknown relatedness within the cohort that can also skew significance tests (Voight et al., 2005).

GWAS Procedures

At this stage, the process of carrying out a GWAS is, essentially, standardized. QC is of critical importance, before any testing can be applied. The steps can be broadly split into those that assess the data on a SNP-by-SNP basis and those that look at the individual samples. For both, tests are carried out for the proportion of missingness, since this can uncover problems with a probe in the array chip or a poorly prepared sample (Marees et al., 2018).

Marker QC

Finding SNPs that could have been affected by some kind of artefact in the equipment used, or some other factor, is important to avoid skewed results. The Hardy-Weinberg Equilibrium (HWE) is a theorem that was developed by two mathematicians regarding the allele frequencies within a population. It had previously been thought that under Mendelian inheritance, recessive alleles should reduce in frequency with the population. However, this has been observed to be untrue. In HWE, it was shown that allele frequencies, p and q remained stable across generations, as in the equation:

$$1 = p^2 + 2pq + q^2 \tag{1.6}$$

In cases where the allele frequencies present don't conform, it is therefore possible that something has gone wrong with the collection of this data. The test has traditionally been carried out using an exact test, however other methods have also been developed (Weinberg, 1908; Hardy, 1908; Emigh, 1980; Wigginton et al., 2005). A threshold for the significance of the departure from HWE is used in order to remove those markers that have allele balances that are too distant from the expected balance. The threshold used differs between studies from around $p < 1e-5$ to a popular GWAS tool's manual, in which it is stated that a cut-off of $p < 1e-50$ should filter out serious genotyping errors (Coleman et al., 2016; *Input filtering - PLINK 1.9* n.d.).

Another commonly used marker filter is MAF. Part of the reason for this is simply having the statistical power to reach a significant p-value with the given number of samples accurately. Statistical power is a measure of the probability that any one significant result is a true positive. This is dependent on the effect size, sample size and the accuracy of the apparatus, however a value of 0.05 for MAF has commonly been used (Sham et al., 2014; Robert J Klein, 2007). Recent research has shown a reduction in the quality of SNP array data calls for MAFs of less than 0.01 (Van Hout et al., 2020).

Sample QC

Sex Chromosome and Heterozygosity

A useful test that is usually implemented takes us back to Wright. He was interested in the quantification of inbreeding in a population and introduced his inbreeding coefficient, F (Wright, 1922; Wright, 1950). Since heterozygosity is less common under inbred conditions, this is the focus of the statistic. It can be calculated in GWAS:

$$\hat{F} = 1 - \frac{O(Het)}{E(Het} \tag{1.7}$$

Taking the observed number of heterozygous SNPs divided by the expected number under HWE at each locus (Laird et al., 2011). This operation can also be carried out to test if the X chromosome material is as expected in males and females. Standard thresholds expect an estimate of F less than 0.2 for females and greater than 0.8 in males (Purcell et al., 2007).

An indicator that is used to indicate either inbreeding or contamination is the rate of heterozygosity in the individuals. As previously stated, low heterozygosity is a sign of potential inbreeding, whereas high heterozygosity is likely due to a mix of different genomes, as in contamination. This can be calculated using the method-of-moments and estimating the f coefficient:

$$\hat{f} = \frac{O(Hom) - E(Hom)}{n_{obs} - E(Hom)} \tag{1.8}$$

The expected counts are based on MAFs and HWE and n_{obs} is the total count of observations (Purcell et al., 2007; Weir et al., 1984). Using this value, samples tend to be filtered out when they fall beyond three standard deviations in either direction from the mean, excluding potentially 0.3% of the population as extreme outliers (Reid, 2010).

Identifying Relatedness and Population Stratification

These two potential confounding factors are important to control for, since failure to do so could lead to many false positive results. This would occur as a result of shared genotypes owing to ascendents or ethnic genotypes that are irrelevant to the trait of interest. However, this also means that these properties can be derived using the genetic data. There have been a number of different approaches suggested for dealing with these, but here we will highlight the most prevalent (Sillanpää, 2011).

Both procedures require some preprocessing of the genetic data. They are carried out under the assumption of linkage equilibrium and as such, those in high LD are pruned out of the dataset. This can be achieved by using *a priori* knowledge of particular regions, testing for LD across the data or both (Weale, 2010).

In order to quantify relatedness, cryptic or otherwise, there are a number of methods available. There are two key terms, identity-by-state (IBS), which refers to a particular allele in an individual that can be compared with another at the same locus and identical-by-descent (IBD), which refers to the sharing of an allele as a result of inheritance from a shared ascendant (Thompson, 1974). A well established metric for relatedness is the kinship coefficient, the probability that a randomly sampled allele on an autosomal chromosome in two individuals is IBD (Sonesson et al., 2005). It is represented by ϕ as a function:

$$2\phi = \frac{\pi_1}{2 + \pi_2} \tag{1.9}$$

where π indicates the probability that the individuals share either 1 or 2 identical states. In GWAS data, heterozygous alleles are impossible to derive since the chromosome each allele is on is not clear and so they are not equivalent. In the software KING this is estimated for GWAS data with the kinship coefficient:

$$\phi = \frac{N_{Aa,Aa} - 2N_{AA,aa}}{N_A a^i + N^j_{Aa}} \tag{1.10}$$

Where N is the total number of SNPs with a certain genotype, $N_{Aa,Aa}$ when both samples are heterozygous, $N_{AA,aa}$ the total number when each sample is homozygous for the opposite allele and $N_A a$ the total number of heterozygous loci for sample i and j.

The program PLINK uses IBD estimation to quantify relatedness, π. This is calculated using $P(Z = z)$ were Z is the IBD state for the entire genome. In turn, these are derived from the count of N loci that have equal IBS of I:

$$P(Z=0) = \frac{N(I=0)}{N(I=0|Z=0)} \tag{1.11}$$

$$P(Z=1) = \frac{N(I=1) - P(Z=0) \times N(I=1|Z=0)}{N(I=1|Z=1)} \tag{1.12}$$

$$P(Z=2) = \frac{N(I=2) - P(Z=0) \times N(I=2|Z=0) - P(Z=1) \times N(I=2|Z=1)}{N(I=2|Z=2)} \tag{1.13}$$

$$\pi = \frac{P(Z=1)}{2} + P(Z=2) \tag{1.14}$$

In each of the previous cases we can assess the relationship, however slightly confusingly, the outcomes are presented slightly differently. The score derived by the Kinship coefficient is equivalent to half of the PLINK IBD assessment (Table 1.2) (Purcell et al., 2007). Often, the next step is to find any relationships from third degree relatives (first cousin, great grandparents, great uncle/aunt) and closer (Anderson et al., 2010). Commonly, the individual with the least missing data or those without the trait being studied are removed. This is to maintaining statistical power, those with more missing data are not included in tests for missing loci and typically cohorts have more controls than cases, so removing controls is preferable.

Finding a population structure within the genetic data is a less precise process, due to admixture of populations over history. Principal component analysis (PCA) is a technique that was introduced to genetics to map human evolutionary history across a selection of loci

Relationship	Kinship Coefficient	PLINK IBD
Parent-Offspring	0.25	0.5
Full Siblings	0.25	0.5
Half Siblings	0.125	0.25
First Cousins	0.0625	0.125
Second Cousins	0.0156	0.0625

Table 1.2: Relatedness measures (Lange, 1997; Anderson et al., 2010)

(Menozzi et al., 1978). It is a technique that can be used for large multivariate datasets in order to reduce the number of dimensions by searching for the maximum possible variance. The direction of most variance is the first principal component (PC) or eigenvector. The second PC is the maximum variance at a right angle to the first PC. This can continue with each subsequent PC explaining less genetic variance. Tools such as SMARTPCA can programmatically find outliers that could be indicative of a very different ancestry group. Alternatively, the principal components can be plotted and outliers manually identified (Patterson et al., 2006; Price et al., 2006).

Association Testing

Standard practice for GWAS has been to test each locus one by one, with a chosen statistical test. Since Risch suggested it, the significance threshold has basically been remained at $5e-8$, a Bonferroni corrected standard α of 0.05 divided by $1e6$. This is based on an estimate of the number of genes and quantity of diallelic alleles found within them, and is the *de facto* standard used in GWAS (Risch et al., 1996; Morris et al., 2007). Many of the problems addressed in QC had long been identified by the time of the first GWAS, with discussion around statistical tests confronting the issues of stratification, relatedness and power (Thomas et al., 2005). As a result, GWAS has often used ethnically homogeneous groups in order to not interfere with a test statistic such as χ^2 (Robert J. Klein et al., 2005).

However, due to the complex nature of the diseases being studied and potential sources of bias, corrected regression is now the normal approach, with linear regression for quantitative traits and logistic regression for binary outcomes. This allows correction for fixed effects, such as PCs generated to account for population stratification and also other environmental confounding factors, such as body mass index (BMI) or age. The covariates can be represented as W as a vector of effect sizes α, with each locus x and effect size β for trait Y, under the assumption of an additive model. Now, the model can be summarized as:

$$Y = a + \alpha W + \beta x \tag{1.15}$$

Given a continuous trait, the coefficient of determination (R^2) can be calculated to measure the variance explained by the model, as well as a p-value and effect size based on an F test for each variable. Given a binary trait, a likelihood ratio test is performed to estimate significance and effect size (Uffelmann et al., 2021; Purcell et al., 2007; Laird et al., 2011).

An additional modification is found in the tool GEMMA and others (X. Zhou et al., 2012). They employ a relatedness matrix. The relatedness matrix contains an eigen-decomposition of the genetic material. This can then be incorporated into a mixed model as a random effect. As a result, sample sizes can be maximized by retaining any related individuals and cryptic relatedness can be combatted (Xiang Zhou et al., 2012; Mbatchou et al., 2021; Astle et al., 2009).

Imputation

Since the loci included in the array are selected within LD windows, other variants that are known to be in linkage with them can be predicted with differing levels of accuracy. This process has greatly benefitted from the knowledge that has been gathered about the human genome across numerous ethnic groups in projects such as 1000 Genomes and the phases of HapMap. Using this information, haplotypes can be inferred, and probabilistic models built, normally using Hidden Markov Models (HMM) (Abecasis et al., 2005; Li et al., 2009). The most recent tools are focussed on making these processes more computationally efficient, to deal with the ever-increasing size of datasets and genomic panels (Rubinacci et al., 2020).

Success of GWAS and Limitations

GWAS have been carried out enthusiastically since their inception. The GWAS Catalog is a central repository for recorded GWAS and the significant hits found (Buniello et al., 2019). As of August 2022, the team had collected results from 5,931 publications and recorded 415,784 associated variants. This includes work from what so-called biobanks, in which individuals provide biological samples such as blood, urine, and saliva. This can then be analysed for metabolites or other standard measures. Some collect magnetic resonance imaging (MRI), electrocardiogram (ECG) and other routine medical tests. Alongside this is demographic information and behavioural data collected from interviews and questionnaires. The largest of these is Biobank Ganz with samples from 20 million individuals (Huppertz et al., 2016). However, Chinese efforts have seen numerous biobanks started with the aim of including around 100 million individuals (Gan et al., 2015). In the UK, the UK Biobank has yielded over 3,000 publications and has a wealth of data from half-a-million people to explore from many angles, including genetics (Allen et al., 2012).

A study in 2000 noted that it takes an average of 17 years for a basic biological discovery to translate to the clinic (Balas et al., 2000). Since the first GWAS was in 2006, in 2023 we should perhaps be seeing practical applications derived from GWAS. Certainly, the wealth of genetic markers identified has offered many biological insights into the aetiology of diseases and traits. There has been a rapid proliferation of studies and common reproducibility of results, often targetting many different traits. This has ushered in the possibility of conducting meta-studies, in which multiple cohorts can be easily combined to increase the statistical power and identify more associations (Zeggini et al., 2009). This has also been extended to demonstrating causal relationships using the technique of Mendelian Randomization (MR) (Sanderson et al., 2022).

One area of development for early detection of diseases is that of polygenic risk scores (PRS). These summarize the different markers with an effect size above a threshold, in order to allow comparisons with individuals outside of the study data used. This is achieved by using weighted scores for individual risk loci, leading to a global score. They can then be validated in other cohorts or in a test portion of the dataset (Wray et al., 2021). Genetic screening has been used for some time, such as in the case of breast cancer, when testing for mutations in the BRCA genes has been carried out successfully for decades (Solomon et al., 2000). However, these polygenic tests broaden the scope of diseases that can be assessed and in more extreme cases, pre-emptive action, such as surgery or prescription of medications, can be taken (Torkamani et al., 2018). An example of how these are being used is identifying individuals who could benefit from statin treatment in the case of coronary artery disease (CAD) (Mega et al., 2015; Natarajan et al., 2017).

The use of the genetic information collected in GWAS has also been applied to drug design and identifying potential therapeutic targets. One method is using an extension to GWAS called Phenotype-Wide Association Studies (PheWAS). In this technique, single loci are tested for association with multiple phenotypes. This is usually used to find pleiotropic loci that are involved in multiple traits. However, it has also been used to find possible drugs that can be repurposed for a different condition based on shared genetics (Yin et al., 2018). The repurposing of the drug ustekinumab for Crohn's disease is an example of this, found by testing the *IL23R* gene (Reay et al., 2021). GWAS associations are also being used to prioritize proteins to target in drug design in CAD (Chen et al., 2021). Finally, there has been an effort to identify biomarkers as early warning signs, for example blood homocysteine levels as a risk indicator for strokes (Otani et al., 2019).

However, there have been limits to the success of GWAS. Many studies have been carried out using SNP arrays, that are not well suited for detecting rare SNPs. When compared to sequencing data, the UK biobank group found that the array calls were 98.7% accurate for SNPs of MAF > 0.01, falling to 73.2% for MAFs < $1e-4$. This was further pronounced for the imputed SNPs, with an accuracy of 95.2% for MAF > 0.01 and 32.2% for MAFs < $1e-4$ (Van Hout et al., 2020). Furthermore, since the loci identified are normally not the causative loci but instead in LD with the causative loci, there is an imprecision of identifying the driver of the disease, be it the gene or the region of the genome. This can also result in causative loci being missed, as the SNP in LD doesn't provide enough statistical power for a significant association (Schaid et al., 2018).

It was understood that complex disease was made up of many common mutations before GWAS (Hirschhorn et al., 2002). However, the mass collection of data and associations to many traits and diseases has elucidated quite how complex the genetic architecture is. What has been found is large numbers of mutations that contribute towards any one trait with only very small effect sizes. One interpretation of this, is the omnigenic model of disease (Boyle et al., 2017). In this model, it is hypothesized that there are a small number of core genes contributing to the disease but that regulatory networks within a cell are so interconnected that a mutation in a non-core gene can have an effect on the outcome. As a result, all genes that are expressed in the affected cell or tissue can have some effect on the trait.

Missing Heritability

In a 2008 study, genetic data were collected from around 30,000 individuals, mostly from Iceland (Gudbjartsson et al., 2008). The aim was to quantify the genetic contributions to height. They found 27 genomic regions with significant associations but small effect sizes and only 3.7% of phenotypic variance explained. Further work extended this to 40 significant variants explaining 5% of heritability, but this still fell well short of estimates for heritability for height, expected to be 80-90%. This is part of a wider problem known as ‘missing heritability’. It seemed that GWAS were finding numerous common variants which only accounted for a small fraction of the overall cause of a trait (Maher, 2008). This has led to speculation as to where this heritability can be found. One theory is that natural selection has led to small effect sizes over time by deselecting large effect sizes that cause disease or, in the case of beneficial mutations, adopting them. Another factor is that rare alleles with MAF below 5% could hold more deleterious variants or association being decreased due to incomplete linkage between causal SNPs and those sampled, thus not reaching the stringent p-value threshold (Manolio et al., 2009). This seems to be part of the problem, since later studies with more power or greater genetic coverage have uncovered causal variants previously missed and so explained much more phenotypic variance (J. Yang et al., 2010; Wood et al., 2014). Other forms of variation are often also missed by using SNP arrays, so rare variants, structural variants and epigenetic factors (Theunissen et al., 2020; Girirajan, 2017; Young, 2019; Trerotola et al., 2015). Aside from variants which simply aren’t captured by the methodology, some of the variation can be there just hidden in epistatic interactions that have not been identified, and so excluded from heritability calculations (Zuk et al., 2012).

1.4.7 Validation and Downstream Analysis

After carrying out a genetic analysis like GWAS, the result will be a number of variants or regions that have been judged to be significant. The task of validation and interpretation is of vital importance to understand the discoveries and place them in a biological context. There has been a considerable effort to collect and store the data that allows this process to be carried out. There are tools that can be used to efficiently retrieve this data and provide many annotations from multiple data sources, such as Variant Effect Predictor (VEP) and ANNOVAR (McLaren et al., 2016; K. Wang et al., 2010).

The most immediate annotations are those at a SNP level, so if the SNP is within the bounds of a gene, the name of that gene. The gene has long been known to consist of a structure, for example of exons and introns, coding and non-coding respectively (Gilbert, 1981). Much work has, of course, been carried out experimentally to define these regions (Brown, 2002). With the availability of more genetic data, it has been possible to automate this process with computational algorithms, such as matching sequences similar to known regions (Zhuo Wang et al., 2004; Mathé et al., 2002). Alongside other tools, such as those that detect splice sites, much data is available to classify the region of the gene in which a SNP is located (Desmet et al., 2009).

The consequence of a particular mutation helps understand how likely it is to cause a loss of function (LoF) in the protein coded for, however it should be noted that GWAS will not necessarily find the causal locus but one in LD with it. Mutations can be silent or missense, regarding whether the mutation changes the amino acid coded for within the peptide chain (Zhen Wang et al., 2001). There are repositories such as ClinVar, an archive of variant-phenotype links based on a wide variety of research that is submitted and curated (Landrum et al., 2018). Another example is Online Mendelian Inheritance in Man (OMIM), a medical genetics resource that catalogues gene-phenotype links, from which ClinVar also draws much of its information (Amberger et al., 2017; "Improving databases for human variation" 2016).

A number of approaches have been devised to predict the pathogenicity of mutations within the genome. The method taken by PolyPhen and SIFT uses protein sequence data from sources such as UniProt and SWISS-PROT to assess if a mutation will cause a change in amino acid, if the change in amino acid is likely to denature the protein and if the site is ligand binding, for example. They then return a probability-based score and classifications from benign to damaging (Adzhubei et al., 2013; Adzhubei et al., 2013). Additionally, there are predictors like Combined Annotation-Dependent Depletion (CADD) and Deleterious Annotation of genetic variants using Neural Networks (DANN) that use over 60 sources of data, including sequencing, epigenetic sites and evolutionary conservation based on other primate genomes (Rentzsch et al., 2019; Quang et al., 2015). The difference being that CADD is based on a support vector machine and DANN is based on a deep neural net model to capture non-linearity . There are numerous other approaches, but perhaps of note is rare exome variant ensemble learner (REVEL), an ensemble approach combining a number of different prediction methods and claims a greater accuracy as a result (Ioannidis et al., 2016).

Data representing the locality of expression of a gene is very useful when lending credence to a mutation affecting a particular trait, since it demonstrates the presence of the protein in a tissue of interest. The Genotype-Tissue Expression project (GTEx) has provided 7,051 expression profiles from 449 samples across 44 organs or cell lines. For each gene, odds ratios were given for expression relative to background levels of genetic expression (Aguet et al., 2017). Similarly, ProteomicsDB contains data for gene expression and other related information (Samaras et al., 2020). More specifically, the Allen Brain Atlas presents RNA sequencing data for numerous regions of the brain (Sunkin et al., 2013).

Another set of resources that help elucidate potential biochemical mechanisms for associations are pathway annotations. The Kyoto Encyclopedia of Genes and Genomes (KEGG) was released in 1995 as an integrated set of databases that has now grown to contain 15 manually curated databases. These hold data for pathways and functional modules, as well as genomic, chemical and medical information, that can be used to annotate genes. There are also useful resources, such as visual pathway maps (Kanehisa et al., 2017). Reactome is a set of relational databases storing biochemical pathways. This draws from numerous other sources, including KEGG and data from model organisms, and provides graphical representation of sets of genes within pathways using directional networks and statistical tests (Fabregat et al., 2017). These resources are useful for demonstrating the possible biological mechanisms by which genetic interactions might take place. A resource that specifically provides information on protein-protein interactions (PPI) is STRING-db. This is a database of evidence for PPI, with the strength of evidence scored, from the weak evidence of two proteins being found in the abstract of a journal article, through to stronger evidence such as lab assays. From these, an overall association score is calculated to represent the strength of evidence for the PPI (Szklarczyk et al., 2021). Although there are many different databases available, BioGRID is another that provides information on genetic interactions, cataloguing data from model organisms as well as 670,000 for humans with a focus on those involved in disease or biochemical pathways (Oughtred et al., 2021).

As of yet we have focussed on databases as a method of storing data on genes. However, much data is held in the form of ontologies. These are data structures, derived from the field of philosophy, that have been utilized in computer science to describe concepts from their highest level in a hierarchical structure down to the most granular, with the nature of the relationship described at each transition from term to term. This is useful because it provides a clear structure to conceptualize how these terms fit together and a framework to make semantic comparisons using these properties (Schuurman et al., 2008). To give a relevant example, we have the Gene Ontology (GO). GO has three separate categories, Biological Processes (BP), Molecular

Function (MF) and Cellular Component (CC). Each of these is a tree, with each non-terminal term being the parent to other terms. For example, taking CC as the root term, there are three child terms - cellular anatomical entity; protein-containing complex; and virion component. Each following child term becomes more specific until all cellular components can be labelled, with relation to overarching categories. Each term contains attributes such as a description and a list of genes for which their product can be labelled with this property (Ashburner et al., 2000). There are numerous other ontologies within the field of bioinformatics that can be leveraged to better understand sets of genes, for example the human phenotype ontology (HPO) that links genes to traits and the Disease Ontology (DO) that links them with particular diseases (Köhler et al., 2021; Schriml et al., 2012).

Gene Set Enrichment Analysis

As we've seen, there are many data sources for annotation of SNPs and genes, however this requires an approach to test these findings for significance. A method that was introduced in 2005 for gene expression data is gene set enrichment analysis (GSEA). This takes a list of genes that have been found to have significant associations with a trait and tests them together. Initially, this list was rank ordered by strength of association in the original methodology. The high-level process is to annotate each gene with a set of terms, such as biological pathways or GO terms, so that a pathway contains a number of proteins coded by a number of genes. A running sum statistic is calculated, increasing if a gene is in the set and decreased if not. From this, an enrichment score was generated, and statistical significance was found using permutation testing with a correction made for multiple testing (Subramanian et al., 2005).

This process has been refined over time with a number of different tools offering services to test sets of genes. A prominent example is g:Profiler, a web-based tool for testing gene sets, with SNP to gene annotation capabilities and graphical representations. It performs ranked and unranked tests, with a wide variety of different datasets, including GO terms, HPO and Reactome pathways. It assesses each term using a cumulative hypergeometric test to generate p-values. In ordered lists it is assumed that the genes higher up the list are more likely to be relevant and so tests are performed containing increasing numbers of genes in the list and the most significant result returned. When testing ontological terms it follows the 'True Path Rule' in which a child term is back propagated to include all parental terms which are also assessed (Reimand et al., 2007).

Finally, it is possible to make comparisons between genes or individuals using semantic similarity. Due to the hierarchical structure, terms are positioned relative to each other and so terms can be related from shared ascendents. It is therefore possible to make comparisons with a view to quantifying this relationship (Holliday et al., 2017). Early methods of conceptualizing semantic similarity involved counting the number of edges in common between two terms or set of terms. However, one improvement was that proposed by Resnik. It is based on the information content shared, with a lower-level terms also representing their ascendents in these counts. This takes a comparison of two collections of nouns called corpora, corpus c_1 and corpus c_2, and tests the number of terms shared:

$$sim(c_1, c_2) = \max_{c \in S(c_1, c_2)} \left[-log p(c) \right] \tag{1.16}$$

In which $S(c_1, c_2)$ is the set of shared concepts between c_1 and c_2. Then $p(c)$ is calculated from:

$$freq(c) = \sum_{n \in words} n \tag{1.17}$$

Where $words$ contains all the words included in c and n is the total occurrences of that word. This allows the final calculation of:

$$\hat{p}(c) = \frac{freq(c)}{N} \tag{1.18}$$

Where N is the total number of terms (Resnik, 1995). However, there are other approaches, such as the GO-specific methodology by Wang (J. Z. Wang et al., 2007).

Chapter Two

Identifying Epistatic Interactions

2.1 Introduction

This chapter looks at the range of epistasis detection methods available and is largely taken from my paper. I was responsible for all writing, as well as design and preparation of figures, with others assisting with the study design and reviewing the text (Russ et al., 2022).

There have been a variety of methodologies and approaches that have been formulated to identify epistatic interactions. This chapter is an overview of these and a general categorization of them. This is followed by a series of simulated benchmarking experiments to determine their ability to differentiate between true interactions and general statistical noise between unrelated loci. These tests represent different types of epistasis, with scaled parameters to adjust the difficulty of detection, so as to better distinguish the efficacy of each method. Finally, these are applied to a small use-case from UK biobank.

As we saw in the previous chapter, the term epistasis has evolved to mean when two genes interact to affect the expression of a particular phenotype, with the outcome of the interaction different from a simple additive effect made up of the total joint individual genetic effects (Cordell, 2002; Elandt-johnson, 1971). These digenic epistatic interactions are typically depicted in penetrance tables, with two loci of diploid chromosomes making up a 3x3 table. Thereby depicting all possible allelic combinations and their phenotypic contributions. These interactions can theoretically be manifested over many different combinations. Considering a binary penetrance table for high-risk or low-risk genotypes, across two loci, there are 512 (2^9) possible configurations. Albeit, whilst symmetrical models can be removed, that still leaves 50 unique conformations (Evans et al., 2006).The distribution of alleles within the population also approximates to the distribution set out in the Hardy-Weinberg Equilibrium (HWE). Accordingly, the count is governed by a function of the minor allele frequency for each locus (A. W. F. Edwards, 2008). The departure from additivity can also present in different ways. So, loci in an interaction will not necessarily have any significant effect individually - 'pure' epistatic models are made up of loci with no main effect on their own, as represented in Fig. 2.1. In such cases, both loci must be considered together to detect the interaction (Ryan J Urbanowicz et al., 2012a).

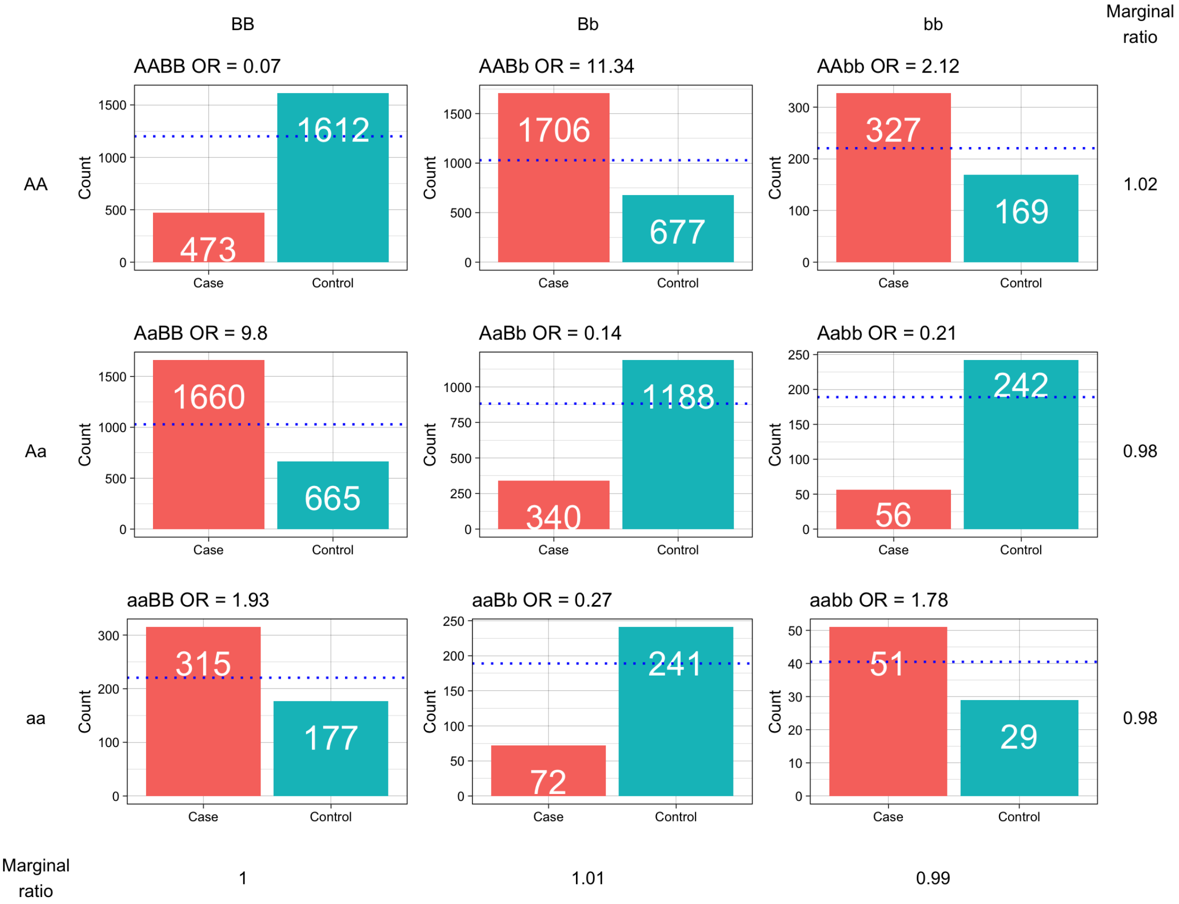


Figure 2.1: An example of the distribution of cases and controls across all possible genotypes as a result of an epistatic interaction for two loci (note that y-axes are different scales). To the right and below are the marginal effects, shown as a ratio of cases to controls. This is an example of a pure interaction because these marginal effects are very minimal. The blue dotted line indicates numbers expected under the Hardy-Weinberg Equilibrium, with capital genotypes representing the major allele and lower case the minor allele. OR is indicating the odds ratio comparing cases and controls at that genotype. (Previously published in Russ et al. 2022)

There are numerous approaches for detecting interactions available. A Web of Science query for ’epistasis detection’, taking methods designed for case/control tasks, retrieved a total of 105 methods published between 2010 and the end of 2020 (full list and references in table A.1). Upon inspection, 59 are currently available for download and formed the basis of the methods included in this study (Table A.1). Many different detection strategies have been implemented, with some testing all possible interactions exhaustively and others using different filtering steps to reduce the number of loci considered, resulting in a reduction in the number of tests required. Broadly, the methods can be categorized as statistical approaches, such as Generalized Linear Models (GLMs) and contingency tables, data mining and machine learning (ML) techniques. Chief amongst the data mining approaches are Multifactor Dimensionality Reduction (MDR), while nature-inspired algorithms are the key representatives of ML-based approaches. When selecting tools, key factors such as runtime and correcting for covariates, as well as detection of higher-order interactions were considered.

Early epistasis detection methods utilized simple statistical techniques, such as χ^2 tests (Carrasquillo et al., 2002) and GLMs for assessing interactions (Millstein et al., 2005; Macgregor et al., 2006). These have endured and are implemented in various tools. For example, Cassi and PLINK employ a Z-score based test, named fast-epistasis, as well as each implementing slightly differing logistic regression approaches. PLINK also has an implementation of the Wan Log-Linear method, BOOST (Ueki et al., 2012; C. C. Chang et al., 2015; Wan et al., 2010a). Another, more recent example is the wtest R package, that applies a novel statistic to compare distributional differences present in the alleles of cases and controls, using a χ^2 distribution (R. Sun et al., 2019).

MDR is a non-parametric data mining method, which accounts for all possible genotypes, for a set of loci as in Fig. 2.1. This is achieved by summarizing the best combinations of genotypes to divide cases and controls. It is similar to a naive Bayes approach, with the genotypes as parameters of a probabilistic classifier (Ritchie et al., 2001; Hahn et al., 2004; McKinney et al., 2006). As a modelling technique, it is flexible and can consider confounding factors as well as non-linear models. The latter renders it advantageous over GLMs. Built in to the methodology are training and testing stages by cross validation. There have been different MDR methods that have been implemented using odds ratios (Chung et al., 2007), fuzzy set theory (C.-H. Yang et al., 2020; Leem et al., 2017) or a support vector machine in SVM-MDR (Fang et al., 2012).

Nature-inspired algorithms aim to mimic successful search strategies in natural phenomena and have been applied to the challenge of efficiently detecting epistatic relationships. Ant Colony Optimization (ACO) algorithms, the most prominent examples of this type of approach, simulate ants that randomly ’explore’ data. They aim to identify the most efficient path to the ’food’, represented here by the division of cases and controls by an interaction. The ants share information and utilize algorithms to make probabilistic decisions in a network of nodes, reinforcing successful decisions by leaving ’pheromones’, assessed using a variety of statistical tests (Dorigo et al., 2006). Strategies for directing ants commonly use contingency table methods, logistic regression, information theory and Bayesian networks or combinations of these in stages (J. Shang et al., 2019). An early ACO algorithm example is AntEpiSeeker (Y. Wang et al., 2010), that uses a two-stage approach with χ^2 tests. More recent implementations include MACOED (Jing et al., 2015), combining the Akaike Information Criterion with a logistic regression variant and epiACO (Y. Sun et al., 2017) that uses mutual information and Bayesian network methods. Further nature-inspired algorithm examples include the particle swarm optimization (PSO) methods (Junliang Shang et al., 2016) as well as genetic algorithms (Yang Guo et al., 2019).

Other prominent examples include the approaches proposed Goudey et al, gain in sensitivity and specificity (GSS) method (Goudey et al., 2013; Chatelain et al., 2018). This approach employs Receiver Operating Characteristic (ROC) metrics to assess prediction performances between the interaction model and that of the two loci individually. It measures the difference in the area under a curve of all nine possible genotypes from two loci against the combined area under the curve of the two loci considered individually, giving six alleles total to be assessed.

Another interesting approach is SNPRuler, which generates association rules from combinations of loci and their genotypes using a derived statistic (Wan et al., 2010b). Also, of note are tools developed in order to harness computational resources efficiently, reducing runtime and memory requirements, such as MPI3SNP. This algorithm can be run in parallel on multiple CPUs or on GPUs for three locus interactions. The genotypes of the individuals are represented in a bitwise fashion and mutual information is calculated (Ponte-Fernández et al., 2020).

With such a variety of algorithms available, it is important to differentiate which will effectively discover epistatic interactions. Ibrahim et al (Ibrahim et al., 2013) compared three algorithms, SNPRuler, SNP Harvester and Ambience using large simulated datasets and multiplicative impure epistatic interactions. They reported that none of the algorithms consistently identified the interactions but identified Ambience to be the most robust and SNPRuler to have the most power, particularly across higher order interactions taking into account more than two loci. Chatelain et al (Chatelain et al., 2018) compared fastepi, GBOOST, SHEsisEpi, DSS and IndOR using four impure epistatic models. They found that DSS and GBOOST were the most powerful methods, with the former being preferential in the presence of limited linkage disequilibrium (LD), for example as a result of pruning variants in tight LD. Finally, Alchamlat et al compared MDR, BOOST, BHIT, KNN-MDR, MegaSNPHunter and AntEpiSeeker using pure interactions generated from real data, reporting KNN-MDR to be the most powerful approach (Abo Alchamlat et al., 2018).

The purpose of this study is to provide a clear, objective, comparison between some of the most prominent epistasis detection methods in a variety of different scenarios. Since there are many different approaches available, the aim is to find the optimal method and rationalize its selection criteria. With so many possible combinations of loci, being able to confidently assert interactions found will cater for the inclusion of epistasis analysis in GWAS studies and uncover a portion of the missing heritability. To account for this, the different scenarios employed in this study have been categorized based on different underlying genetic conditions, incorporating pure and impure epistasis.

2.2 Materials and Methods

2.2.1 Study Design

Since the objective of this study was to differentiate the ability of various tools to correctly predict epistatic interactions between two or three loci, a number of detection methods were selected, based on their merit or the type of methodology they adopted. Several tools were considered that have been reviewed in previous studies. This included SNPRuler as considered by Ibrahim et al (Ibrahim et al., 2013), DSS and BOOST of Chatelain et al (Chatelain et al., 2018) and KNN-MDR as used by Alchamlat et al (Abo Alchamlat et al., 2018). However, due to lack of availability, two substitutions were made with GSS used in place of DSS and KNN-MDR with a different MDR.

The methods chosen were assessed based on their ability to uncover specific interactions. Consequently, approaches that ranked features, such as the Relief-based methods (e.g. multiSURF and TuRF) were not considered (Ryan J. Urbanowicz et al., 2018). Tools were selected as representatives of common statistical approaches, such as logistic regression (e.g. Cassi and PLINK's epistasis) as well as of novel statistics approaches, for example wtest. Data mining approaches are represented by SNPRuler, MDR and GSS that have previously demonstrated efficacy. Nature-inspired algorithms, primarily represented by ACO-based approaches, were also considered. Included were AntEpiSeeker (Y. Wang et al., 2010), the most cited method in this domain, as well as epiACO (Y. Sun et al., 2017), a representative of more recent approaches. Finally, CINOEDV (Junliang Shang et al., 2016) was considered as an alternative type of nature-inspired algorithm, utilizing a PSO framework.

All tools were assessed based on simulated data, generated in a reproducible way, with a given set of parameters selected to differentiate between tools performances. For pure epistatic models, GAMETES was used to simulate models with different heritability and detection difficulty settings. Furthermore, five impure epistatic models, conforming to specific genotypes, were simulated using the EpiGEN tool. These models were situated in a LD structure for the noise loci. All scenarios have been generated for both two locus and three locus interactions and thirty replicates per scenario, with the aim of producing a robust testing strategy. Furthermore, these data can be used for future comparisons. The construction of these datasets as a testing regimen aims to cover, therefore, both marginal and non-marginal epistatic models across a range of detection difficulty degrees. The models used for testing impure epistasis, shown in Table 2.3, have been chosen based on penetrance tables from Evans et al (Evans et al., 2006). In the case of joint-dominant, joint-recessive and modular forms, these have been included to provide biologically-likely scenarios. In order to broaden the detection challenge, diagonal and X-OR models were also tested to provide a more complex, but plausible, genetic landscape.

The comparison between the tools' performance was dependent on their ability to identify the ground-truth epistatic interactions. Since they were assessed across a range of potential interactions, a ranked list of identified interactions for each tool was created. The rank of the True Positives, in each analysis, were retained when their ranked position was ≤ 50. This number was used to allow a range of ranks for statistical comparison of distributions, and was informed by inspection of the positions of true positives generated. Also, consideration was given to the relevance of results to researchers, since for a feature space of 2,500 there are greater than three million possible combinations. These lower ranked results are of diminishing interest and wasteful in terms of computational resources. When any True Positive was ranked position > 50 or not detected, it was assigned a rank of 51. This ensured that any tool detecting a small number of correct interactions, ranked highly, was not highlighted as outperforming other tools that were consistently identifying a higher number of correct interactions that were ranked at different levels. Given the purpose of this study is to differentiate the relative ability of the tools to detect epistatic interactions, the most important metric is finding the true interaction to be the most important pair, since a rank below first is hidden by random noise. Rank is the only practical method to assess them, since some of the tools don't provide a comparable statistic. However, the true interaction may have been ranked below first, and as such it is therefore useful to compare the distributions as well. As such, key indicators that have been used for these comparisons are percent of true interactions ranked first and a p-value for the comparative distributions. Due to the ranks presenting as a non-normal distribution, a lower tailed Mann-Whitney U test was used to compare each tool against the distribution of all the other tools combined, to generate these p-values.

All experiments were carried out using the University of Birmingham's BlueBEAR HPC service, which provides a High-Performance Computing service to the University's research community. See http://www.birmingham.ac.uk/bear for more details. The code and the settings applied available on GitHub, see Appendix. The versions of all software used in this study are provided in Table 2.1.

Table 2.1: Versions of software used

Tool	Version	Available
R	3.6.0	https://www.r-project.org/
Java	11.0.2	https://www.oracle.com/uk/java/technologies/javase/jdk11-archive-downloads.html
GAMETES	2.2	https://github.com/UrbsLab/GAMETES
EpiGEN	8/10/2020	https://github.com/baumbachlab/epigen
AntEpiSeeker	1.0	http://nce.ads.uga.edu/~romdhane/AntEpiSeeker/index.html
Cassi	2.5.1	https://www.staff.ncl.ac.uk/richard.howey/cassi/index.html
CINOEDV	2.0	https://cran.r-project.org/src/contrib/Archive/CINOEDV/
epiACO	1	https://sourceforge.net/projects/epiaco1/files/epiACO.rar/download
GSS	02/07/2014	https://github.com/bwgoudey/gwis-stats
MDR	3.0.2	https://sourceforge.net/projects/mdr/
MPI3SNP	1.0	https://github.com/chponte/mpi3snp/
PLINK	1.9b6.17	https://www.cog-genomics.org/plink/
SNPRuler	1	https://mybiosoftware.com/snpruler-predictive-rule-inference-epistatic-i html
wtest	3.2	https://cran.r-project.org/web/packages/wtest/index.html

2.2.2 Data Generation

There are a number of challenges related to generating simulated epistasis datasets. From a biological perspective, in the genetic environment alleles loosely conform to HWE, which is a function of the Minor Allele Frequency (MAF) and are anchored locally in a LD structure. Finally, any interaction involved in causation of a phenotype will explain a quantity of the trait's heritability, but likely leave much unexplained. Further practical challenges relate to runtime, efficiency and scalability. There are at least seven different tools available for generating simulated epistasis datasets (Blumenthal et al., 2020). In this study, GAMETES was used to simulate pure epistatic models and EpiGEN to create impure, defined models within LD.

GAMETES randomly generates pure epistatic models using a 'Sudoku' method. Similar to the puzzle game, in which rows and columns must hold a set of numbers. Here, the rows and columns of the 3x3 allelic table (in the case of two locus interactions) must have equal numbers of cases and controls. As any genotype is generated, there are constraints placed on the numbers that can appear at each other genotype. This is additionally determined according to the HWE distribution and heritability explained by the interaction (Ryan J. Urbanowicz et al., 2012b). EpiGEN works differently, with functionality to generate specific epistatic models, either using in built settings or via customizable model files. It makes no attempt to mask the main effects of the individual loci, and so creates impure epistatic models. These are placed within a genetic context, including a simulated LD structure, based on HapMap3 data. All epistatic models from both tools were accompanied by a fixed number of randomly generated noise loci. These had a range of MAFs and conformed to HWE. They are intended to provide True Negative outcomes to differentiate the ability of the tools tested. Both pieces of software are free to use and require no proprietary software to operate (Blumenthal et al., 2020).

These two algorithms are being applied due to their differing functionality - GAMETES generates pure epistatic models, whereas EpiGEN is suited for impure epistatic models. This means the test set can differentiate the ability of different tools for different situations. Both are effective tools for simulating genetic material, with noise variants following rules such as HWE and a range MAFs, as well as random placement of the interacting variants within these. A downside of GAMETES is that the genotypes that contain the interaction are randomly dispersed and so don't comply with any potential Mendelian conformation. EpiGEN has the functionality to select genotypes that affect the phenotype, so that specific interactions can be created by the user. However, in order to create the desired effect size it modifies the number of cases and controls, which could introduce some bias due to this imbalance.

GAMETES was used to configure the MAF and the heritability of the modelled interactions. The MAF selected was 0.4 in order to better populate each genotype, especially interactions with more minor alleles. Heritability is a measure of the variability attributable to the phenotype, in this case, the interaction. In this study, it was set at 0.02, 0.01 and 0.005 so to adjust the interaction effect sizes. Additionally, GAMETES provides a metric termed the Ease of Detection Measure (EDM), which indicates an assessed level of difficulty for a tool to detect the interaction. GAMETES randomly generates a set number of models, ranks them according to their EDM, and returns a user-defined number of models by percentile. Here, three models were requested, the easiest disregarded and the remaining two used for testing (*GAMETES* 2021; Ryan J Urbanowicz et al., 2012a). Noise loci were randomly generated with MAFs between 0.05 and 0.5, with each dataset consisting of 1,000 cases and 1,000 controls. Thirty replicates were made for each scenario with randomly generated differences. This procedure was carried out for second and third-order interactions. The settings applied in these experiments are summarized in Table 2.2.

Table 2.2: Configurations used for GAMETES generated models. In all datasets there are 1,000 cases and controls and 30 replicate

Loci in Interaction	Heritability	Ease of Detection Measure	Total loci
2	0.02	2	2500
2	0.01	2	2500
2	0.005	2	2500
2	0.02	1	2500
2	0.01	1	2500
2	0.005	1	2500
3	0.02	2	500
3	0.01	2	500
3	0.005	2	500
3	0.02	1	500
3	0.01	1	500
3	0.005	1	500

EpiGEN was used to generate a corpus of synthetic genetic data based on 2000 samples modelled on Chromosome 22, this was chosen for speed of processing. Several models were included, based on possible epistatic combinations, as detailed in Evans et al (Evans et al., 2006). Models were built with a number of modes of epistatic interaction, namely joint-recessive, joint-dominant, modular, diagonal, and XOR (Table 2.3). These particular models were selected in order to give a range of epistatic models for comparison. Joint-recessive and joint-dominant models represent simple Mendelian interactions, with the modular and XOR interactions as conceivable variants on these. The diagonal interaction is a less biologically inspired conformation in order to provide an alternative, plausible challenge. This was achieved by altering the case-control ratio at the genotypes involved in the interaction, changing the penetrance. Exploratory data analysis was used to find a range of ratios that demonstrated a differentiation between the tools being tested. Noise loci made the feature space up to 500 in total and were limited to MAFs between 0.05 and 0.5. The datasets were set to aim for 1,000 cases and 1,000 controls, however these numbers did fluctuate to fit the models. Thirty replicates were made for each scenario with randomly generated differences. This procedure was repeated for third-order interactions, except with a feature space of 100 loci, including three interacting loci and 97 noise loci. For penetrance tables, see Table 2.4.

Table 2.3: EpiGEN Penetrance models with capital genotypes as the major allele

Joint Dominant	**AA**	**Aa**	**aa**	**Joint Recessive**	**AA**	**Aa**	**aa**
BB	0	0	0	**BB**	0	0	0
Bb	0	1	1	**Bb**	0	0	0
bb	0	1	1	**bb**	0	0	1
Modular	**AA**	**Aa**	**aa**	**XOR**	**AA**	**Aa**	**aa**
BB	0	0	0	**BB**	0	0	1
Bb	0	0	1	**Bb**	0	0	1
bb	1	1	1	**bb**	1	1	0
Diagonal	**AA**	**Aa**	**aa**				
BB	1	0	0				
Bb	0	1	0				
bb	0	0	1				

Table 2.4: EpiGEN Penetrance models with capital genotypes as the major allele. Third dimension indicated as shown in final section for C allele.

Joint Dom.	**AA**	**Aa**	**aa**	**Joint Rec.**	**AA**	**Aa**	**aa**
BB	0/0/0	0/0/0	0/0/0	**BB**	0/0/0	0/0/0	0/0/0
Bb	0/0/0	0/1/1	0/1/1	**Bb**	0/0/0	0/0/0	0/0/0
bb	0/0/0	0/1/1	0/1/1	**bb**	0/0/0	0/0/0	0/0/1
Modular	**AA**	**Aa**	**aa**	**XOR**	**AA**	**Aa**	**aa**
BB	0/0/0	0/0/0	0/0/0	**BB**	0/0/0	0/0/0	0/0/1
Bb	0/0/0	0/0/0	0/0/1	**Bb**	0/0/0	0/0/0	0/0/1
bb	0/0/1	0/0/1	0/0/1	**bb**	0/0/1	0/0/1	1/1/0
Diagonal	**AA**	**Aa**	**aa**	**C allele**	**AA**	**Aa**	**aa**
BB	1/0/0	0/0/0	0/0/0	**BB**	CC/Cc/cc	...	...
Bb	0/0/0	0/1/0	0/0/0	**Bb**	CC/Cc/cc	...	...
bb	0/0/0	0/0/0	0/0/1	**bb**	CC/Cc/cc	...	...

The data were converted to various required formats using PLINK and R. The versions of all tools used in this study are detailed in Table 2.1. Default settings were used unless stated otherwise or expanded upon.

2.2.3 Tools Benchmarked

Statistical Approaches

The first category of approaches focuses on using straight-forward statistical tests and simply test every combination exhaustively. These approaches include contingency table methods, log-linear regression and logistic regression as well as novel statistics in order to assess each potential interaction.

PLINK: fast epistasis

Three PLINK (C. C. Chang et al., 2015) epistasis detection methods were assessed. The first one, fast-epistasis, takes the number of individuals for cases and controls at all genotypes for two loci (Table 2.5) and condenses them to a 2 x 2 table (Table 2.6).

Table 2.5: Two locus 3 x 3 table for allelic combinations, here *n* is equal to the number of individuals with each genotype

Minor Allele Dose Per Locus	**0**	**1**	**2**
0	n_{00}	n_{01}	n_{02}
1	n_{10}	n_{11}	n_{12}
2	n_{20}	n_{21}	n_{22}

Table 2.6: Reduced 2 x 2 table for two loci, here n is equal to the number of individuals with each genotype

Allele Per Locus	Major	Minor
Major	$A = 4n_{00} + 2n_{01} + 2n_{10} + n_{11}$	$B = 4n_{20} + 2n_{01} + 2n_{12} + n_{11}$
Minor	$C = 4n_{20} + 2n_{21} + 2n_{10} + n_{11}$	$D = 4n_{22} + 2n_{21} + 2n_{12} + n_{11}$

Using the matrix in the Table 2.6, log odds and variance can be calculated as shown below, where OR is odds ratio, v is variance, and A-D refer to cells of the matrix:

$$OR = \log \frac{AD}{BC} \tag{2.1}$$

$$v = \frac{1}{A} + \frac{1}{B} + \frac{1}{C} + \frac{1}{D} \tag{2.2}$$

Lastly, the χ^2 test statistic is calculated comparing cases and controls:

$$T = \frac{(OR_{case} - OR_{control})^2}{v_{case} + v_{control}} \tag{2.3}$$

PLINK: BOOST

The second is an implementation of BOOST, a log-linear or Poisson regression model generating a contingency table, similar to the Table 2.6 but with an additional dimension to divide cases and controls, given as k. An observed count is denoted as n_{ijk} with i and j the genotype at two loci. This is the result of the random variable N_{ijk} that is assumed to follow a Poisson distribution. The probability that an observation falls into any one cell is denoted as π_{ijk}, with the sum of all π being 1. The mean is given by:

$$\mu_{ijk} = n\pi_{ijk} \tag{2.4}$$

The likelihood is calculated with the equation:

$$f(\mu) = \prod_{i,j,k} \frac{e^{-\mu ijk} \cdot \mu_{ijk}^{nijk}}{n_{ijk}!} \tag{2.5}$$

This is transformed to the log-likelihood, represented as:

$$L(\mu) = \sum_{i,j,k} [n_{ijk} \log(\mu_{ijk}) - \mu_{ijk} - \log(n_{ijk}!)] \tag{2.6}$$

This methodology allows for fast, exhaustive testing.

PLINK: epistasis

The final PLINK method, called simply ’epistasis’ assesses each interaction using logistic regression. This uses an additive model with an extra multiplicative term to test for an epistatic interaction. The $\hat{\beta}$ for the interaction term is estimated and tested for significance. This method also allows for covariates to be adjusted for.

Cassi

Cassi (Ueki et al., 2012) provides an alternative logistic regression approach to the PLINK logistic regression model. Initially, the phenotype is regressed for an additive and a multiplicative model. Then a likelihood ratio test is performed for models with and without the interaction term. Thus, the deviance from the additive model is attributed to the interaction. Cassi also allows for covariates to be adjusted.

wtest

Another contingency table method is wtest (R. Sun et al., 2019), which can be used to test one to many loci based on k combinations of interacting loci and a binary phenotype, making a 2 x *k* table. The method assesses distributional differences between cases and controls. Taking the *ith* column of *k*, the numbers of n_{1i} cases and n_{0i} controls and the total number of N_1 cases and N_0 controls, the conditional probabilities can be estimated, $\hat{p}_{1i} = n_{1i}/N_1$ and $\hat{p}_{0i} = n_{0i}/N_0$. The full formula uses a χ^2 distribution with *f* degrees of freedom:

$$W = h\sum_{i=1}^{k}\left[\log\frac{\hat{p}_{1i}/\left(1-\hat{p}_{1i}\right)}{\hat{p}_{0i}/\left(1-\hat{p}_{0i}\right)}/SE_i\right]^2 \sim \chi^2_f \tag{2.7}$$

With the scalar h and f as covariance matrices of the log odds ratios, estimated from bootstrapped samples under the null hypothesis. The Standard Error is calculated as follows:

$$SE_i = \sqrt{\frac{1}{n_{0i}} + \frac{1}{n_{1i}} + \frac{1}{N_0 - n_{0i}} + \frac{1}{N_1 - n_{1i}}} \tag{2.8}$$

An option to filter the loci by main effect p-value is also provided. Since this option for this set of experiments would result in all pure interactions being missed, the filter was set to include all loci with a p-value less than 1. Beyond exhaustive statistical tests, approaches have been devised to optimize a search strategy before assessing the interaction.

Nature-Inspired Approaches

AntEpiSeeker

AntEpiSeeker (Y. Wang et al., 2010) uses a standard ant colony optimization framework over i iterations, in which m ants are guided through a path of n interacting genetic loci, selected dependent on a probability density function (PDF) for each locus *k*. In these models, the ants leave pheromones to indicate favourable paths, which evaporate over time. The PDF is calculated as follows:

$$p_{ki} = \frac{\tau_{ki}^{\alpha}\cdot\eta_k^{\beta}}{\sum_{j=1}^{L}\tau_{ji}^{\alpha}\cdot\eta_k^{\beta}} \tag{2.9}$$

η^{β} represents some prior 'attractiveness' information, which is set to 1 here. As a result, the PDF represents the pheromones at locus k divided by the sum of the pheromones at each of j locus from a set of L total loci. α represents the weight of the pheromone at the locus and β is the weight of the heuristic information. The pheromone levels are updated according to the χ^2 test score for the interacting loci, as such:

$$\tau_{k(i+1)} = \left(1-\rho\right)\tau_{ki} + \Delta\tau \tag{2.10}$$

Where ρ is a number between 0 and 1, representing the pheromone evaporation rate and $\Delta\tau$ is the change in pheromones at locus k at iteration i and is equal to $0.1\chi^2$. This is repeated for all m ants over i iterations.

In the documentation for AntEpiSeeker there are suggested values for most of the parameters and these were used in this experiment. However, for two parameters, iTopModel and iTopLoci, this guidance was not given. This is presumably because they are dependent on the number of loci assessed since they define the number of possible models and the number of loci with the maximum quantity of pheromones. For second order interactions they were set to 1,000 and 200 respectively. So as to reflect the smaller number of loci assessed for the third order interactions, these values were set to 50 and 10.

epiACO

The epiACO (Y. Sun et al., 2017) approach adopts the same framework as AntEpiSeeker, shown in Eq 2.9. However, it employs a different method for testing interactions and parameter values. The interaction test statistic used, termed the SValue, uses Mutual Information (MI), the entropies of S loci and Y phenotype, and a Bayesian metric, the K2 Score in a logarithmic form:

$$MI(S;Y) = H(S) + H(Y) - H(S,Y) \tag{2.11}$$

$$SValue = \frac{MI}{K2score_{log}} \tag{2.12}$$

There are a number of search strategies adopted in order to make the search more effective. The path selection strategy involves ants taking either a probabilistic route or a stochastic one. The probability P of ant k selecting locus i at iteration t is defined as:

$$P_k^i(t) = \begin{cases} R & q <= q_0 \\ \\ S & q > q_0 \end{cases} \tag{2.13}$$

In which q is a randomly generated number from a uniform distribution of [0,1] and q_0 is the iteration divided by the total number of iterations. The probabilistic path is defined as:

$$R = \begin{cases} \frac{\tau_i(t)^\alpha \cdot \eta_i^\beta}{\sum\limits_{u \in U_k(t)} \tau_u(t)^\alpha \cdot \eta_u^\beta} & i \in U_k(t) \\ \\ 0 & otherwise \end{cases} \tag{2.14}$$

where $\tau_i(t)$ is the pheromones at locus i and η_i is the heuristic information at the locus. $U_k(t)$ is the set of not-selected loci by ant k. α represents the weight of the pheromone at the locus and β is the weight of the heuristic information. The stochastic path strategy follows:

$$S = \begin{cases} 1 & i = rand(V_k(t)) \\ \\ 0 & otherwise \end{cases} \tag{2.15}$$

where all loci at iteration t are sorted in descending order by pheromones, with the latter half being represented by $V_k(t)$. This allows for a wider search space at lower iteration numbers. The pheromone updating strategy at iteration i and locus k can then be defined as:

$$\tau_{i(t+1)} = (1-\rho)\tau_{i(t)} + \Delta\tau_{i(t)} + \Delta\tau^*_{i(t)} \tag{2.16}$$

With $\Delta\tau_{i(t)}$ a pheromone increment for an ant visiting and $\Delta\tau^*_{i(t)}$ being a bonus increment for those that belong to candidate solutions based on the S-value calculated. $\Delta\tau_{i(t)}$ is found for ant a of m total ants:

$$\Delta\tau_{it} = \sum_{a=1}^{m} \Delta\tau_{it}^{a} \tag{2.17}$$

The default settings of the epiACO implementation return only the top three interactions and, in order to make its output comparable with the other assessed methods, the code was modified to return the top 50 interactions.

CINOEDV

CINOEDV (Junliang Shang et al., 2016) uses particle swarm optimization (PSO). A number of particles are simulated and distributed across a k loci space with the aim to position themselves at the strongest interaction. After initial random placement, they adjust their velocity and position. This is dependent on shared information with nearby particles and an assessment of their local space using a novel co-information method based on entropy to assess interactions. At each iteration g, the position (S) and velocity (v) of each particle is updated:

$$\tilde{v}_{qk}^{g+1} = W_{qk}^{g} \cdot v_{qk}^{g} + c_1 \cdot r_1 \cdot \left(PS_{qk}^{g} - S_{qk}^{g}\right) + c_2 \cdot r_2 \cdot \left(GS_{qk}^{g} - S_{qk}^{g}\right) \tag{2.18}$$

for the *qth* particle. W_{qk}^{g} is an inertia term, which takes into account the local scores against those elsewhere. The c terms are acceleration constants, and r terms are random values between 0 and 1. *PS* and *GS* represent the particle's most favourable position it has visited and that of the whole swarm, respectively. The position is then updated as such:

$$\tilde{S}_{qk}^{g+1} = S_{qk}^{g} + v_{qk}^{g+1} \tag{2.19}$$

The result is that the particles move into groups, centred around the interactions that produce the best outcomes.

Data Mining Approaches

MDR

The MDR (Hahn et al., 2003) algorithm splits the data into a training set and a test set at a ratio of 9:1. During the training stage, the probabilities for each genotype, at two or more loci, given the status of case or control, are calculated. The interaction is defined by the probability that each genotype is a case. The number of dimensions considered for a two locus problem is thereby reduced from nine genotypes to the product of those genotypes. This evaluation is a naive Bayes classifier, defined by distributions of cases and controls across the possible genotypes:

$$v_{NB} = \underset{v_i \in V}{\arg\max}\, p\left(v_j\right) \prod_{i=1}^{n} p\left(a_i | v_j\right) \tag{2.20}$$

Where v_j is one of a set of V phenotypic classes and a_i is an attribute describing each multi-locus genotype present. The output is a binary variable ascribing the presence of a genotype associated with cases. An assessment of accuracy is carried out using the reserved test set with the mean of sensitivity and specificity:

$$Accuracy = 0.5 \times \left(\frac{TP}{TP+FN} + \frac{TN}{TN+FP}\right) \tag{2.21}$$

By splitting the genotypes into high-risk and low-risk genotypes, MDR reduces the multi-locus space into a binary variable and assesses the accuracy that interaction assigns true-positives and true-negatives.

GSS

Gain in sensitivity and specificity (GSS) (Goudey et al., 2013) is a method which employs measures of sensitivity and specificity to compare an additive and interaction model. The additive model is calculated by assessing each of the genotypes of the individual loci for sensitivity and specificity and taking the largest area under the Receiver Operator Characteristic (ROC) curve. This is compared to the area under the ROC curve for each of the nine genotypes represented in the interaction model. Hence, any gain in these measures can be attributed to the interaction and quantified using a p-value calculated from a min-max optimization. By considering the difference between the predictive power of the loci individually against the nine possible genotypes from two loci, the sensitivity and specificity are optimized.

SNPRuler

SNPRuler (Wan et al., 2010b) uses predictive rule learning to find possible epistatic interactions. Rules are generated in trees for different allelic configurations, based on classification of cases or controls. Any individual SNP can be defined multiple times for different genotypes, to account for more complex epistatic models. However, any additional rule must increase predictive power. This is assessed using a derived χ^2 measure to assess potential additional rules, which must achieve this increase to be appended to the current rule. The upper bound, UB, of the potential addition to the rule is calculated:

$$UB = \frac{(Rm - (b - \min(b, d')))^2}{(m + (b - \min(b, d')))(\gamma m - (b - \min(b, d') - m))} \tag{2.22}$$

Here, Rm is the ratio of cases to controls in the current rule, b is the number of cases represented in that rule, whilst d' is the number of the b case that do not have the potential new genotype. The term, m refers to the minimum value between the number of controls that adhere to the current rule and those in the new genotype and γ is the total number of samples divided by the number of controls for the current rule being built. Finally, all rules generated are tested using a χ^2 test to calculate a statistic and p-value. This two stage approach aims to find any possible interactions quickly before applying a more rigorous statistical test to rank those found.

MPI3SNP

MPI3SNP (Ponte-Fernández et al., 2020) is specifically designed for fast detection of third order interactions. In order to be most computationally efficient, individuals are represented in a bitwise fashion by their genotype, with a table for cases and one for controls. Using this division, probabilities for a combination of SNPs can be quickly ascertained for use in a Mutual Information (MI) equation:

$$I(X,Y) = H(X) + H(Y) - H(X;Y) \tag{2.23}$$

The MI is calculated by the addition of the entropy of X, the genotype, and Y, the phenotype, followed by the subtraction of the joint entropy. This is repeated exhaustively for all combinations, with the option of using CPU or GPU parallelization in order to further minimize the run time.

Table 2.7: Tool Summary

Tool	Category	Reference
AntEpiSeeker	Nature-Inspired	https://doi.org/10.1038/npre.2012.6994.1
Cassi	Statistical	https://doi.org/10.1371/journal.pgen.1002625
CINOEDV	Nature-Inspired	https://doi.org/10.1186/s12859-016-1076-8
epiACO	Nature-Inspired	https://doi.org/10.1186/s13040-017-0143-7
GSS	Data Mining	https://doi.org/10.1186/1471-2164-14-S3-S10
MDR	Data Mining	https://doi.org/10.1093/bioinformatics/btf869
MPI3SNP	Data Mining	https://doi.org/10.1177/1094342019852128
PLINK:BOOST	Statistical	https://doi.org/10.1086/519795
PLINK:FastEpi	Statistical	https://doi.org/10.1086/519795
PLINK:epistasis	Statistical	https://doi.org/10.1086/519795
SNPRuler	Data Mining	https://doi.org/10.1093/bioinformatics/btp622
wtest	Statistical	https://doi.org/10.1186/s12920-019-0638-9

For details of each algorithm tested, please see Tables A.1, 2.10 and 2.7. In all cases, default or recommended values were applied.

2.2.4 Atrial Fibrillation

The detection tools were employed to identify potential interactions associated with Atrial Fibrillation (AF) using patients selected from the UK biobank cohort (Bycroft et al., 2018). This presents a real-world application of these methods, with the assessment of the validity of the results conducted using *a priori* knowledge. However, it also functions as a limited experiment into possible interactions present in AF.

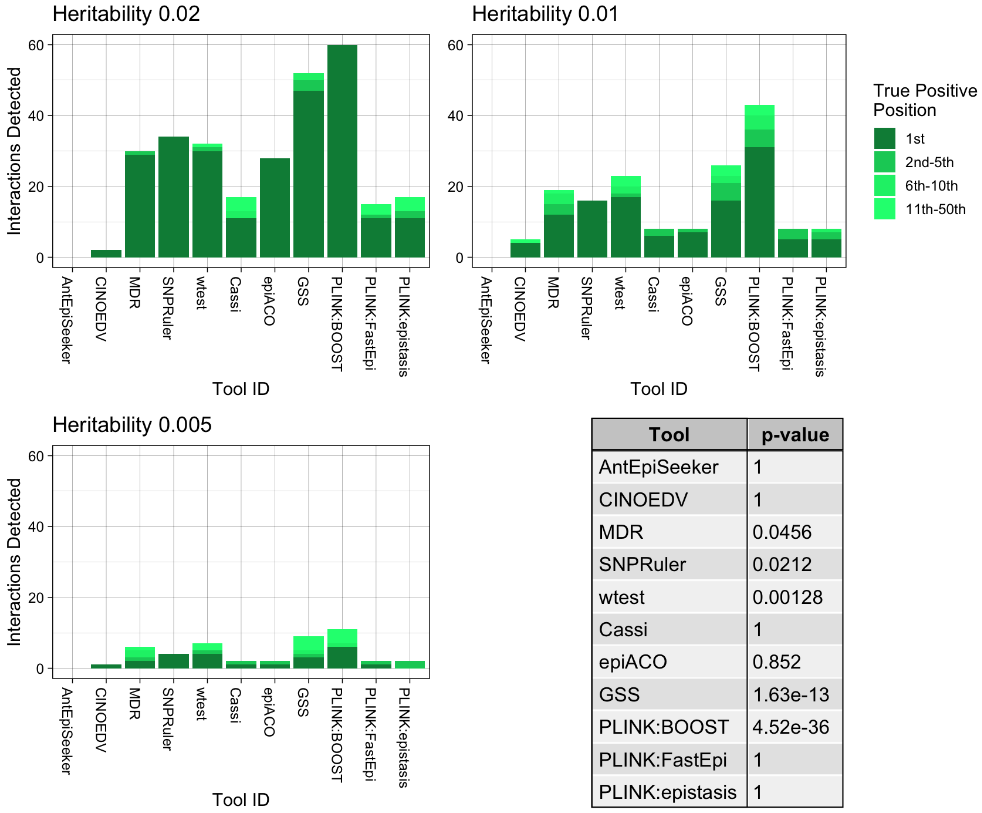
There are a number of considerations when applying these tools to real data. The methods tested, as a group, have a number of limitations. These include memory use, runtime, inability to deal with missing values and a lack of facilities to account for environmental factors, as such requiring careful selection of samples. As a result, the cohort and number of loci considered was filtered to work within the scope of these tools. UK biobank contains genetic and phenotype data for 486,445 individuals, including 33,492 reported to have AF and 93,095,623 genotyped and imputed loci. Individuals were removed if they had been flagged as outliers for genetic missingness and heterozygosity, had a sex chromosome mismatch or evidence of sexual aneuploidy. Only Caucasians were included in the study and patients were excluded if they had a second degree relative or closer participating in the UK biobank cohort, leaving 335,400 samples. This included 23,178 individuals remaining with AF who were paired with an equal number of controls that had no diagnosis of AF, using random sampling. This left a final cohort size of 46,356. Loci were included if they had an imputation INFO score of 1 and a MAF of at least 0.495, leaving 2,478 loci. After removing those with missing values, 1592 loci remained. By using a high MAF, it ensures that under HWE each genotype is represented by many samples, thereby increasing the statistical power of the assessment. Gene annotations were derived from Variant Effect Predictor (McLaren et al., 2016). To assess if interacting genes are over-represented for Gene Ontology classes, they profiled using Bonferroni correction in gProfiler (Reimand et al., 2007). STRING was used for protein-protein interactions, keeping default settings with interactions >0.400 being significant (Szklarczyk et al., 2019).

2.3 Results

Figs. 2.2-2.5 present each tool's rank of the simulated epistatic interaction. This is limited to a rank of 50, with those interactions, which were not detected or ranked lower than 50, represented as translucent points at a provisional rank of 51. Since some methods exhaustively test all combinations, it is fair to introduce a limit beyond which a discovery is considered as 'not detected'. Each plot represents a different set of parameters depending on the way the data were generated. For pure interactions, the graphs are split by heritability, a measure of how penetrant the interaction, whilst for the impure epistatic datasets they are split into different scenarios based on defined models. There are also different levels of detection difficulty for each model. Pure interactions are assessed by the simulation software, while impure interactions' detection is based on the fraction of cases and controls that have the affected genotypes. For each experiment, thirty tests were carried out per set of parameters, with random differences between them. Experiments were carried out for both second-order and third order interactions, including interactions of two loci and three loci respectively, as shown below.

2.3.1 Second Order Interactions

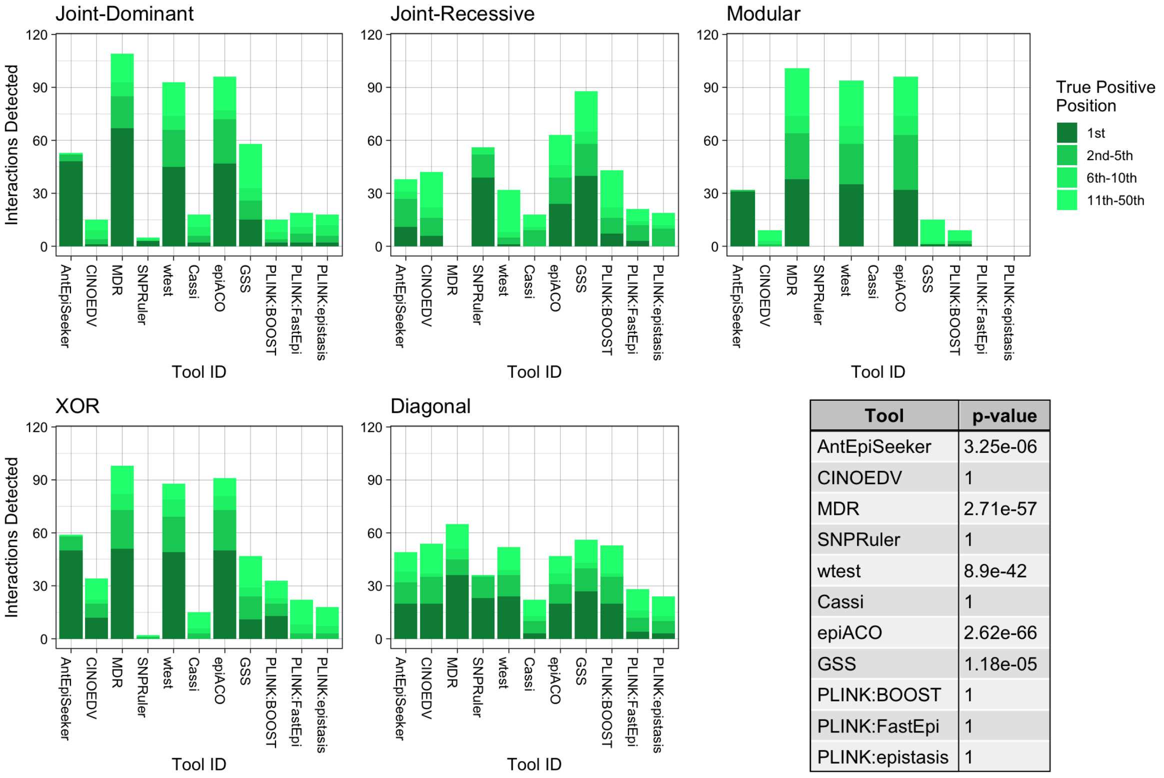
In the case of pure epistatic interactions (Fig. 2.2), the PLINK implementation of BOOST exhibited the most effective performance ($p = 4.52e-36$), identifying all the interactions at a heritability of 0.02, with 71.7% and 18.3% of the interactions correctly identified for 0.01 and 0.005 heritability levels respectively. GSS achieved the second-best performance ($p = 1.63e-13$) followed by wtest, SNPRuler and MDR. Looking specifically at the correct interaction being top ranked, BOOST found 53.9% of the true interactions, followed by GSS, SNPRuler and wtest.



Tool	p-value
AntEpiSeeker	1
CINOEDV	1
MDR	0.0456
SNPRuler	0.0212
wtest	0.00128
Cassi	1
epiACO	0.852
GSS	1.63e-13
PLINK:BOOST	4.52e-36
PLINK:FastEpi	1
PLINK:epistasis	1

Figure 2.2: Summary of the pure epistasis results for second order interactions. The three bar charts show the number of True Positive interactions discovered and the position that the algorithm ranked it amongst combinations with noise loci. Each chart shows a different heritability for the interaction, with higher heritability explained making it more prominent against random noise. The table shows the results of a Mann-Whitney U Test comparing the non-normal distribution of True Positive ranks for a single tool against the distribution of true positive ranks for all other tools. (Previously published in Russ et al. 2022)

When assessing impure two locus models of epistasis (Fig. 2.3), for all models except from the joint-recessive model, MDR exhibited the best performance ($p = 6.31e-90$), identifying between 54.2% and 90.9% of interacting loci. However, for the joint-recessive model, this method has the lowest performance, failing to detect any of the interactions. In this case, GSS correctly identified most of the interacting loci, demonstrating significantly superior detection ability ($p = 1.04e-26$). Notably, across all models, epiACO achieved the highest mean proportion of identified interactions at 65.5% ($p = 2.62e-66$) compared to MDR at 62.2% ($p = 2.71e-57$), followed by wtest with a 59.8% discovery ($p = 8.90e-42$). Finally, for ranking the correct interaction first, MDR exhibited the best performance, identifying 32.0% of the correct interactions as most likely, followed by epiACO, AntEpiSeeker and wtest.

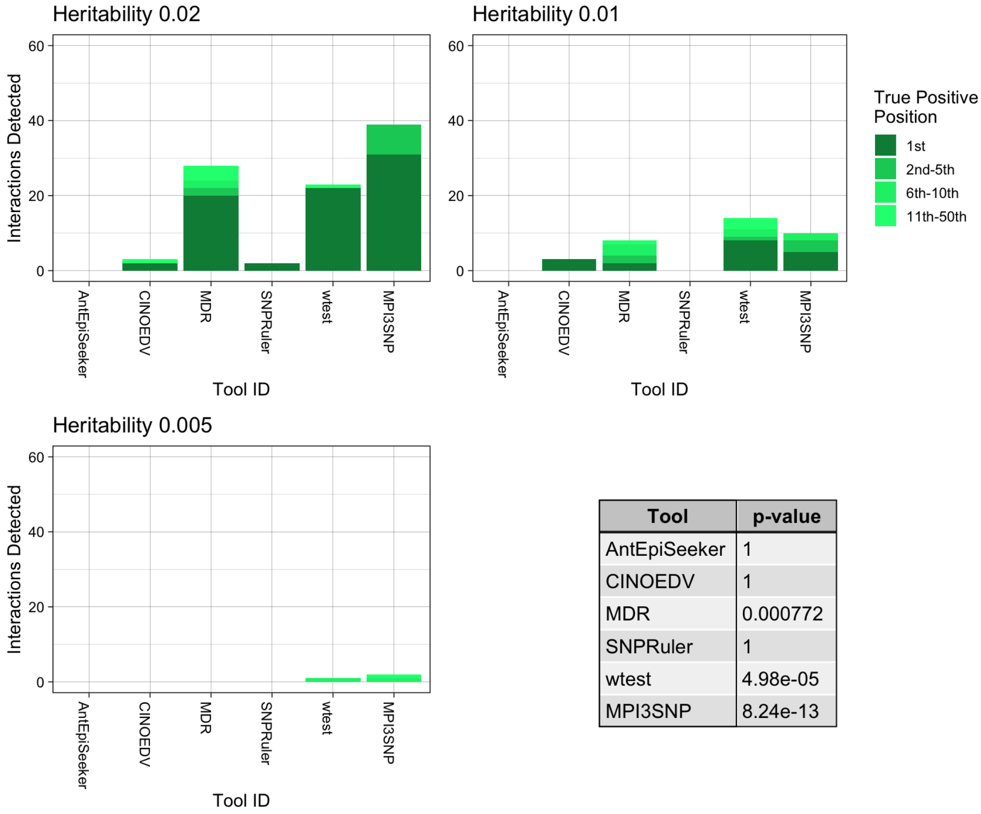


Tool	p-value
AntEpiSeeker	3.25e-06
CINOEDV	1
MDR	2.71e-57
SNPRuler	1
wtest	8.9e-42
Cassi	1
epiACO	2.62e-66
GSS	1.18e-05
PLINK:BOOST	1
PLINK:FastEpi	1
PLINK:epistasis	1

Figure 2.3: Summary of the impure model results for second order interactions. Each bar chart shows the number of True Positive interactions discovered and the position that the algorithm ranked it amongst combinations with noise loci. Each chart shows a different interaction models (see Table 2.3). The table shows the results of a Mann-Whitney U Test comparing the non-normal distribution of True Positive ranks for a single tool against the distribution of True Positive ranks for all other tools. (Previously published in Russ et al. 2022)

2.3.2 Third Order Interactions

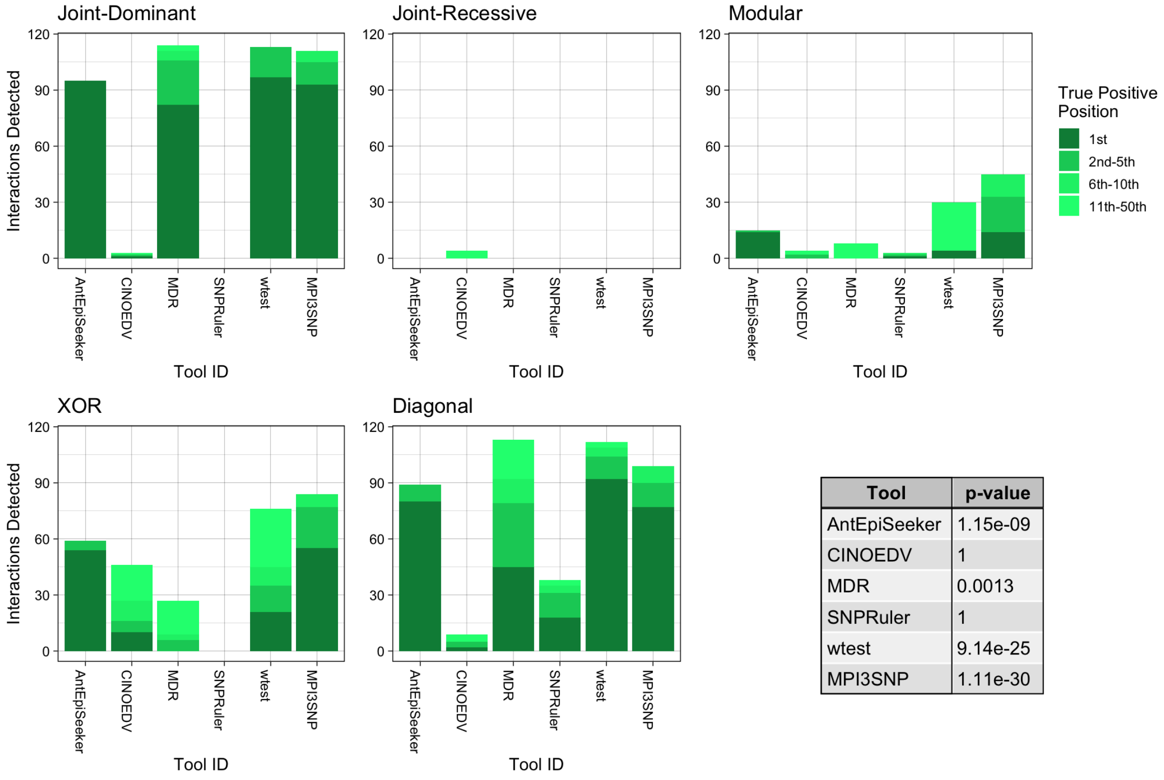
Some of the tools assessed cannot identify higher order interactions and, as such, were not included in the following experiments. Fig. 2.4 presents the tools' interaction detection performances for pure epistatic three-locus models. MPI3SNP identified 28.3% of correct interactions ($p = 8.24e - 13$), followed by wtest which retrieved 21.1% ($4.98e - 05$), and MDR with 20.0% ($p = 7.72e - 4$). Furthermore, MPI3SNP identified 20.5% of the correct interactions as most important, compared to wtest and MDR with 17.2% and 12.2%, respectively.



Tool	p-value
AntEpiSeeker	1
CINOEDV	1
MDR	0.000772
SNPRuler	1
wtest	4.98e-05
MPI3SNP	8.24e-13

Figure 2.4: Summary of the pure epistasis results for third order interactions. The three bar charts show the number of True Positive interactions discovered and the position that the algorithm ranked it amongst combinations with noise loci. Each chart shows a different heritability for the interaction, with higher heritability explained making it more prominent against random noise. The table shows the results of a Mann-Whitney U Test comparing the non-normal distribution of True Positive ranks for a single tool against the distribution of True Positive ranks for all other tools. (Previously published in Russ et al. 2022)

For the case of impure epistatic models (Fig. 2.5), wtest detected one more interaction than MPI3SNP, retrieving 56.7% and 56.5%, respectively. However, the Mann-Whitney U Test demonstrates that the rank distribution of those identified by MPI3SNP was superior, achieving a p-value of $1.11e-30$ compared to $914e-25$. These were followed by MDR with 44.3% ($p = 1.3e-4$) and AntEpiSeeker with 43.0% ($p = 1.15e-09$). In terms of identifying the correct interaction at the highest rank, AntEpiSeeker found 40.5% of such interactions, followed by MPI3SNP with 39.8%, wtest with 35.7% and MDR with 21.2%. Similar to the previous experiments, the joint recessive dataset was the least well identified, with CINOEDV achieving the best performance with four accurate hits, while the other methods failed to detect any interactions.



Tool	p-value
AntEpiSeeker	1.15e-09
CINOEDV	1
MDR	0.0013
SNPRuler	1
wtest	9.14e-25
MPI3SNP	1.11e-30

Figure 2.5: Summary of the impure model results for third order interactions. Each bar chart shows the number of True Positive interactions discovered and the position that the algorithm ranked it amongst combinations with noise loci. Each chart shows a different interaction models (see Table 2.3). The table shows the results of a Mann-Whitney U Test comparing the non-normal distribution of True Positive ranks for a single tool against the distribution of True Positive ranks for all other tools. (Previously published in Russ et al. 2022)

Considering the run-time of each tool (Table 2.8), PLINK was the fastest for the two locus problems, its implementations of BOOST and Fast Epistasis necessitated roughly the same time with different numbers of features. All other tools required more time with additional features, apart from CINOEDV which unexpectedly required less time for greater numbers of features. However, this appears to be due to a small number of outliers that took up to 45 minutes to complete. GSS exhibited the slowest performance, taking over 46 hours to assess 2500 features. When assessing three locus problems, MPI3SNP and SNPRuler exhibited the fastest performance, with SNPRuler being fractionally slower with fewer features but requiring almost 50.0% less time for the larger feature set. The tools with the slowest performance were CINOEDV and wtest, with the latter expending over 14 hours to assess a single dataset. Notably, AntEpiSeeker appears to have performed the three locus problem quicker than two locus problems, but in fact different parameters were used for both datasets as a result of the differing numbers of features. This is because it is recommended that the number of iterations carried out over two stages is a function of the number of SNPs. The first stage, uses a tenth of the number of iterations to the number of SNPs and the second stage, half the number of the first. Since the three locus problems datasets use 500 SNPs compared to 2500, that means there are many fewer iterations required.

The memory requirements (Table 2.8) demonstrate a marked increase with the number of features and how many loci are involved in the interaction. wtest is the most memory intensive tool, necessitating over 14 GB of RAM for the 500 variable, three locus interaction tasks. The PLINK implementations and AntEpiSeeker demonstrated particularly efficiency.

Table 2.8: Average time taken in minutes/RAM used in MB per method at different numbers of loci for a di-locus or tri-locus search. *AntEpiSeeker performed with slightly different settings in three-locus experiments (see Methods).

	Two Locus Detection		**Three Locus Detection**	
Tool	**500 loci**	**2500 loci**	**100 loci**	**500 loci**
AntEpiSeeker	1.89/6	6.35/16	1.32*/6	1.79*/9
CINOEDV	16.20/205	15.70/272	48.18/211	422.18/1346
MDR	0.26/1	6.01/291	0.37/1	22.59/2092
SNPRuler	0.34/1	0.66/306	0.56/1	0.46/2
wtest	0.41/1	7.47/2096	6.73/199	851.33/14302
MPI3SNP	-/-	-/-	0.22/1	1.23/69
Cassi	0.50/1	1.76/156	-/-	-/-
epiACO	13.20/665	20.85/697	-/-	-/-
GSS	29.45/682	2791.15/843	-/-	-/-
PLINK				
Fast Epistasis	0.23/1	0.21/1	-/-	-/-
PLINK BOOST	0.22/1	0.22/1	-/-	-/-
PLINK Epistasis	0.25/1	2.40/1	-/-	-/-

2.3.3 Demonstrative Second Order Interactions in Atrial Fibrillation

Two-locus interaction analysis was performed on UK biobank participants with and without AF (Table 2.9). AntEpiSeeker, CINOEDV and epiACO all found a partner for rs730072, with CINOEDV finding a relationship with rs4668136, as opposed to AntEpiSeeker and epiACO, which both identified rs1152591. MDR and wtest also found a potential interaction with rs1152591, but instead the other locus was rs3792234. GSS and BOOST also found the same pair of SNPs most explanatory, with a relationship predicted between rs9346918 and rs4342945. Similarly, Cassi and epistasis found the same pairs in rs1608994 and rs3809775. This SNP set was significantly over-enriched for the Gene Ontology terms, zinc ion binding ($p = 4.04e-2$) and RNA polymerase II transcription regulator complex ($p = 4.03e-2$). Moreover, there is evidence in the STRING database for protein interactions (Szklarczyk et al., 2019) showing that homologs of *DTNB* and *SYNE2* in *Mus musculus* are co-expressed and interact in various assays, an interaction that was found by two independent tools. This could potentially be a fruitful line of enquiry given that *SYNE2* is known to be associated with AF (Ellinor et al., 2012). Additionally, *LRP2*, which epistatically interacts with *DTNB*, has been linked to arrhythmias in multiple conditions from genomic screens and proteomic assays (Lygirou et al., 2018; Theis et al., 2020).

Table 2.9: Epistatic interactions predicted by each tool as most important for Atrial Fibrillation. An * denotes that the SNP was found to be intergenic and this is the nearest gene

Tool	**SNP1**	**Gene1**	**SNP2**	**Gene2**
AntEpiSeeker	rs730072	*DTNB*/*ARNILA*	rs1152591	*SYNE2*/*ESR2*
Cassi	rs1608994	*MSR1*	rs3809775	*HOXB8*/*HOXB9*
CINOEDV	rs730072	*DTNB*/*ARNILA*	rs4668136	*LRP2*
epiACO	rs730072	*DTNB*/*ARNILA*	rs1152591	*SYNE2*/*ESR2*
GSS	rs9346918	*PRKN*	rs4342945	*PPP2R2D*
MDR	rs3792234	*STON1*	rs1152591	*SYNE2*/*ESR2*
PLINK:epistasis	rs1608994	*MSR1*	rs3809775	*HOXB8*/*HOXB9*
PLINK:FastEpi	9:140746691	*EHMT1*	rs56018060	*CA12*/*LINC02568*
PLINK:BOOST	rs9346918	*PRKN*	rs4342945	*PPP2R2D*
SNPRuler	rs6754266	*LOC105373398**	rs12627212	*RUNX1*
wtest	rs3792234	*STON1*	rs1152591	*SYNE2*/*ESR2*

2.4 Discussion

This study has systematically assessed the performance of several state-of-the-art tools, including popular implementations, such as logistic regression, nature-inspired algorithms and multifactor dimensionality reduction. The evaluation of their performance was carried out for two key categories of epistasis, namely pure epistatic interactions with no main effect, which were randomly distributed and impure epistatic interactions with some main effect, that conformed to several set epistatic models. This evaluation strategy was repeated for both second-order and third-order interactions. The purpose of this has been to inform future research to uncover genetic interactions, reduce the impact of missing heritability and ultimately to better inform use of genetic data as we strive towards personalized medicine.

The experiments revealed that the performance of each tool varies depending on the task that is being assessed. The pure epistatic models represent the most difficult interactions to detect, since the lack of marginal effects necessitate all loci in an interaction to be considered to identify any effect. Each locus individually will not have a main effect and as such its effect would have insignificant p-values in a standard univariate GWAS, with p-values around 1 (Fig. 1). However, the tools which detected the most pure epistatic models were not the same as those that found the most impure models, indicating that tool selection based on significance is a valid strategy.

In the case of pure two locus models, the PLINK implementation of BOOST had the best performance. It was also the fastest to complete the task (Table 2.8). This is perhaps expected, since PLINK is one of the first tools in this space with continuous maintenance and improvements including speed improvements, utility of bitwise operators, multithreading and other techniques. This is similar to MPI3SNP that is also compiled in C++ and appears to have well optimized code for the task as seen in its runtime. A disadvantage of BOOST lies with the lack of a function to include covariates, necessitating their correction in post-processing steps.

Since PLINK only assesses pair-wise interactions, it is not a viable option for third order interactions. For these cases, MPI3SNP detected a higher number of pure epistatic interactions than wtest, achieving a superior distribution of ranks when tested. Of the interactions found, those ranked first made up 72.5% for MPI3SNP, compared to wtest with 81.6% of instances. In terms of runtime and memory requirements, MPI3SNP performed excellently, particularly when compared to the prohibitive requirements of wtest.

Considering impure epistatic interactions, the MDR was most successful at detecting two locus problems in all test cases, except for the joint recessive model. This type of model is the hardest to detect since the interaction only occurs at only one genotype. Since MDR considers all possible genotypes, it is likely that random fluctuations in other genotypes detract from the effect of the affected genotype, which contains the fewest total individuals under HWE. For a minor allele frequency (MAF) of 0.4, with a total of 2,000 samples, there are approximately 50 samples with this genotype in a two locus problem and 7 for three loci. This figure is then split into cases and controls, rendering distinguishing a three locus joint recessive scenario very difficult, even for the instances that all or almost all of the samples with that genotype were cases. GSS retrieved the most of the two loci cases, whilst CINOEDV found four higher order interactions. However, it is questionable how feasible searching for these interactions is without much larger sample sizes. Higher order impure epistatic models returned mixed results, with MPI3SNP again performing well with higher dimensionality but ranking fewer at first place than AntEpiSeeker. wtest and MDR were also notable for detection ability, and perhaps again a joint searching strategy could be employed to give a greater combination of speed and accuracy.

Evidently, this assessment, given that it encompasses a survey of a wider range of tools than previous reviews of this nature, clearly demonstrates that there is no one best tool for all interaction types. GSS has previously been reported as the most powerful algorithm by Chatelain et al (Chatelain et al., 2018) and although the results showed a similar trend, GSS was not the optimum solution for any of the scenarios that were assessed, exhibiting substantially long run times and memory requirement. Alchamlat et al (Abo Alchamlat et al., 2018), reported MDR as the best performing tool and the results indicate that, although it did not exhibit the best performance across the datasets that were used, its performance was the most consistent. Furthermore, there was little to suggest an ensemble of tools would provide better identification of interactions. If the interaction was to be retrieved by any tool, the most proficient tool, in each scenario, was almost always the one that retrieved it and as such, multi-tool ensemble-based approaches were not found to be useful (for key features of each tool see Table 2.10).

Table 2.10: Key features for each tool

Tool	Statistical Test	Exhaustive Search	Mem. Use	Higher Order	Missing Data	Covariates
AntEpiSeeker	χ^2	No	Low	Yes	No	No
CINOEDV	Co-information	Both	High	Yes	No	No
MDR	Permutation Testing	Yes	Med	Yes	Yes	Yes
SNPRuler	χ^2	No	Low	Yes	Yes	No
wtest	W-test	Yes	High	Yes	No	No
Cassi	Logistic Regression	Yes	Low	No	Yes	Yes
epiACO	Mutual Information	No	High	No	No	No
GSS	Min-Max Optimization	Yes	High	No	No	No
PLINK:BOOST	Log-Linear Regression	Yes	Low	No	Yes	No
PLINK:FastEpi	Z-Score	Yes	Low	No	Yes	No
PLINK:epistasis	Logistic Regression	Yes	Low	No	Yes	Yes
MPI3SNP	Mutual Information	Yes	Low	Yes	No	No

The application of the tools to the UK biobank AF data highlighted limitations found in some of the tools, such as the inability to handle missing genotypes and the memory requirements needed to run wtest when applied to a larger dataset. Given *SYNE2* has previously been associated with AF (Ellinor et al., 2012), that does lend some credence to the interactions involving that gene, that was found by four tools. Interestingly, there was evidence in STRING (Szklarczyk et al., 2019) for an interaction between homologs of *DTNB* and *SYNE2*, being both co-expressed in *Mus musculus* and interacting within various assays. This perhaps supports the use of AntEpiSeeker and epiACO. However, more research would need to be carried out. The experiment did demonstrate some patterns amongst the tools. Unsurprisingly, both PLINK's epistasis and Cassi identified the same interactions, since they are both variants of logistic regression. Interestingly, there was also exact agreement between GSS and BOOST, as well as wtest and MDR and a high-level of similarity amongst the nature-inspired methods. This indicates that perhaps some concordance between how these methods approach the data that has led them to rank highly the same interactions.

A key limitation to this paper is the focus on detection capabilities of tools at a smaller scale than a regular GWAS. The next step will involve transitioning to larger feature spaces of around 750,000 initial loci and assessing different strategies for searching efficiently, especially since the time and memory requirements for spaces of this size will be very large. This will therefore require testing strategies that can be applied at a larger scale, and is therefore beyond the scope of this paper. Using a faster more efficient tool initially, followed by a quicker more accurate one could be a potential solution. Alternatively, individual features could be ranked first using a Relief-based algorithm, such as multiSURF (Ryan J. Urbanowicz et al., 2018). Furthermore, since different tools perform better for pure and impure epistatic problems, a division of the data can be made using the p-value of individual loci generated by logistic regression. Finally, this research is limited to second and third-order interactions. Further work is planned to be carried out for higher-order interactions, but establishing feasibility at lower levels has been carried out preferentially due to the power constraints which are increased at each dimension. Since the loci exist in HWE, their genotypes containing fewer samples is unavoidable, an effect that is compounded with additional dimensionality.

It is notable that there are a large number of epistasis detection tools in existence, and therefore it is plausible that some of them would exhibit better performances than the selection assessed here. The aim was to provide an independent, objective performance assessment over a representative selection of available tools. The inclusion criteria considered whether it was feasible for a tool to be assessed under the study constraints, and hence tools designed for quantitative phenotypes as well as tools that relied upon the wider genome context were not included. Examples of the latter would be Eigen-epistasis (Stanislas et al., 2017) that calculates eigenvectors for sections of the genome or GenEpi that uses gene boundaries to group loci (Y.-C. Chang et al., 2020).

The practical application of this research is dependent upon further development of these approaches in order to create workflows to find interactions at scale for many diseases and phenotypes. In the clinic, the additional genetic markers have the potential to be used in diagnostic applications and as targets for therapeutic remedies. Owing to the substantial effort expended to carry out GWAS, there is a significant drive to ensure there are beneficial medical advances as a result(Shu et al., 2018; Lau et al., 2020). Quantifying the heritability and effect size of the interactions allows for their incorporation with known single loci involved in the disease. Inclusion of interactions within models for PRS has been shown to affect overall accuracy(Dai et al., 2020). Additionally, *a priori* data about interactors can guide further research into particular diseases and enhance biological interpretability. Knowledge can be revealed using powerful resources, such as Reactome and KEGG datasets for pathway analysis, Gene Ontology for protein function, processes and cellular location and STRING Database for protein-protein interactions(Fabregat et al., 2017; Kanehisa et al., 2017; Ashburner et al., 2000; Szklarczyk et al., 2019).

The approach revealed that different tools are optimum for different challenges. For detecting pure, two locus interactions, BOOST, as implemented by PLINK, was most effective, with low runtime and memory requirements. For impure epistatic interactions, MDR retrieved the highest number of correct interactions. For the more computational costly cases, MDR offers options for covariate correction. Finally, for detecting three locus interactions, MPI3SNP exhibited the best performance, with the minimal computation requirements notable for this challenge. How these can be applied to genome-wide datasets is the next challenge.

Chapter Three

Detection on a Genome-Wide Scale

3.1 Introduction

3.1.1 Genome-Wide Epistasis Detection

An effect of the proliferation of GWAS and large biobanks, is the abundance of data produced. This presents enormous opportunities, but also the possibility to partially squander these if insights are missed. It is of vital importance to ensure that disease drivers are identified so as to maximize the resulting medical benefits. However, it is not common practice to include searches for genetic interactions in GWAS. This is due to lack of best practices, onerous resource requirements and technological hurdles, with Ritchie comparing the process to finding an 'epistasis needle' in a 'Genome-wide haystack' (Ritchie, 2015). There are two broad strategies for epistasis detection from these datasets, to deal with these challenges. The first is using *a priori* knowledge to curate genetic targets to test, and reduce the problem of high dimensionality. This can be useful if a specific condition is being studied and resources are limited. A downside of this is that it will introduce an historical bias, in which genes previously associated with a phenotype are tested under a reduced significance threshold. This leads to less rigorous testing, despite a greater amount of data being available that has simply been excluded. The second approach is untargeted, fully utilizing the large GWAS datasets. As discussed in the previous chapter, this introduces high significance thresholds under multiple testing but can implicate genes previously unassociated with a phenotype and find unknown genetic relationships, as well as provide a greater surety in the associations discovered.

In the prior chapter, methods were compared using a smaller dataset. In this chapter, the aim is to increase the scale to biobank size datasets and apply the most effective workflow to biological data. The first stage is a simulation study carried out on Chromosome 22, as extracted from UK biobank, with models of epistasis fitted in to the linkage structure. Following this, the most successful pipeline is applied to three diseases, Atrial Fibrillation, Parkinson's Disease and Alzheimer's Disease. Finally, analysis of the most significant relationships is carried out to understand the mechanisms behind the associations.

3.1.2 Challenges and Considerations

Genetic Structure

In the previous chapter, the factor of chromosomal structure was largely avoided. However, conserved inherited sequences of DNA provide a specific challenge when testing for epistasis, but also potential opportunities. The non-random patterns of LD across a chromosome are structured broadly into haplotype blocks. These are areas in which recombination during meiosis rarely happens, resulting in a set of associated groups of genotypes, known as haplotypes. These haplotypes form part of the species' evolutionary history, with mutations retained as a result of random genetic drift or some beneficial evolutionary aspect(Clark, 2004).

The challenge posed by LD is essentially confined to loci on the same chromosome, since outside of translocation of a section of a chromosome to another, genetic material is not generally shared between different chromosomes (Paththinige et al., 2019; Vasilevska et al., 2013). However, as stated in equations 1.3 and 1.5, LD is the correlation of the genotypes of different loci. As such, any relationship between the loci is likely due to their intrinsic correlation rather than a genetic interaction. This can present over a long range, with LD observed at loci 1,747,249 base pairs apart (Park, 2019). However, it is also possible that the simultaneous conservation of mutations is due to a critical epistatic relationship of the conformations of those genes affected. This has been observed, for example, in mutations in the gamete recognition genes $ZP3$ and $ZP3R$ (Rohlfs et al., 2010).

As we have seen previously, it is in part due to this structure that GWAS has been so successful. When proposed by Risch, it was in utilizing LD that genetic markers could be used as proxies for a region (Risch et al., 1996). This has also allowed the process of imputation of proximal variants to be implemented from panels made up from data such as the HapMap project, 1000 Genomes and others (Altshuler et al., 2005; Auton et al., 2015). Similarly, a variant can be used as a proxy for its local region via LD when assessing interactions. This means that the variants assessed for interaction may not be the causative loci but ones that are correlated, with that level of correlation defined by the level of LD between the causative loci and their proxies. The issue with this is a problem of deteriorating power. In GWAS, the most significant variant found is unlikely to be the causative variant, but one in LD with that variant. As such, there is the potential that the association is weakened compared to if the variant driving the trait had been tested. This is a limitation of SNP arrays when compared to sequenced data.

The method of determining haplotype blocks is achieved using a process called LD pruning. This is carried out in a moving 'window' of variants, with its size determined by the physical distance or a given number of variants. As the genetic material is scanned, a metric such as r^2 is calculated to determine the structure in that window and group variants that are in LD by a set amount. This way, blocks can be built along the chromosome. After the blocks have been determined, a process called clumping is often applied. This is a way of selecting a variant from the block by taking the one that achieved the greatest monogenic association, when assessed against the trait in question. The resulting collection of loci are then 'representatives' of their block (Purcell et al., 2007). Indeed, this method of dimensionality reduction has been utilized in network-based analyses so that a node represents a correlated region of the genome, scans of HapMap phenotypes and as part of the suggested protocol of tools such as MB-MDR (Lareau et al., 2015; Becker et al., 2012; Joiret et al., 2019).

Models of Epistasis

As discussed in the previous chapter, we can model epistasis as pure, or impure, dependent on the monogenic significance of the individual variants. We have observed that different tools are better adapted to each of these situations, and as such, a strategy to tailor tools to models they perform better on can then be employed. This can be achieved using the variant's significance when assessed against the trait, and those below a pre-determined significance can be assessed by one tool and those above by another. This is an extension of the method suggested by Evans and Kooperberg, in which variants are assessed if they have a monogenic significance below a selected threshold (Evans et al., 2006; Kooperberg et al., 2008).

As previously discussed, there are hundreds of possible models, from more ordered conformations adhering to Mendelian Inheritance, such as joint dominant, whereby a minor allele of two genes is require to cause the phenotype to less well ordered conformations that are more stochastic. These don't fit well under a linear model interaction as a product of two variants, since they are non-linear patterns. In the previous chapter we saw that tools that assessed interactions by looking at each combined genotype performed better, and so in these simulations a collection of different models should be generated to assess the approaches.

Multiple Testing

The principle of multiple testing correction is required because the standard significance threshold used of 0.05 is a probability. As more hypotheses are tested, a false positive result is therefore more likely, so that threshold must be adjusted. The number of tests that are carried out in exhaustive epistasis detection is a function of the number of variants tested. The genome-wide *p-value* significance threshold of $5e-8$ assumes $10e^6$ independent regions of the genome. This means that the number of tests required in an exhaustive search is a two element subset of $10e6$ elements, denoted ${}^{1,000,000}C_2$, which is slightly fewer than $5e^{11}$ combinations. This means that for the Bonferroni adjusted significance threshold, 0.05 is divided by the number of tests, giving a threshold of $1e-13$ ("Etymologia" 2015). The Bonferroni correction method is seen as being prohibitively stringent, with a common method to reduce the burden being the Benjamini-Hochberg False Discovery Rate (FDR) (Benjamini et al., 1995). Proposed to reduce the number of erroneous rejections, false negative results, it is an extension to the Bonferroni correction. Resulting $p-values$ are ranked from smallest to largest, through iterations i and the adjusted *p-value* is calculated:

$$P(i) = \alpha \times \frac{i}{m} \tag{3.1}$$

with α being the type I error rate, normally 0.05 and m being the number of tests carried out. With this method, the most significant result is assessed using the Bonferroni correction, but the adjustment is less severe moving through the ranked list. The limit of a value's correction is the significance of the previous item in the list. As such, if statistical significance is not reached by the first value, it will not be reached by any.

3.1.3 Applications for Atrial Fibrillation, Alzheimer's Disease and Parkinson's Disease

In the previous chapter, there was a small-scale example of epistasis detection in atrial fibrillation (AF). There are quite a number of loci known to be associated with AF, with one study noting 138 variants (Roselli et al., 2020). In part, the discovery of so many variants is due to the prevalence of the disorder, affecting around 2% of Europeans, and also the severity, as patients suffering from AF are more likely to have strokes or die, so medical interventions are required (Zoni-Berisso et al., 2014). There has been a small study looking at epistasis in atrial fibrillation, identifying a potential interaction between $T174M$ and $M235T$ (Moore et al., 2006). Additionally, there have been some studies that have identified potential interactions from biological experimentation, one finding a relationship between $PDE4DIP$ and DES in a familial study and another proposing epistatic effects of rare mutations in K^+ channels (Abou Ziki et al., 2021; Mann et al., 2012).

To extend this study further, we will also apply the pipeline to the two neurodegenerative disorders, Alzheimer's Disease (AD) and Parkinson's Disease (PD). These are conditions that cause great suffering, affecting an estimated 33.9 million and 6 million people worldwide, respectively (Barnes et al., 2011; Bloem et al., 2021). AD has been the focus of a project specifically aimed at identifying interactions, called the Epistasis Project. One paper was targeted at $BDNF$, DBH and $SORT1$, finding some evidence of interactions(Belbin et al., 2019). This approach, of targeting mutations in specific genes also yielded a potential interaction between HFE and TF, as well as $IL-6$ and $IL-10$ (Lehmann et al., 2012; Combarros et al., 2009). However, they use a very loose significance threshold, that may make sense within their experiments, but in terms of genome-wide significance it is perhaps not sufficient. Other groups have continued this practice, focussing on genes associated with late-onset AD, with interactions claimed between a number of genes (Raghavan et al., 2017). Evidence for epistasis in PD is less common, although some studies have been carried out in a similar targeted way, such as with genes in the $mTOR$ pathway (Bandres-Ciga et al., 2020; Fernández-Santiago et al., 2019).

3.2 Materials and Methods

3.2.1 Study Design

The study was designed to follow two phases, firstly a simulation study examining different approaches to epistasis detection over Chromosome 22, in order to assess strategies for finding interactions. Chromosome 22 was used since a significance threshold is being tested against, so interactions can be assessed empirically, and as the shortest chromosome it is less computationally intensive for hundreds of simulation tests. This was to understand the most effective strategies and assess the feasibility of the approaches in terms of computational and time requirements. Secondly, the methodology that was deemed to be most effective was applied to three real case-control datasets from UK biobank. Interactions found were then assessed for their contribution towards the disease model.

3.2.2 Chromosome 22 Simulation

Data Generation

The basic genetic material used in the experiments was from GWAS data obtained from UK biobank, to which simulated models of epistasis were added programmatically. A random selection of 100,000 individuals was used, with all being classed as white British, unrelated, having no sex mismatch or aneuploidy and not having been flagged for outlying heterozygosity levels or missingness. To limit the feature space and test for epistatic models which would have a more even spread of allelic combinations likely, a minimum MAF of 0.01 was used. This is because the genotype consisting of entirely minor alleles for two variants would contain $0.01^4 \times 100,000 = 0.001$ individuals, so is unlikely to be populated at this level for noise variants. A genotyping filter for missingness was used of no greater than 5% and all variants having a p-value for departure from HWE of no less than $5e-8$. After these filters, 9,644 variants remained. PLINK was used to calculate r^2 values for LD across the chromosome The ld-window and ld-window-kb flags that set the size of the moving window when calculating LD statistics, were set large enough to capture the entire chromosome. This was carried out so that the interaction models could be placed within LD structures, but whilst being located in unassociated haplotype blocks.

Epistatic models for two loci were generated via two methods. For the pure-epistatic models GAMETES v2.2 was used, generating models with heritability scores of 0.005, 0.003 and 0.001, MAF of 0.1, 0.2, 0.3 and 0.4 as well as three levels of Ease of Detection Measure. A custom R script was written to generate impure model-based interactions at three different odds ratios, at 1.2, 1.4 and 1.6, as well as MAF 0.1, 0.2, 0.3 and 0.4. The models used were based on Mendelian forms of inheritance, in five different configurations. Since HWE dictates the distribution of individuals across the 3-by-3 grid, each of the four rotational versions of the models were generated.

The models were fitted into the chromosomal material using an approach coded in R. Here, the first variant in the interaction is matched to an existing variant by genotypic similarity, as in LD, then the phenotype is generated and the second interacting variant is fitted to a remaining variant, outside of LD from the first interacting variant. This placed both variants within the haplotype structure of the chromosome.

Detection Methods

Five methods were used to scan the data, utilising the differing strengths of MDR and BOOST, as the two best performing algorithms from the previous chapter (Figure 3.1). With MDR it is not feasible to conduct an exhaustive search, even using the parallel option, however with BOOST this method was used. For both, default parameters were used. As mentioned in the introduction, reducing the dimensionality can be achieved by dividing the chromosome into haplotype blocks. This was done using PLINK's –blocks flag with the no-small-max-span option, across a maximum range of 500 kbp. Afterwards, a random variant from the block was selected for the experiment. Any which had not been detected as part of the haplotype blocks were included as well. This resultant datasets were tested by both BOOST and MDR individually. In Chapter 2, BOOST had performed best at finding pure epistatic interactions, whereas MDR was better for impure interactions. From this insight, another test was carried out by separating the loci using a monogenic significance of 0.05. Those not reaching that threshold were tested using BOOST, and those that did were assessed with MDR. This allows each to be targetted at the types of interactions they are more likely to locate as a combined approach. Finally, the latter approaches were combined, so that the monogenic significance was calculated, and haplotype blocks formed in a chimeric setup.

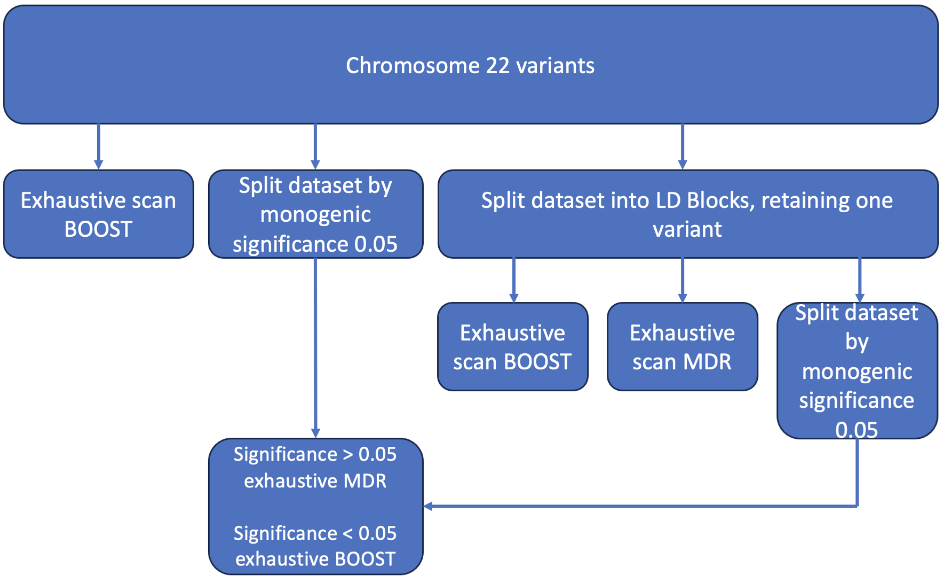

Chromosome 22 variants
Exhaustive scan BOOST
Split dataset by monogenic significance 0.05
Split dataset into LD Blocks, retaining one variant
Exhaustive scan BOOST
Exhaustive scan MDR
Split dataset by monogenic significance 0.05
Significance > 0.05 exhaustive MDR
Significance < 0.05 exhaustive BOOST

Figure 3.1: Flowchart to show data segmentation for different testing strategies implemented

Two significance thresholds were implemented in the simulation study. Since the test was being carried out on Chromosome 22, a Bonferroni corrected value based on the total number of loci included. A second threshold was also used for genome-wide significance. Calculated again with a Bonferroni correction and based on the standard GWAS significance threshold of ($5e^{-8}$):

$$0.05 \times \left(\frac{0.05}{5e^{-8}} \times \left(\frac{0.05}{5e^{-8}} + 1 \right) \right)^{-1} \tag{3.2}$$

The basis of this equation is the standard GWAS multiple threshold value as defined by Risch of $1e6$ (Risch et al., 1996).

3.2.3 Application to Real Datasets

To expand the application of this pipeline beyond that in chapter 2, as well as a more comprehensive test of AF, AD and PD have also been included for examination. These have been the focus of previous studies for epistasis, as noted in the introduction, so are appropriate candidates here. The UK biobank data was filtered to include only those that have no 3rd degree relatives, are categorized as 'White British', are not flagged outliers for heterozygosity or missingness and do not have sex chromosome aneuploidy. The variants were filtered with a minimum MAF of 0.05, maximum missing rate of 0.05 and departure from HWE p-value $5e-20$. The cases were selected based on recorded ICD codes in the diagnosis fields, as well as the date of diagnosis fields (WHO, 2003). The codes used were based on the mappings provided with phecodes (P. Wu et al., 2019) but in summary, Alzheimer's Disease is covered by F00.0, F00.1, F00.2, F00.9, G30.0, G30.1, G30.8 and G30.9, Parkinson's Disease by G20.0 and G21.4 and AF by I48.0, I48.1, I48.2, I48.3, I48.4 and I48.9 and the available equivalents with the date of the diagnosis.

The data were first tested for non-epistatic, monogenic associations using regenie, with covariates included of age, sex, array type and ten principal components (Mbatchou et al., 2021). The SNP array data were first assessed using BOOST exhaustively, retaining any that were assessed to have a p-value lower than 0.001. A database that was created from data extracted from PhenoScanner, which is based on 1,000 Genomes data with LD data for all pairs of variants with a minimum r^2 of 0.8. From this any variants that were in linkage with those found by BOOST to have an interaction were included from the imputed data (Kamat et al., 2019). From this, the imputed genotypes of UK biobank were included, with a minimum INFO score of 0.3 and MAF of 0.01. BOOST was used to test this exhaustively. Then, with only variants that were involved in an interaction with a p-value less than 5e-4, a new dataset was created with variants all in linkage equilibrium, using the PLINK function –indep-pairwise through a window size of 1 Mbp, moving 80 variants at a time and with an r^2 threshold of 0.8. From this smaller dataset, MDR could be executed (see Figure 3.2).

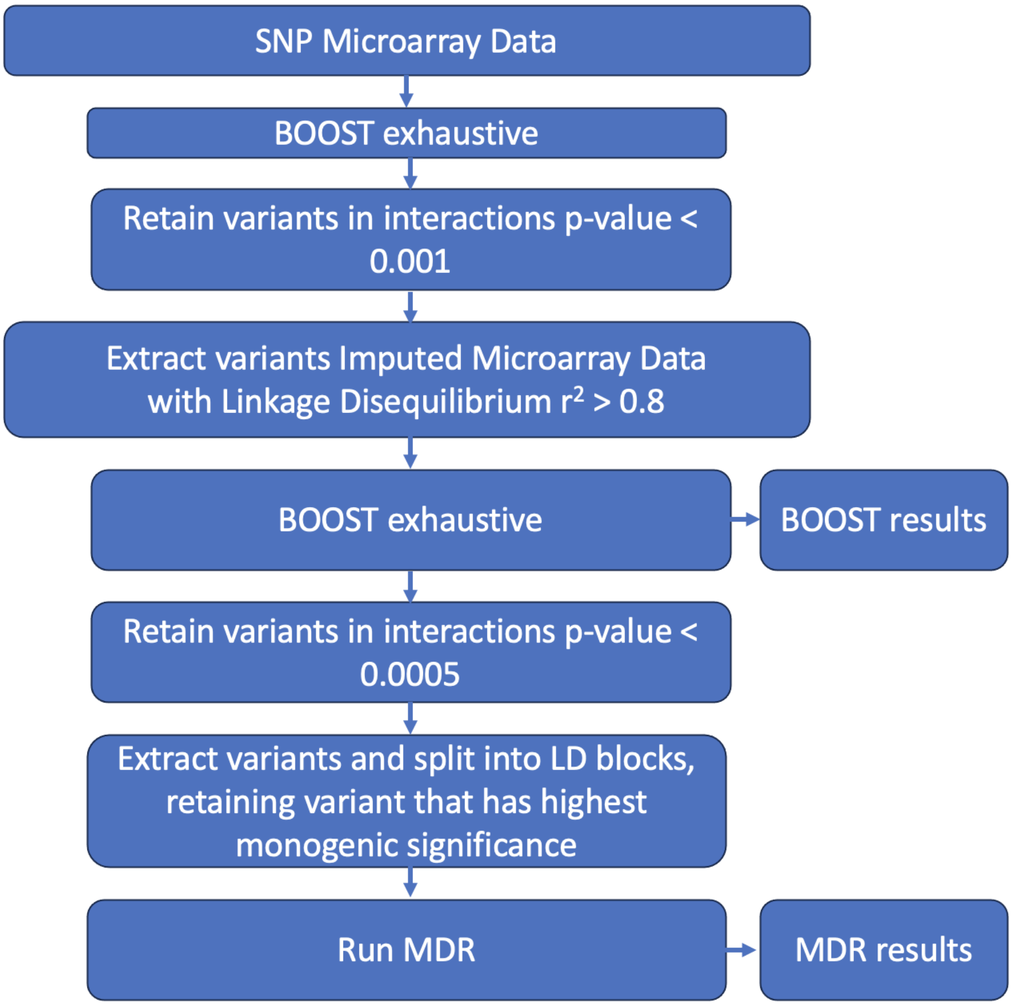

SNP Microarray Data
BOOST exhaustive
Retain variants in interactions p-value < 0.001
Extract variants Imputed Microarray Data with Linkage Disequilibrium r² > 0.8
BOOST exhaustive
BOOST results
Retain variants in interactions p-value < 0.0005
Extract variants and split into LD blocks, retaining variant that has highest monogenic significance
Run MDR
MDR results

Figure 3.2: Flowchart to show stages involved applying pipeline to UK biobank

When using GWAS data, it is important to understand that the variant that is detected is most likely not the causative locus, but a proxy that is in linkage with it. As a result, loci that were identified were scanned against the 1000 Genomes panel using LDlinkR (Myers et al., 2020). Any positions that were not annotated using this were found using BioMart (Smedley et al., 2009). Each pair of interacting loci were assessed using g:Profiler, stipulating that both loci must have the tag found and a significant *p-value* (Reimand et al., 2007). The String database was then queried to find any known interactions in the results (Szklarczyk et al., 2021). Finally, the interaction was assessed, genotype-by-genotype. Each of the possible genetic configurations were tested to find key statistics such as the interaction's heritability, as variance explained found using a pseudo R^2 test to find variance explained and odds ratio. The interaction was defined as any of the genotypes in which the odds ratio at the lower 2.5% tail of the confidence interval exceeded 1 as calculated by a Fisher's Exact Test in R.

3.3 Results

3.3.1 Chromosome-Wide Simulation

Detection Methods

The simulation experiments confirmed what we saw in Chapter 2, BOOST performed best when considering pure epistatic models but is surpassed by MDR when assessing impure models (Fig. 3.3). When both were applied but split by monogenic significance, performance was only minimally affected in the pure experiments and outperformed BOOST on its own. The approach was also fairly robust when split into LD blocks in impure experiments. Experiments in which the data were divided into LD blocks were unable to detect many of the interactions. The patterns observed in Chapter 2 were still in evidence in these experiments, however, with BOOST performing better than MDR for pure epistatic models and MDR detecting more impure interactions.

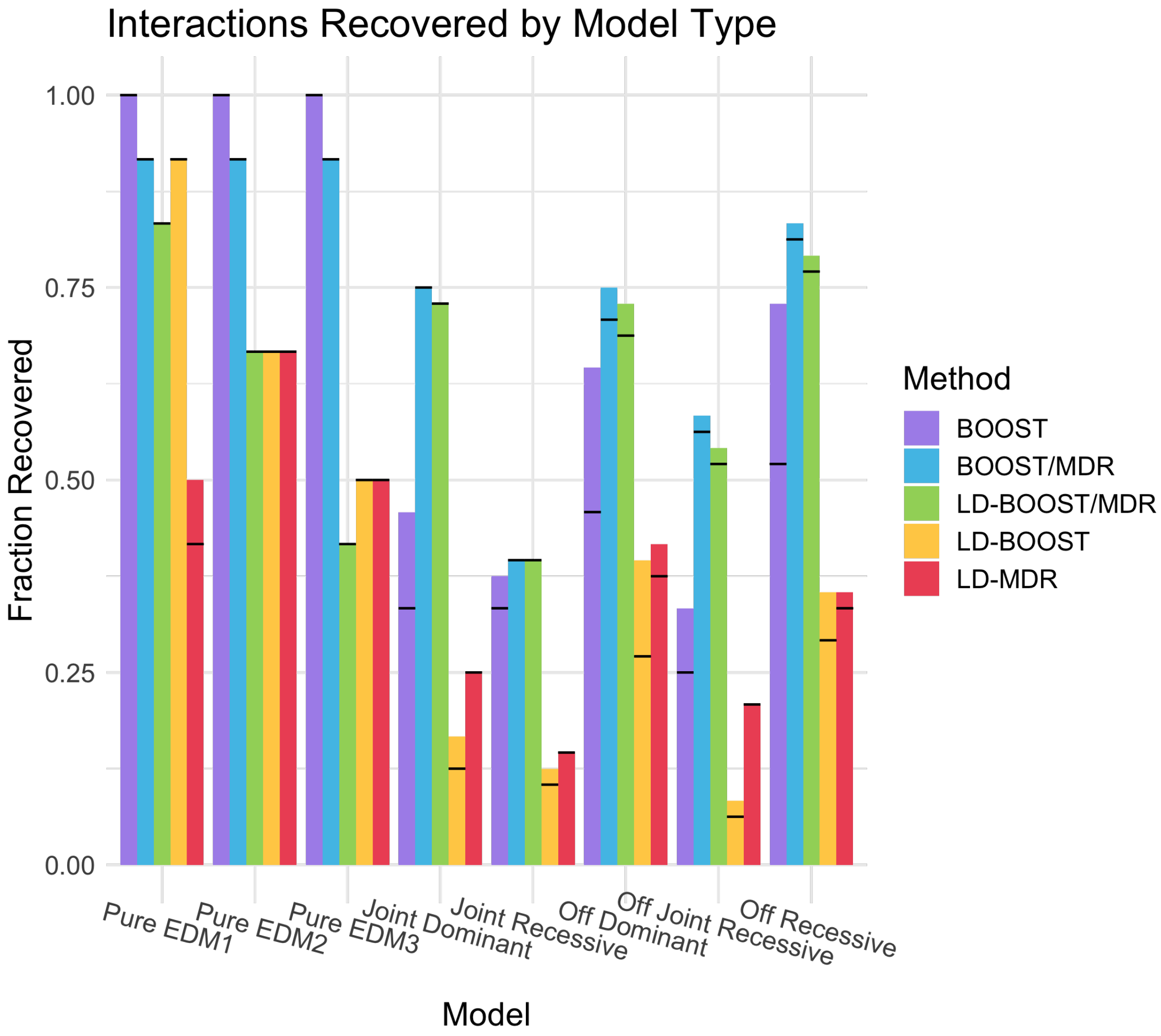


Figure 3.3: Bar chart showing the fraction of epistatic interactions for pure models at three different detection difficulties and five impure models based on Mendelian models. Each bar shows the number of interactions detected at the chromosome 22 significance threshold, with the black line representing the number recovered at whole genome significance threshold

In Figure 3.4, the fraction of total, true interactions detected is given, against a *p-value* threshold for detection, between $5e-2$ to $5e-15$. This showed that BOOST alone was superior at detecting interactions at lower *p-values*, but at thresholds less than $5e-4$ it rapidly stopped detecting interactions, mostly those that were impure models. However, when MDR was used to detect variants with a marginal effect as well as BOOST for those without a marginal effect, this reduction in performance was less apparent.

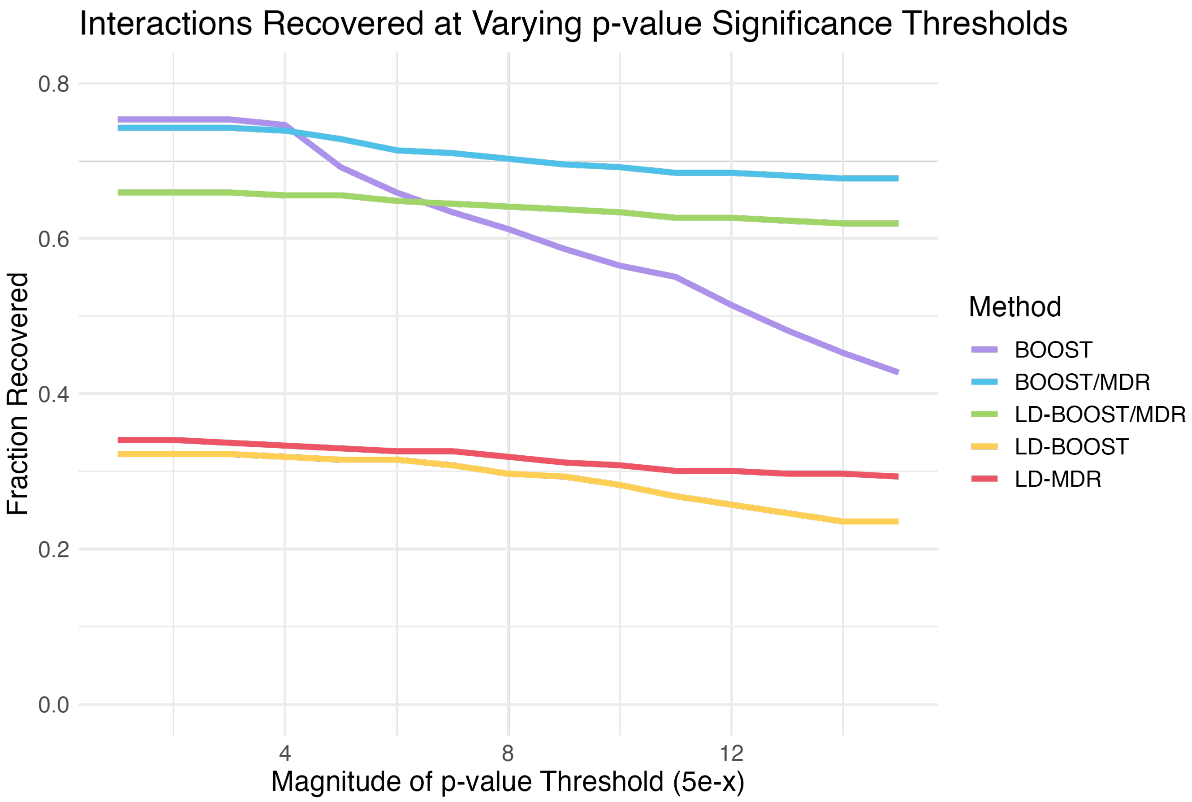


Figure 3.4: Chart showing the fraction of interactions recovered per strategy for different significance thresholds. The x axis $5e-x$ refers to the changing significance threshold to include interactions as detected from $5e-2$ to $5e-15$

We can see in Table 3.1 that methods combining the two algorithms performed best. However, the combined approach that has been split into LD blocks misses 16 interactions, compared to the exhaustive approach. These two methods together only fail to detect 10 interactions, that were uncovered by other methods. Of these, five were found by MDR as split by LD blocks, three by an exhaustive scan with BOOST and two by both of these methods. This indicates that some were missed by splitting the data by monogenic significance.

Table 3.1: Table Showing Interactions Discovered per Method, Excluding Those Missed By All Methods

Method	Total Detected	Total Undetected	Total Undetected By Others
BOOST	127	71	3
BOOST & MDR	188	10	3
LD-BOOST & LD-MDR	172	26	0
LD-BOOST	66	132	0
LD-MDR	82	116	5

The QQ-plot in Figure 3.5 shows the distributions of $-log10(p-values)$ for real interaction pairs, those including half of the pair and those of 'noise' loci. The chart is a bit lopsided, since a cut-off has been applied to more clearly demonstrate how the variants within the interaction present with greater significance than those containing one of the pair. In this particular case, even given this was limited to just chromosome 22, it is notable that the significance achieved by the noise loci is far below the significance threshold.

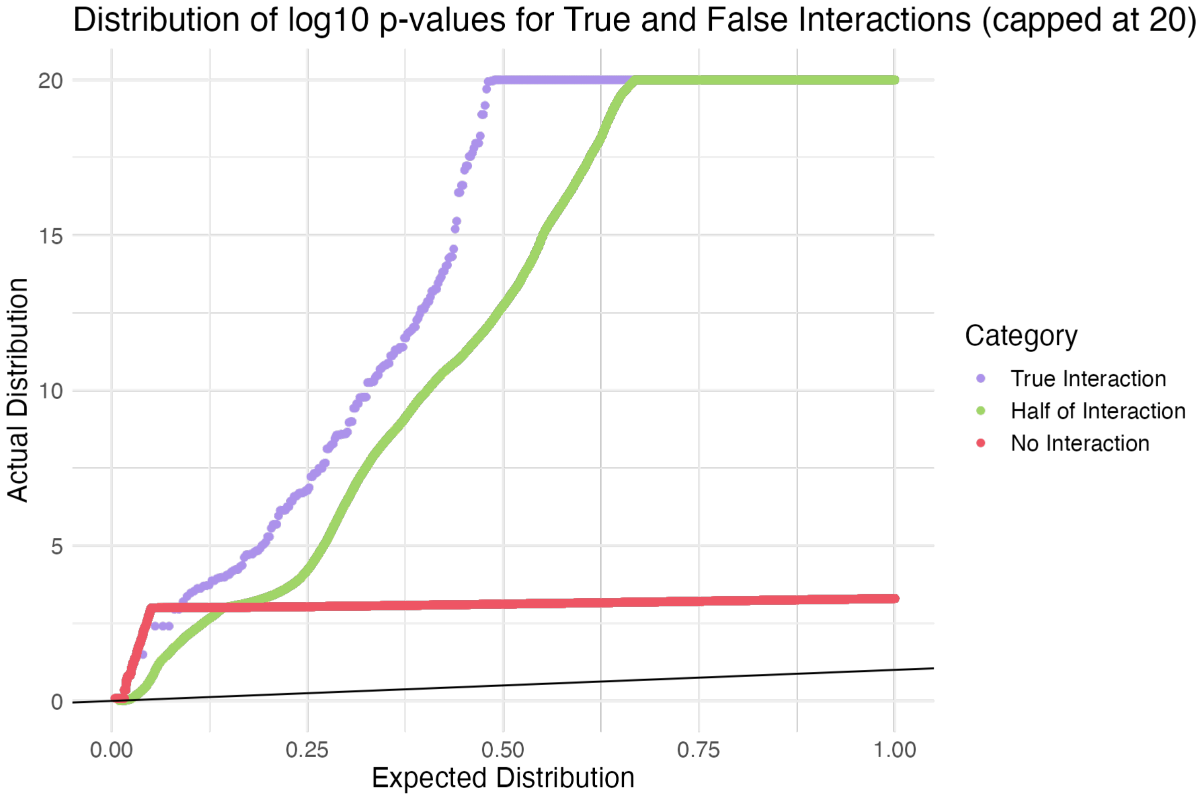


Figure 3.5: Q-Q plot showing the distribution of -log10 p-values for true interactions, those that contain one of two interacting variants and false interactions (capped at 20)

3.3.2 Atrial Fibrillation

Since the combined method using both BOOST and MDR performed the best, it was implemented in tests across all autosomes. The AF experiments with BOOST did not achieve statistical significance but did demonstrate a clear increase in heritability explained as in Figure 3.6. There were nine interactions with significant terms found by g:Profiler. This included an interaction between $AK9$ and $RELN$ that were tagged with the terms GO:1902078 for 'positive regulation of lateral motor column neuron migration' and GO:0006756 'AMP phosphorylation'. There is also the suggestion of neural activity in the interaction between $FSTL4$ and $MYO5B/LIPG/SNHG22$ with 'negative regulation of brain-derived neurotrophic factor receptor signalling pathway' (GO:0031549). Two interactions contained terms related to the immune system, $APC6$ and $LILRB4$ involved in 'negative regulation of cytotoxic T cell differentiation' (GO:0045584), as well as $NSUN2$ and $IL22RA2$ with 'interleukin-22 receptor activity' (GO:0042018). Perhaps also of note is the interaction between $DCHS2$ and $LDHD$, that are both involved in 'D-lactate dehydrogenase activity' (GO:0008720). STRING-db found a medium confidence interaction between $ARL6$ and $PLEKHG4B$.

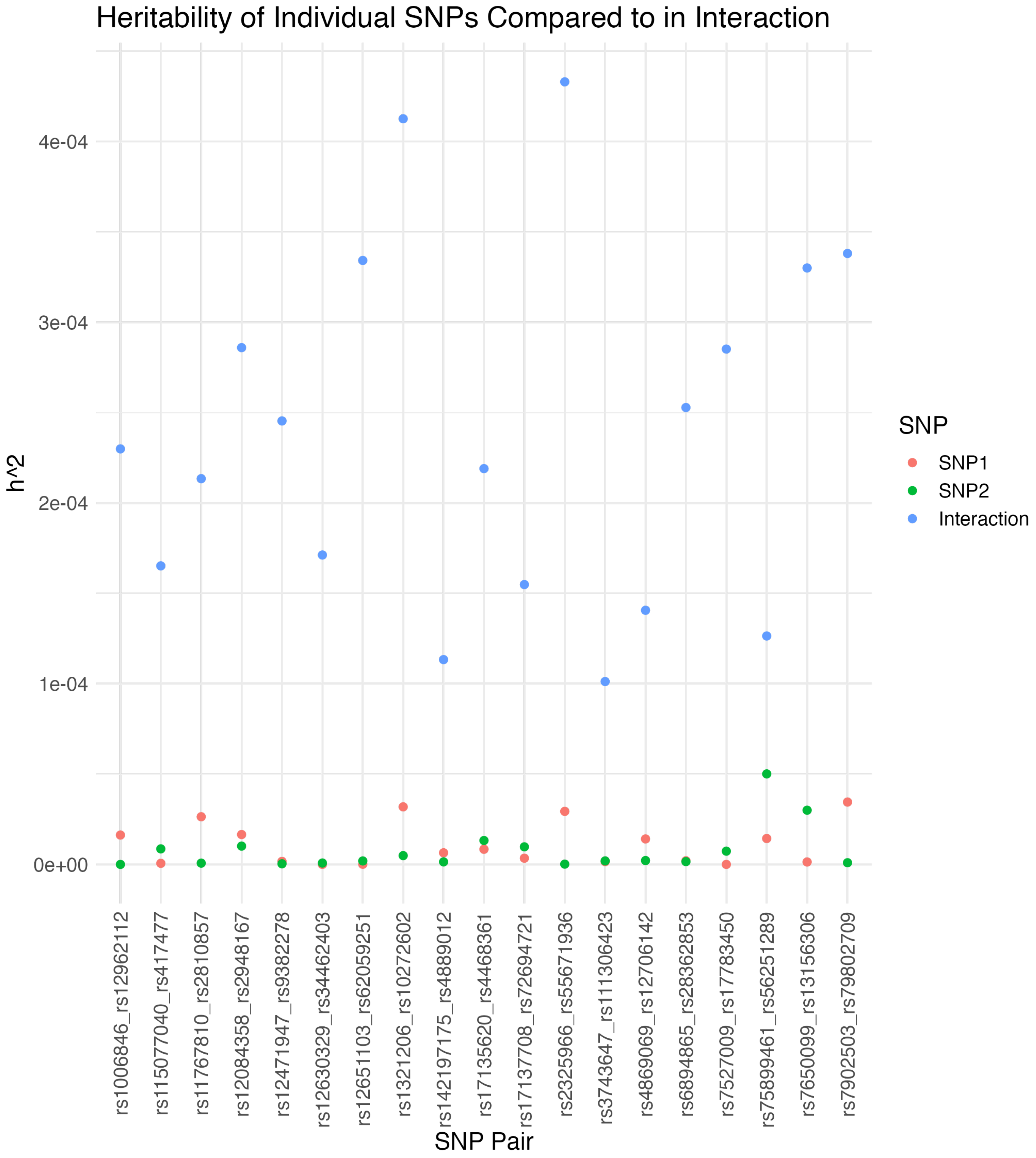


Figure 3.6: Difference between heritability explained of interaction compared to two SNPs in Atrial Fibrillation as detected by BOOST

This trend was less clear when looking at results from the MDR in Figure 3.7, in some cases a variant alone explaining more variance than the interaction. Similar observations were true regarding $p-values$ odds ratios. Interestingly, MDR recovered an interaction that BOOST did also, between $RIOX2/EPHA6$ and $SLC9A3$, which are both implicated in 'Elevated fecal sodium' (HP:0032484). Another interaction between $FZD10-AS1$ and $CORO2B$ was linked by 'talin binding' (GO:1990147).

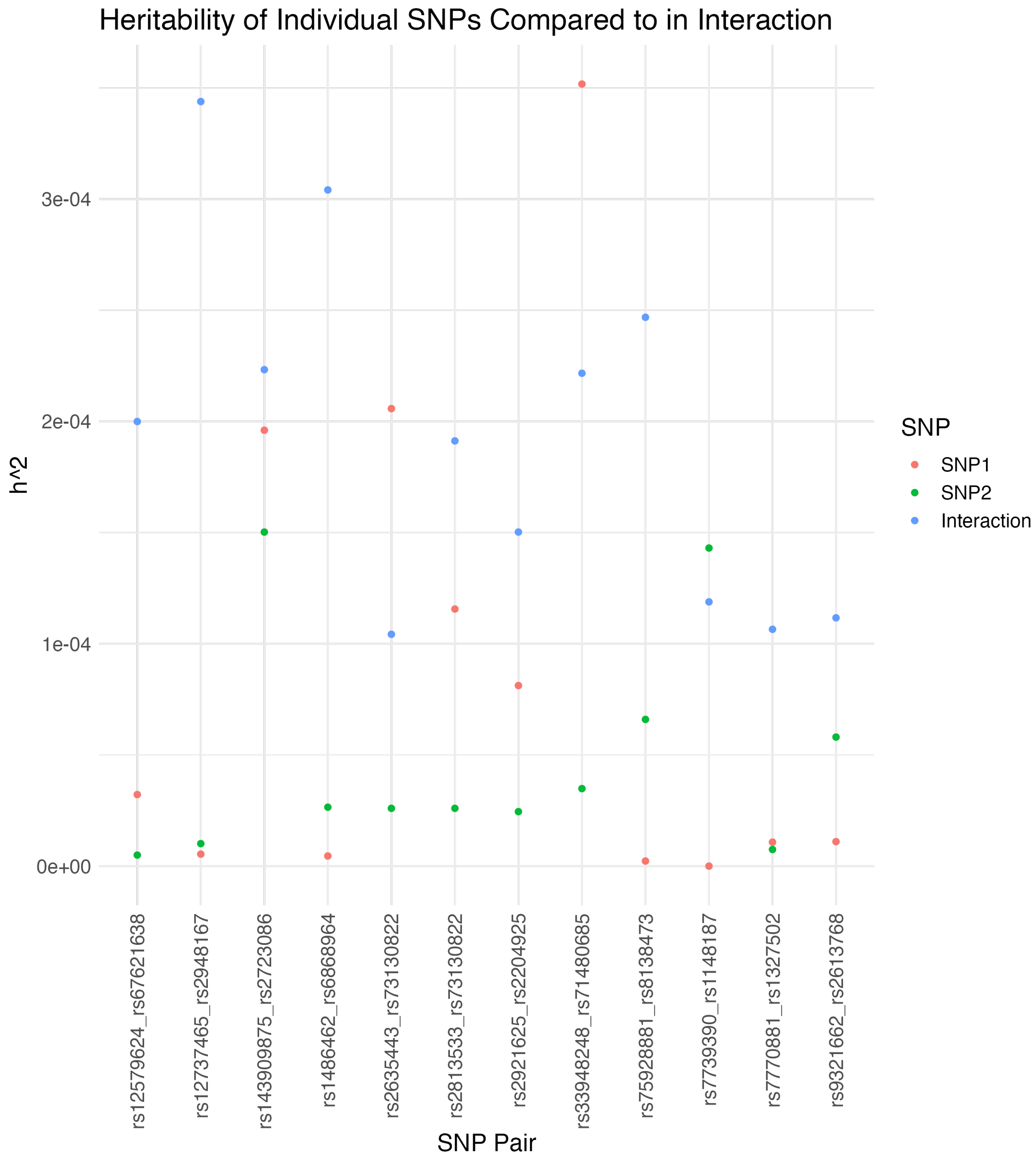


Figure 3.7: Difference between heritability explained of interaction compared to two SNPs in Atrial Fibrillation as detected by MDR

3.3.3 Alzheimer's Disease

As with AF, we can see that BOOST has found interactions with increased heritability across all interactions. There was an interaction found in STRING-db between $SLC35A1$ and $TMEM5$ with medium confidence (0.595) based on having been co-mentioned in PubMed abstracts both in $H.Sapiens$ and other organisms. These shared the terms 'CMP-N-acetylneuraminate transmembrane transporter activity' (GO:0005456), 'Decreased platelet glycoprotein Ib' (HP:0031156) and 'Defective SLC35A1 causes congenital disorder of glycosylation 2F (CDG2F)' (REAC:R-HSA-5619037). There were nine interactions with annotations. However, these were not found to be relevant to the condition. For the MDR, no useful interactions were found.

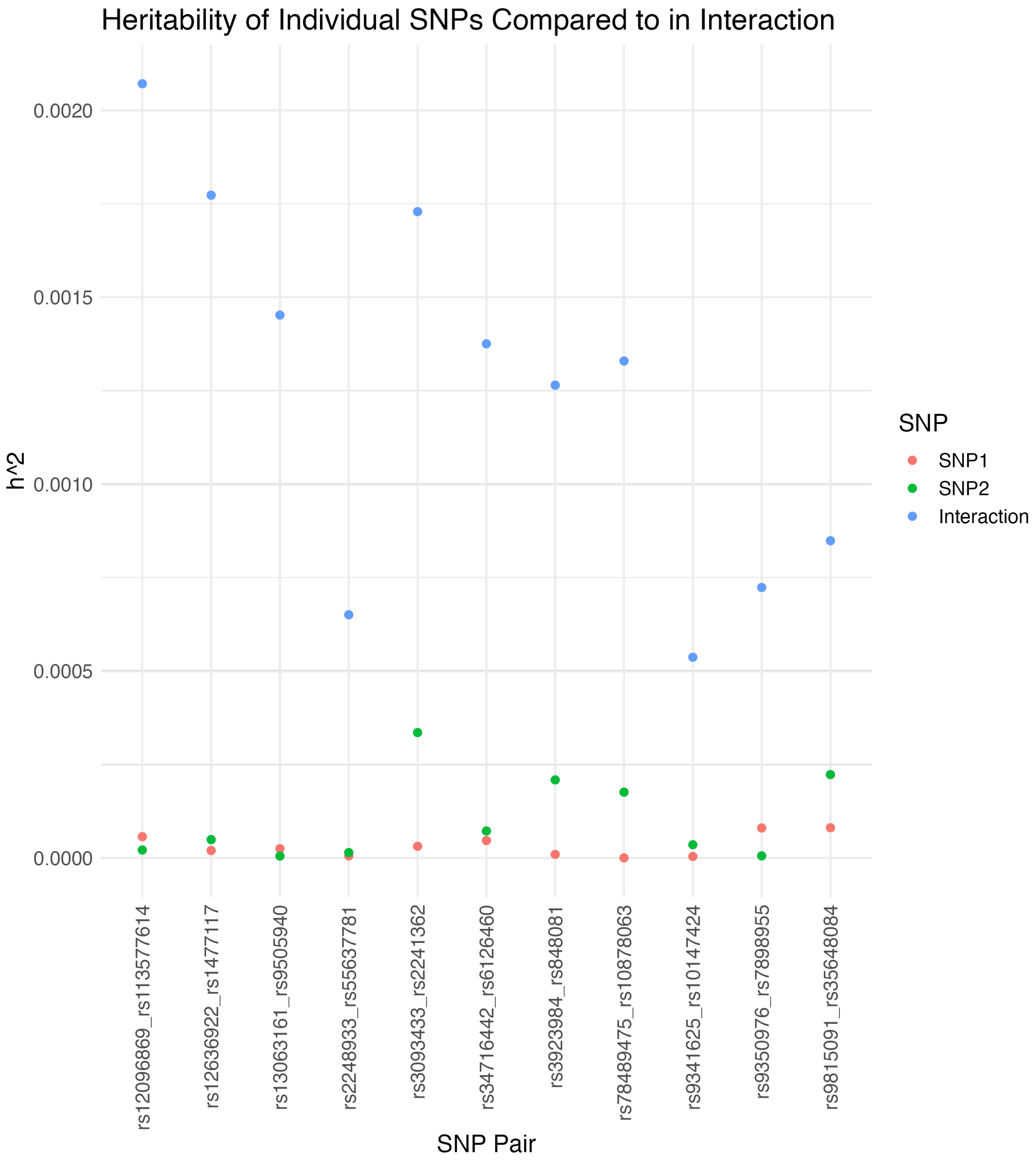


Figure 3.8: Difference between heritability explained of interaction compared to two SNPs in Alzheimer's Disease as detected by BOOST

3.3.4 Parkinson’s Disease

The BOOST search returned interactions that included now interactions with direct links to PD. There were two interactions that were characterized by a link involving unrelated neural activity, with TFG and $LATS2$ sharing a link to ’Lower cranial nerve dysfunction’ (HP:0410262) and $VCAM1$ and $MICAL3$ both linked with ’cardiac neuron differentiation’ (GO:0060945).

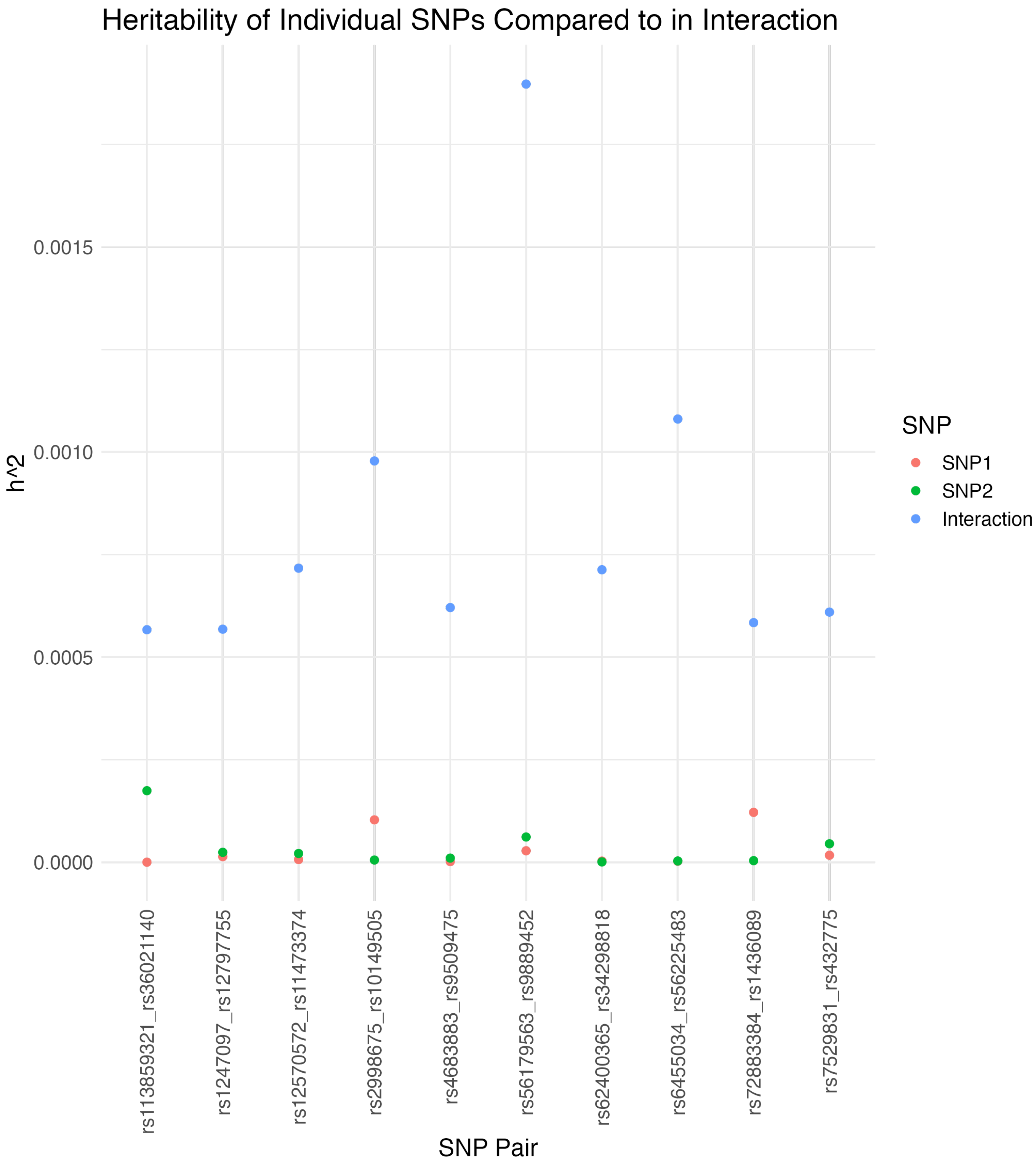


Figure 3.9: Difference between heritability explained of interaction compared to two SNPs in Parkinson's Disease as detected by BOOST

The MDR results returned one more relevant possible interaction, in $FSTL4$ and $SMAD6$. These shared the term ’negative regulation of brain-derived neurotrophic factor receptor signalling pathway’ (GO:0031549).

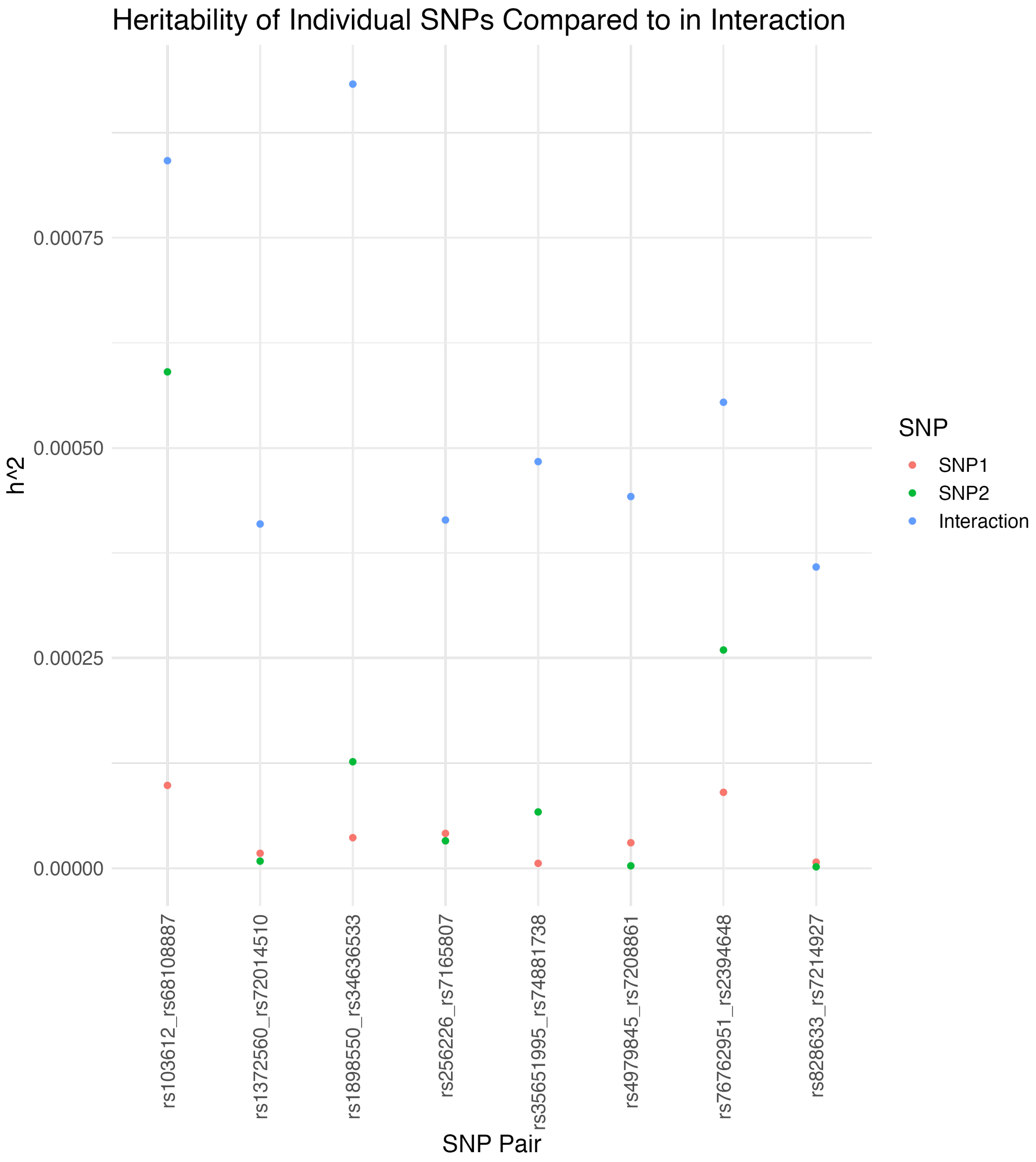


Figure 3.10: Difference between heritability explained of interaction compared to two SNPs in Parkinson's Disease as detected by MDR

3.4 Discussion

The simulation study showed the usefulness of combining strategies, with BOOST clearly better at finding pure epistasis and MDR better for impure interactions, as was seen in Chapter 2. At this scale, it was not possible to run MDR exhaustively, but it seems reasonable to conclude that this division would have held. It was interesting to see that the combined strategy, with data divided by monogenic significance, was almost as successful as BOOST in the pure tests and was consistently the best performer in the impure tests. Selecting variants within LD blocks does reduce the number of interactions recovered, greatly in the case of the single method approaches. This was not so much the case when both methods were applied, perhaps due to a greater number of loci being tested, since the data was divided into blocks after the division by significance.

A valuable observation from the simulation experiments was the number of interactions detected by BOOST when seen against significance threshold. Since BOOST is implemented within PLINK and is an extremely efficient, fast algorithm, this allowed a search strategy to be formulated but carrying it our in three phases. The first using the SNP array loci, then those that interacted at a significance of less than $5e-4$, local imputed loci were added to the data to potentially include variants that were in tighter LD with the causative SNP to be included. These were then tested by BOOST, with hose reaching the significance threshold being included in the set tested using MDR, after the data was reduced using candidate variants from LD blocks.

The results for AF were different between Chapters 2 and 3, which is perhaps unsurprising given the limited data that was used in the previous iteration. Observations for AF were of interest, as some of the terms found between genes are linked to AF in the literature. There is a growing body of evidence that inflammation plays a part in the aetiology of AF, so a link with cytotoxic T cells and IL-22 could play a role (Xiaoxu Zhou et al., 2020). Indeed, $CD4^{+}CD28^{null}$

T cells are known to play a role in the development of AF (Hammer et al., 2021) and IL-22 is found in elevated levels in the atria of patients with AF (Yongxin Wu et al., 2020). The role of D-lactate hydrogenase is a potential area for further research, given people with AF tend to have increased levels of atrial lactate that causes oxidative stress within the tissue (Xu et al., 2013). Included in the results were references to some neural activity, however since the lateral motor column and Brain-Derived Neurotrophic Factor are both removed from the heart tissue this is maybe a less useful insight. The findings from MDR were less obviously linked, in particular elevated fecal sodium levels, since hypertension is such a strong predictor of AF you would expect that sodium being lost from the body would be a benefit. This has been shown in a pharmacological test in rats (Linz et al., 2020). There is some research to link talin to AF, in that talin has been shown to be key to cardiomyocyte growth, as seen in $D.melanogaster$ (Bogatan et al., 2015).

The results were less interpretable for AD, although the protein-protein interaction in STRING-db confirms that an interaction is present. Checking both genes for expression in GTEx, $SLC35A1$ is highly expressed in the cerebellum and cerebellar hemisphere, whilst $TMEM5$ is poorly expressed in the brain, but most present in the spinal cord. However, given $TMEM5$ is thought to be involved with neural tube defects, there may an unidentified link (Vuillaumier-Barrot et al., 2012). Similarly for PD, interpretation of the results was limited, with a potential link to brain-derived neurotrophic factor possibly an avenue for further study given its centrality in brain development and overall health (Miranda et al., 2019).

None of the interactions tested were able to reach a *p-value* that was significant. This indicates the problem of statistical power when trying to demonstrate epistasis. Particularly in the case of AF where the $p-values$ were less than $1e-11$, many of those found were likely to be true negatives as a result of statistical chance. In this project, it was notable that the findings for AF had many more possible explanations and links to the disease than those for AD and PD, which could be because the greater number of samples provided more reliable results that had a better likelihood of being false negatives. It would also have been interesting to see some pathways represented across interacting pairs, but this was not present in the results.

Chapter Four

Searching for Epistasis in Rare Conditions

4.1 Introduction

Rare variants are defined as those with a MAF of less than 0.01 (Goswami et al., 2021). They are attractive targets for deleterious mutations, since their low incidence within the population indicates either a more recent mutation or one that is not readily spread through the population as a result of it hindering the individual's chances of passing it on. In the study of rare disease, it is often low prevalence, high penetrance mutations that cause these diseases (McCarthy et al., 2008). They are incorporated into this work from two perspectives, firstly deleterious rare variants will not always have full penetrance, that is a person can have a particular rare variant known to be associated with a condition but not have the trait. As a result, there is the potential for interactions between mutations giving rise to the disease in question. Secondly, rare variants have been identified as a part of the problem of missing heritability, and so the expansion of the study into rare variants expands the scope of this investigation into the drivers of disease (Girirajan, 2017).

In the case of rare diseases, different approaches must be taken to identify potential causative variants. Cohorts that are the size of GWAS are more difficult to assemble when a disease is less prevalent in society. In smaller cohorts, pedigree data is sometimes used by sequencing trios of mother, father, and proband. In this case, a Transmission Disequilibrium Test (TDT) can be used to leverage this (Ruiz-Narváez et al., 2004). GWAS is also generally focused on common variants, in part because they are available in large enough numbers to reach statistical significance, as well as SNP arrays performing poorly when detecting rare variants (Goswami et al., 2021; Mn et al., 2021).

Scans for candidate mutations can be performed by learning from previous work, as held in a variety of different databases and ontologies. However, integrating this can be difficult. A solution is held in PhenomeNET, a large knowledge graph, integrating cross species phenotypic data using ontologies. This is structured with multiple species' phenotypes mapped together, linked by their similarity. By using orthologous genes, disease prediction can be made from, for example, similar mouse phenotypes (Hoehndorf et al., 2011).

This has then been built upon with the variant prioritization algorithm, PhenomeNET Variant Predictor (PVP). This prioritizes variants from a single individual's genetic material held in Variant Call Format (VCF). It combines genotype information, with OMIM mode of inheritance, various pathogenicity scores and the data in PhenomeNET. A deep neural network is then used to collate all this information and generate a predictive score by which mutations are ranked. This allows the user to assess the mutations involved in giving rise to the phenotype (Boudellioua et al., 2019). Included within this software is another facet, OligoPVP. This tool uses the same background information as DeepPVP but also incorporates data from STRING-db to create a scoring system that integrates semantic similarity, pathogenicity, and known protein-protein interactions (Boudellioua et al., 2018).

4.1.1 Developmental Disorders and Hypothyroidism

Following the work carried out examining neurological conditions in Chapter 3, the Deciphering Developmental Disorders dataset was obtained (Firth et al., 2011). This data contains trios and whole exome data, with the child having been diagnosed with a developmental disorder. However, an interesting rare condition that is in this dataset in hypothyroidism, a disorder that has known links with brain development and psychiatric conditions (Uchida et al., 2021). Congenital hypothyroidism occurs in around one in two to four thousand newborn infants, varying depending on the population and screening procedures. Alongside brain development and Down's syndrome, hypothyroidism is also associated with many physical malformations (Rastogi et al., 2010). There are a number of known genetic factors that can be causative for hypothyroidism, and so it appears to be an interesting case study for variant prioritization (Stoupa et al., 2021).

4.2 Materials and Methods

4.2.1 Whole Exome Sequencing and Annotation

This work used two strategies. Firstly, a variant prioritization approach to uncover rare mutations that could explain the condition observed. This is required since these variants are unlikely to be present in many individuals, and so we lack the statistical power to use tests outlined in previous chapters. Secondly, a modified version of the pipeline used in Chapter 3 will be used for the common variants present (Figure 4.1).

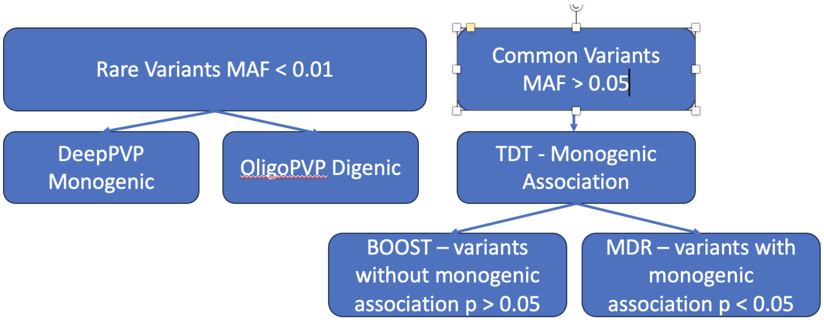

Rare Variants MAF < 0.01
Common Variants MAF > 0.05
DeepPVP Monogenic
OligoPVP Digenic
TDT - Monogenic Association
BOOST – variants without monogenic association p > 0.05
MDR – variants with monogenic association p < 0.05

Figure 4.1: Flowchart to show testing strategies implemented

The data for this work were acquired from the European Genome-Phenome Archive (EGA), using their in-house python downloader package, pyEGA. There were 102 samples with hypothyroidism, with two parents each. The format was as bam files, aligned to the genome build b37d5. This was first converted to fastq files using samtools, then realigned to GRCh38 using bwa's mem function. For variant calling, a few methods were trialled, but Octopus gave the most reliable results. Octopus uses a Bayesian model to build haplotypes based on the reference genome and aligned reads. A trio-based analysis is built in, so all three genomes are used to infer variants (Cooke et al., 2021). The default hard filters were used for *de novo* and regular variants, with any that weren't marked as 'PASS' removed.

Real Time Genomics tools (rtg) was used to annotate any *de novo* variants. VEP was used to annotate the variants including the –everything flag as well as a custom annotation from the gnomAD genomes dataset for MAFs. Variants were filtered so that only rare alleles with a MAF less than 0.01 in both the gnomAD exome and genomes data were included. PVP requires variants to be mapped to the GRCh37 genome build, so a liftover was carried out using picard tools. PVP was then used to prioritize variants for the Hypothyroidism HPO term 'HP:0000821'.

Variants were further annotated with GTEx data to demonstrate gene expression in the thyroid. Pathogenicity was defined using a combination of a number of annotated sources from VEP – the consequence as defined by the Sequence Ontology (SO) Impact groups of 'High',

’Medium’, ’Low’ and ’Modifier’ categories as well as the PolyPhen and SIFT definitions. Using these, a new pathogenicity categorization was created using all of these. ’Very High’ is defined as being in the most damaging category for all definitions. ’High’ is defined as any in the top two categories, but not all. ’Low’ is a variant that is classified as tolerated, benign and low or modifier impact. ’Medium’ is those that don’t fall into the above categories. This is then combined with a CADD score, a quantitative pathogenicity rating, to categorize variants as pathogenic. Variants were retained in this way if they were classed as ’Very High’ and CADD score greater than 20, ’High’ and greater than the category mean CADD Score or ’Medium’ and CADD score of greater than 30. Each sample was then examined for *de novo* variants, digenic variants with a link to hypothyroidism either within a panel by Genomics England or in the HPO database. Any potential interactions were checked to see if the parent’s also have both mutations, to demonstrate if this could be the source of the condition.

4.2.2 Evaluation of Common Variants

A search for interactions was also carried out for common variants, using the pipeline set out in Chapter 3. Firstly, the VCF files for all of the trios were merged. From this, BEAGLE was used to imputed loci that were not included in the data, using the GRCh38 mappings and 1000 Genomes panel. The data was converted into PLINK format. Variants were excluded if they had a missing rate of greater than 0.05, a MAF of less than 0.05 or a *p-value* for departure from HWE of less than $5e-8$. The data was pruned to keep only variants in linkage equilibrium using PLINK. The software KING was then used to confirm that none of the probands were related (Manichaikul et al., 2010). PCA was then carried out to use the PCs as covariates using flashpca (Abraham et al., 2017).

4.3 Results

4.3.1 Variant Prioritization

To find pathogenic variants, a combination of PolyPhen, SIFT and Impact was used, with CADD also incorporated after classification. In Figure 4.2 there is generally the expected trend with 'Very High' and 'High' risk variants having the greatest distribution of CADD scores, with this reducing through the classifications. This also shows why CADD is used as well, since there are potentially low risk loci even in the 'Very High' risk category.

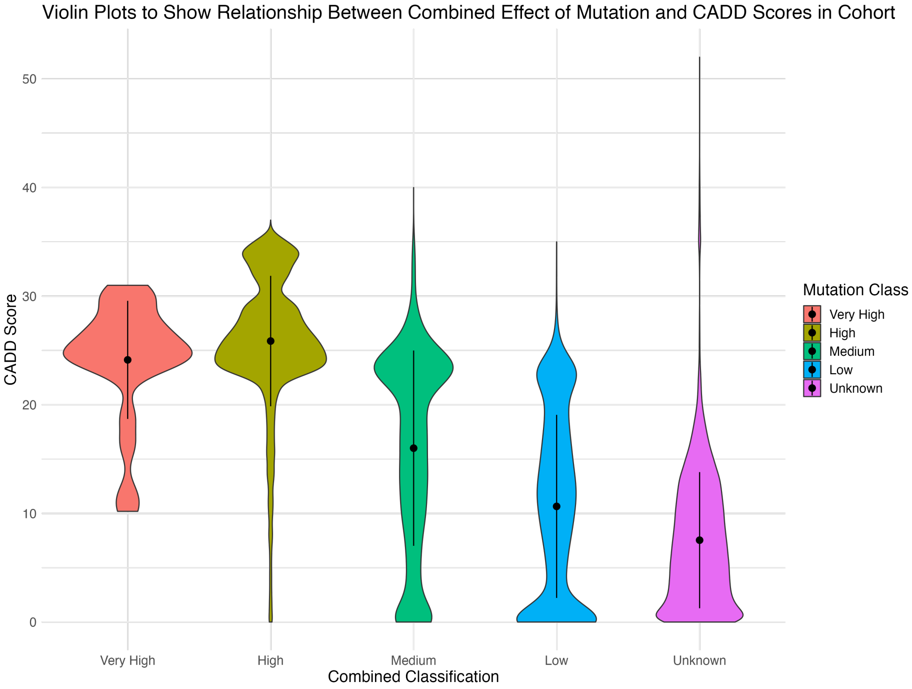


Figure 4.2: Violin plot to show the distribution of CADD scores present at each of the combined pathogenicity categories

In the cohort, four individuals had deleterious *de novo* mutations in relevant genes, as shown in Table 4.1. Two of these genes were in the Genomics England panel for hypothyroidism, while the other two were linked to hypothyroidism in the HPO database. In terms of monogenic mutations, five had a mutation in the *TG* gene, two in *DUOX2*, two in *SLC26A4* and one in *TPO*. All of these are known genes linked to hypothyroidism, however these were not *de novo* mutations so cannot be fully accountable. It should also be noted that one individual, whose mother had hypothyroidism, also shared a homozygous, deleterious genotype in TPO, which is noted for its 'Autosomal recessive inheritance' in HPO.

Table 4.1: *de novo* mutations that are linked to hypothyroidism

Gene	Evidence
PAX8	Genomics England Panel
PRKAR1A	Genomics England Panel
KAT6B	HPO db link
CACNA1C	HPO db link

When assessing the genomes for digenic interactions, there were six candidates that were pathogenic and contained a gene that was explicitly linked to hypothyroidism, as seen in Table 4.2. Of these, only one was not also present in one of the parents between *PDIA4* and *TG*. These are predicted to be interacting in STRING-db at high confidence (0.808), due to co-expression of homologs, biochemical interactions of homologs in other organisms and co-mentions in PubMed abstracts for these genes and also homologs in other organisms.

Table 4.2: digenic mutations that are linked to hypothyroidism

Gene	Evidence	Parents
PDIA4_TG	Genomics England Panel (TG)	No
APOB_PTCH1	HPO db link (PTCH1)	Yes
DNAH1_DNAH8	HPO db link (DNAH1)	Yes
GDNF_SALL1	HPO db link (SALL1)	Yes
MTOR_TSC2	HPO db link (TSC2)	Yes
IFIH1_CHUK	HPO db link (IFIH1)	Yes

4.3.2 Implementation of Epistasis Detection Approaches

Experimentation on common variants through a GWAS pipeline with logistic regression yielded a non-significant association to rs10184375 with a *p-value* of 0.0003685. Comparatively, TDT performed well, achieving a significant result at the SNP rs1519139, with a *p-value* of $1.28e-08$. This variant is located within the gene *LOC105372041*, which is a lncRNA with little known about it. However, GTEx does report some samples expressing the gene in the thyroid, albeit not in most.

The use of epistasis detection tools for common variants in this cohort yielded interactions that did not achieve statistical significance. The most significant result from BOOST (*p-value* $4.66e-11$) was between rs111501662, an intergenic variant and rs56084170, which was in linkage ($R^2 = 0.629$) *AP4S1* and *HEATR5A*, neither of which appear to have any clear link to thyroid function. The second most significant result was between rs1496554 in *MORN1* and rs1609475 in *DEAF1*. Testing with MDR revealed a possible interaction between rs4874118, which is in LD with *ZC3H3* ($R^2 = 1$) that is highly expressed in the thyroid in GTEx and rs1478612311, an intronic mutation in *PLAAT3*, which is demonstrably expressed in the thyroid, but below the levels of in other organs. The second-strongest interaction was between rs11234871, an intron variant of *PRSS23* and rs11078306, which is intronic to *TVP23C*.

4.4 Discussion

In this chapter, a different approach to uncovering missing heritability has been taken. By focussing on rare variants, it was possible to find *de novo* mutations in four children, that would not have been identified in a GWAS. Two of these genes were a part of the Genomics England hypothyroidism panel. This represents a gold standard for identification of genetic mutations that are involved in a given disease. This resource takes research from a diverse selection of database resources, such as OrphaNet, OMIM and ClinVar, curated evidence from within the NHS, crowdsourcing of expert knowledge, publications and continuous feedback for improvements. This keeps the resource up to date and gives a level of confidence in the resource (A. R. Martin et al., 2019).

When a germline mutation is not found, then a parent must have any other allele present in the child. There was one parent, a mother, who had hypothyroidism and, as such, a shared genotype that the father didn't have could potentially be the cause. No single heterozygous mutation was found that could be explanatory, so the search was expanded to heterozygous mutations, of which a homozygous alternative allele in the gen *TPO*, that is included in the Genomics England panel was found.

A similar rationale is behind the search for digenic mutations - in the absence of a *de novo* mutation, it seems apparent that there must be a combination of mutations not present in the parents. The digenic search in PVP was, thus, of great use in identifying possible interactions. Of the six digenic interactions found, only one was not apparent in the parents and so is the only one that could potentially provide a 'solution'. The putative link between *TG* and *PDIA4* appears in the STRING-db with a high confidence score (0.808) based on homologs being coexpressed and interacting in other organisms, as well as being co-mentioned in PubMed abstracts regarding *H. sapiens.* Although none of these papers is discussing thyroid function, the protein thyroglobulin produced by *TG* is undoubtably closely related to hypothyroidism, acting as a substrate in the production of the thyroid hormones T3 and T4 (Citterio et al., 2019). The gene *PDIA4* is also noted as being a part of the 'Thyroid hormone synthesis' pathway in KEGG.

This approach to finding potential causative variants does face some serious limitations though. On the one hand, it is not possible to test these mutations statistically, since they are only present in one individual in the cohort. As such any possible interpretation is difficult to validate, aside from that documented in previous studies. On the other, the disorder is rare and so finding other cohorts to validate against is also more difficult. The UK biobank does have some individuals with congenital hypothyroidism though, so this would be a potential future cohort to explore.

The part of the study focussed on common variants was fairly severely underpowered. However, this does make it interesting that the TDT test found a significant variant, those this does appear to be a fairly normal sample size when using this particular test (Ruiz-Narváez et al., 2004). It is possible though, given PLINK's implementation uses a χ^2 test, that uncorrected for factors were biasing the result. Of the interactions found, it was perhaps interesting that one of the genes involved was *DEAF1*, given that it is highly involved in problems of intellectual development and microcephaly, given that these are traits in high abundance in the cases but not the parents that were used as controls (Faqeih et al., 2014).

Chapter Five

Conclusions and Future Work

The idea behind epistasis heralds back to the earliest days of genetics, of Bateson, Punnett and Saunders growing a wide variety of crops and rearing animals in different breeding experiments near Cambridge. The observation was made that different genetic factors could suppress the expression of others, and thus the term was born. Conceptually it has passed through some of the most eminent scientists involved in the founding of population genetics, in Fisher, Wright and Haldane, people whose work is still a mainstay of many genetic approaches carried out today.

In the last century, genetics has become a driving force of biology. As foreseen by Wright and Fisher, it can now be clearly observed that organisms are made up of complex biochemical processes and metabolic pathways with yet unknown relationships and drivers. Understanding how these systems go wrong in diseases, identifying the diverse issues found in complex diseases, and translating these findings into personalized medicinal treatments is becoming a reality.

The completion of the Human Genome Project finalized a reference for genetic diversity, and further projects like HapMap and 1000 Genomes were able to catalogue the diversity of mutations in the genome. With the growth of technology in SNP arrays and sequencing machines, large cohort association studies were possible, as outlined about a decade earlier by Risch. As the equipment has become cheaper and funding moved into projects like the Wellcome Trust Case-Control Consortium or UK biobank, there is now a glut of data available for common diseases. Problems have arisen along the way though, there have been some serious problems with false positive results being reported from GWAS, but also the enormous number of small effect size associations has raised concerns. Identified as the 'problem of missing heritability', serious questions have been asked about why such a small proportion of phenotypic variance has been explained compared to estimates produced in twin studies before. A number of possible answers have been put forth, such as the lack of inclusion of rare variants and also, the presence of genetic interactions that are not being looked for. This thesis has been an investigation into the possibility of overcoming this shortfall in heritability and aimed to put forward an efficient set of methods that can be used to assess genetic interactions and their role in complex disease.

In Chapter 2, a review was carried out into the various methods that had been devised for the task of detecting interactions in case-control GWAS datasets. These were broadly categorized into three types; statistical; swarm intelligence; and data mining, as well as some that carried out exhaustive searches and others that attempted to optimize the search. A simulation study was created to test their detection abilities based on either pure epistasis, that in which no main effects are seen and impure epistasis, in which the individual variants have some statistical association with the trait. This was carried out for both two and three locus models. Three approaches stood out from these tests, PLINK's implementation of the BOOST algorithm, the MDR and MPI3SNP. The problems each of these worked best with were different, so BOOST found pure interactions most effectively, MDR the impure two locus interactions and MPI3SNP the third-order interactions. Assessments were also made for computational performance, with BOOST and MPI3SNP both being fast and memory efficient and MDR, coded in Java rather than C++ being somewhat slower. These approaches were applied to a demonstrative 'toy' dataset based on the AF data in UK biobank, but greatly pared back. It raised some issues that were not seen in the simulation study, such as how the tools handled missing data.

These findings informed the thrust of Chapter 3, the aim of which was to apply the best methods to biobank scale data for GWAS interaction detection. For this, MDR and BOOST were selected, and the experimentation limited to pairwise interactions. Three strategies were assessed. Firstly, an exhaustive approach, testing all pairs. Secondly, a division of the data by strength of association to the trait, BOOST applied to those with little association and MDR to those with a single-variant main effects. Thirdly, the data was divided into LD-based haplotype blocks, with the aim being to reduce the feature space as could be required at a larger scale. Finally, a combined approach from the final two was used, splitting by LD and tailoring the data to the method by association. A simulation study was carried out using data from chromosome 22 in the UK biobank, across 100,000 samples split evenly between cases and controls. The first observation was that MDR could not be used exhaustively at this scale across over 9,000 variants, so its contributions were at a smaller feature space in the second and third strategies. As was observed in Chapter 2, BOOST performed best with the pure interactions and MDR with the impure interactions. Interestingly, the combined method performed very well and as such a version of this was deemed to be the most applicable to all autosomal chromosomes. Another observation was that at a greatly reduced significance threshold, BOOST was able to detect almost all of the interactions captured by the other approaches. Since BOOST is a very fast and highly parallelizable program, this offered a possibility of using it to filter the dataset early in the process.

With a pipeline established, three diseases were selected to test it on. These were Atrial Fibrillation, Alzheimer's Disease and Parkinson's Disease. The detection approach was in three phases to get the most from the data. First, the genotyped SNP array data was used to indicate the regions of interest. Then variants from the wider imputed data in LD with those that had been implicated in potential interactions were extracted and tested again with BOOST. Finally, the data was split into haplotype blocks and those with a monogenic main effect were tested using MDR.

Once potential interactions were found, a novel workflow was implemented. Since each variant in GWAS represents an area of the genome, all genes that the variant was in tight linkage with were annotated to the variant. Then each combination of genes was scanned with the R package gProfileR to find shared annotations from a selection of ontologies. Next, STRING-db was searched for any evidence of a protein-protein interaction. Finally, the interaction was defined by assessing each possible genotype from combinations of the two loci. With the interaction defined, key statistics could then be calculated, in particular the odds ratio and variance explained, the measure of heritability.

One of the great problems when searching for interactions is surpassing the burden of proof to reject the null hypothesis within the frequentist paradigm. It is completely reasonable to be concerned about the problem of multiple testing. Indeed, a prevalence of false positive results hampered the reputation of GWAS in its early days. While the Bonferroni adjustment is seen as overly stringent, the popular Benjamini-Hochberg correction still applies the same threshold but for the most significant result only, with a step-down in proceeding values. As such, at least one result must surpass the stringent adjustment set out by Bonferroni. This presents a problem when testing interactions because so many tests are carried out. Whilst none of the experiments in AF, AD or PD achieved significant results, this was in a large part due to this stringent threshold and so true positive results were likely mixed with negative results. Perhaps a reduced ’suggestive’ threshold could be implemented as well, but in a field so often beset with reproducibility issues, I feel it would be unwise to compromise on an established mode of correction.

There were some very positive signs though, with the heritability calculations showing that, from BOOST especially, a great deal more heritability was explained when taken as an interacting pair. This held true for MDR in general, but less so. There were two interactions that were seen in STRING-db as protein-protein interactions. As well as possible avenues of research highlighted by shared terms found in the gene set enrichment stage. There were more interpretable results for AF, and this is perhaps due to the greatly increased number of samples available for that condition.

Finally, in Chapter 4, we had a practical case using the tool PVP, that aims to prioritize variants with the aim of locating causative loci. This focused on the rare disease of congenital hypothyroidism. As a result, we could also find rare variants that would be missed in a GWAS. Sequencing was carried out and four $denovo$ variants discovered in hypothyroidism linked genes. A number of possible interactions were also found, but only one was not also present in one of the parents. This was a known interaction with a high confidence in STRING-db and both proteins were involved in the KEGG pathway for synthesizing thyroid hormones. Only 102 cases were available, so the power was very low for common variants. The epistasis detection pipeline was applied, but with uninterpretable results.

Although this thesis has not provided any significant results for diseases, the aims of the study that were set out have been achieved and a robust methodology for epistasis detection laid out. This is accompanied by downstream analysis that takes into account the LD surrounding the locus and draws from many annotation sources to better understand how the interaction relates to the trait of interest. This pipeline uses tools that have been shown to perform best amongst peers in simulation studies, and when applied to real data were able to find interactions that explained more heritability than the variants alone.

5.1 Limitations

The pipeline developed through Chapters has some flaws that will need addressing. Most traits are affected by various covariates and these should be taken into account when searching for interactions, or risk presenting results that are actually a result of some segmentation of the data or an unrelated but correlated trait. This could perhaps be achieved by testing genetic loci for those significantly correlated with any expected confounding factors and either removing them or taking note of them for if they are present in the resulting interactions. Another approach would involve using a quantitative epistasis detection algorithm. Binary traits can then be translated into quantitative traits by regressing covariates against them and using the residuals as the new dependent variable.

The lack of incorporation of covariates into these tests also relates to an important weakness of this study, the lack of diversity possible within the cohort of study. The problem is twofold, first that as has been observed in other studies, many genetic studies are focussed on people of European ancestry, leading to less generalizable findings and also that potential associations can be missed (Popejoy et al., 2016). Secondly, the volunteers gave their time, information and biological samples to UK biobank. It is best to get the most from their data to advance medical research as they intended.

Chapter 4 was inherently restricted by the reduced number of samples, as is the case for rare diseases. As a result, the predominant method for identifying potential causative mutations and interactions was using variant prioritization. This cannot be tested statistically in a cohort, unless the disorder is mono- or digenic. One method would be to approach the problem as a Linkage study, leveraging pedigree data in order to generate greater statistical power. This would require access to data from a family group though. Another method would be to attempt to validate the mutations using another dataset. UK biobank contains 111 individuals who are reported to have congenital hypothyroidism, this could be a good group to explore for similar mutations.

5.2 Future Work

An area that was neglected in this study is quantitative traits, this could also help overcome the limitations caused by not including covariates in the analysis. The statistical options are slightly different from in binary outcomes, so for example, a contingency table, based approach is no longer possible due to the lack of categorization. However, there are still MDR approaches, one of which is model-based MDR (MB-MDR), which is a well maintained and optimized program that implements an MDR based on a genomic model (CATTAERT et al., 2011). It also has a built-in multiple testing correction called gammaMAXT, based on permutation testing, that would be worth investigating as a possible way to increase the sensitivity of the pipeline created in this thesis (Lishout et al., 2015).

UK biobank has recently released Whole Genome Sequencing data for around 200,000 individuals. This could be utilized to incorporate rare and common variants, such as using this data to attempt to validate mutations found in the hypothyroidism cohort. Since rare variants are more likely to be highly pathogenic than common variants, they could be used to stratify the data and look for interactions, deleterious or otherwise, with the common variants. Also, by using data with more coverage, it may be possible to locate the causative variant in interactions between common variants. This would likely increase the predictive power and achieve more significant results.

In order to assist users to apply this epistasis detection pipeline to their data, it would be useful to create a Galaxy pipeline that allows users to carry out these tests without having to know everything about each of the programs involved or deal with intermediate files. The downstream processing could also be collected into an R package, allowing users to test potential interactions by identifying associated genes and finding potential explanatory factors using ontologies and STRING-db, amongst others.

Experimental validation of results would be another area of interest. Most methods for doing this involve testing for protein-protein interactions. One method is co-immunoprecipitation (Lee, 2007). This involves fixing an antibody for a protein of interest to a gel bead, this is incubated with matter from cell lysis. Proteins that have bound to the protein can then be identified. Pull down assays are similar, except instead of an antibody, the protein itself is used as 'bait' (Louche et al., 2017). Another alternative is far western blotting (Y. Wu et al., 2007). In this, proteins from cell lysate are separated using standard western blotting. Then a tagged bait protein is added. If it binds to the protein being assayed, then this can be detected using antibodies that associate with the tags. An alternative is to look at the transcriptome. Within the tissue of interest, evidence for an interaction could be co-expression (Paci et al., 2021). This data can be collected using a SNP microarray.

Appendix One

Appendix

For additional online appendix, please see https://github.com/domruss87/thesis_appendix/tree/main

A.1 Chapter 2

Table A.1: Available epistasis detection methods up to 2020

Tool	DOI	Available
ABCDE	https://doi.org/10.1515/sagmb-2012-0074	-

Continued on next page

Tool	DOI	Available
AFT UM-MDR	https://doi.org/10.5808/GI.2016.14.4.166	-
AntEpiSeeker	https://doi.org/10.1038/npre.2012.6994.1	http://nce.ads.uga.edu/~romdhane/AntEpiSeeker/index.html
AntMiner	https://doi.org/10.1007/s13258-012-0003-2	https://sourceforge.net/projects/antminer/files/
BEAM	https://doi.org/10.1038/ng2110	https://sites.fas.harvard.edu/~junliu/BEAM/
Bhit	https://doi.org/10.1186/s12864-015-2217-6	http://digbio.missouri.edu/BHIT/
BiForce	https://doi.org/10.1093/nar/gks550	-
BridGE	https://doi.org/10.1038/s41467-019-12131-7	http://csbio.cs.umn.edu/bridge
CAPE	https://doi.org/10.1371/journal.pcbi.1003270	https://github.com/marta-vidalgarcia/CAPE
CASSI	https://doi.org/10.1371/journal.pgen.1002625	https://www.staff.ncl.ac.uk/richard.howey/cassi/installation.html
CINOEDV	https://doi.org/10.1186/s12859-016-1076-8	https://cran.r-project.org/src/contrib/Archive/CINOEDV/

Continued on next page

Tool	DOI	Available
COE	https://doi.org/10.1089/cmb.2009.0155	http://www.csbio.unc.edu/epistasis/client-coe2.php
Cox MDR	https://doi.org/10.1093/bioinformatics/bts415	-
CSE	https://doi.org/10.1038/hdy.2014.4	-
DECMDR	https://doi.org/10.1093/bioinformatics/btx163	https://drive.google.com/file/d/0B93CxNBXL-MyR1NBZEhMNGRYcUE/view
Deep Mixed Model	https://doi.org/10.1186/s12859-019-3300-9	-
DPEH	https://doi.org/10.1038/s41598-018-24588-5	-
DualWMDR	https://doi.org/10.1002/humu.23951	http://mlda.swu.edu.cn/codes.php?name=DualWMDR
eCEO	https://doi.org/10.1093/bioinformatics/btr091	-
Eigenepistasis	https://doi.org/10.1186/s12859-017-1488-0	https://github.com/vstanislas/GGEE
Encore	https://doi.org/10.1002/gepi.21739	http://insilico.utulsa.edu/index.php/encore/
epiACO	https://doi.org/10.1186/s13040-017-0143-7	https://sourceforge.net/projects/epiaco1/files/epiACO.rar/download

Continued on next page

Tool	DOI	Available
Epiblaster	https://doi.org/10.1038/ejhg.2010.196	https://www.mybiosoftware.com/epiblaster-1-0-two-locus-epistasis-detection-s html
EpiForest	https://doi.org/10.1186/1471-2105-10-S1-S65	-
EpiGTBN	https://doi.org/10.1186/s12859-019-3022-z	http://122.205.95.139/Epi-GTBN/
epiGWAS	https://doi.org/10.1371/journal.pone.0242927	https://cran.r-project.org/web/packages/epiGWAS/index.html
epiMODE	https://doi.org/10.1371/journal.pgen.1000464	-
epiNEM	https://doi.org/10.1371/journal.pcbi.1005496	https://github.com/cbg-ethz/epiNEM
epistasis	arXiv:1710.00894v2	https://cran.r-project.org/web/packages/epistasis/epistasis.pdf
FAACOSE	https://doi.org/10.1155/2017/5024867	-
FAM MDR	https://doi.org/10.1371/journal.pone.0010304	-
fastChi	https://doi.org/10.1142/9789812836939_0050	-

Continued on next page

Tool	DOI	Available
FastLMM	https://doi.org/10.1038/nmeth.1681	https://fastlmm.github.io/FaST-LMM/
FCME	https://doi.org/10.1109/TFUZZ.2019.2914629	https://gitlab.com/yudalinemail/fcmemdr
FDHE-IW	https://doi.org/10.3390/genes9090435	-
FSMDR	https://doi.org/10.1016/j.artmed.2019.101768	-
GAIN	https://doi.org/10.1371/journal.pgen.1000432	https://github.com/insilico
GeneGeneInteR	https://doi.org/10.18637/jss.v095.i12	https://bioconductor.org/packages/release/bioc/html/GeneGeneInteR.html
GenEpi	https://doi.org/10.1186/s12859-020-3368-2	https://github.com/Chester75321/GenEpi
GENIE	https://doi.org/10.1186/1756-0500-4-158	-
GENN	https://doi.org/10.1002/gepi.20307	-
Glide	https://doi.org/10.1159/000341885	https://github.com/BorgwardtLab/Epistasis-GLIDE
GMDR	https://doi.org/10.2174/138920291766616051310261 2	http://ibi.zju.edu.cn/software/GMDR/download.html

Continued on next page

Tool	DOI	Available
GWIS	https://doi.org/10.1186/1471-2164-14-S3-S10	-
HS-MMGKG	https://doi.org/10.2174/1574893614666190409110843	-
IACO	https://doi.org/10.1007/978-3-319-42297-8_3	-
iLOCi	https://doi.org/10.1186/1471-2164-13-S7-S2	https://www.mybiosoftware.com/iloci-snp-interaction-prioritization-technique-html
IndOR	https://doi.org/10.1002/sim.5364	http://emily.perso.math.cnrs.fr/IndOR/IndOR/IndOR.html
Interaction Trees	https://doi.org/10.1109/ICMLA.2012.114	-
IOBLPSO	https://doi.org/10.1155/2015/524821	-
IPSO	https://doi.org/10.1371/journal.pone.0037018	-
JS-MA	https://doi.org/10.3389/fgene.2020.507038	-
KCCU	https://doi.org/10.1186/1471-2156-13-83	-

Continued on next page

Tool	DOI	Available
KNN-MDR	https://doi.org/10.1038/nature05911	-
lampLINK	https://doi.org/10.1093/bioinformatics/btw418	http://a-terada.github.io/lamplink/
LINDEN	https://doi.org/10.1093/nar/gkx505	http://compbio.case.edu/omics/software/linden/
log-linear MDR	https://doi.org/10.1093/bioinformatics/btm396	-
MACOED	https://doi.org/10.1093/bioinformatics/btu702	http://www.csbio.sjtu.edu.cn/bioinf/MACOED/
MapReduce	https://doi.org/10.1080/00207160.2014.1000882	-
MBMDR PC	https://doi.org/10.1089/sysm.2019.0003	http://bio3.giga.ulg.ac.be/index.php/software/mb-mdr/
MBS	https://pubmed.ncbi.nlm.nih.gov/21346997/	-
MECPM	https://doi.org/10.1093/bioinformatics/btp435	https://www.cbil.ece.vt.edu/ResearchOngoingSNP.htm
MegaSNPhunter	https://doi.org/10.1186/1471-2105-10-13	https://gaow.github.io/genetic-analysis-software/m/megasnphunter/

Continued on next page

Tool	DOI	Available
MERF	https://doi.org/10.1038/ncomms8432	https://github.com/PMBio/limix
model based MDR	https://doi.org/10.1111/j.1469-1809.2010.00604.x	-
MODEMDR	https://doi.org/10.1038/s41598-017-16210-x.	-
MP-HS-DHSI	https://doi.org/10.1093/bioinformatics/btaa215	https://github.com/shouhengtuo/MP-HS-DHSI
MPI3SNP	https://doi.org/10.1177/1094342019852128	https://github.com/UDC-GAC/mpi3snp
MultiSuRF	https://doi.org/10.1007/978-3-642-37189-9_1	https://epistasislab.github.io/ReBATE/
npdr	https://doi.org/10.3389/fgene.2020.00784	https://github.com/insilico/npdr
OR based MDR	https://doi.org/10.1093/bioinformatics/btl557	-
Parallel MDR	https://doi.org/10.1093/bioinformatics/btl347	-
PGMDR	https://doi.org/10.1016/j.ajhg.2008.09.001	-
PIAM	https://doi.org/10.1371/journal.pgen.1001338	-

Continued on next page

Tool	DOI	Available
PLINK	https://doi.org/10.1086/519795	https://www.cog-genomics.org/plink2/
PRLR	https://doi.org/10.1093/biostatistics/kxz011	https://github.com/kingqwert/R/tree/master/PRLR
Random Jungle	https://doi.org/10.1093/bioinformatics/btq257	-
Random Survival Forests	https://doi.org/10.1214/08-AOAS169	https://cran.r-project.org/web/packages/randomForestSRC/index.html
ranger	https://doi.org/10.18637/jss.v077.i01	https://www.jstatsoft.org/article/view/v077i01
ReliefF	https://doi.org/10.1007/3-540-57868-4_57	https://epistasislab.github.io/ReBATE/
REMMA (now GMAT)	https://doi.org/10.1093/bioinformatics/bty017	https://github.com/chaoning/GMAT
Robust MDR	https://doi.org/10.1111/j.1469-1809.2010.00624.x	-
Screen and Clean	https://doi.org/10.1002/gepi.20459	-
SHEIB-AGM	https://doi.org/10.1109/ACCESS.2020.2969465	-
Shesis	https://doi.org/10.1038/cr.2010.68	-

Continued on next page

Tool	DOI	Available
ShesisPLUS	https://doi.org/10.1038/srep24095	-
singleMI	https://doi.org/10.1007/s10586-017-0938-9	https://github.com/sleeepyjack/singlemi
SIXPAC	https://doi.org/10.1101/gr.137885.112	http://www.cs.columbia.edu/~snehitp/sixpac/
SMMB	https://doi.org/10.1093/bioinformatics/bty154	https://uncloud.univ-nantes.fr/index.php/s/bhFskQlsPmb5rXp
SNPharvester	https://doi.org/10.1093/bioinformatics/btn652	http://bioinformatics.ust.hk/SNPHarvester.html
SNPInterForest	https://doi.org/10.1186/1471-2105-12-469	https://gwas.biosciencedbc.jp/SNPInterForest/index.html
SNPrank	https://doi.org/10.1093/bioinformatics/btq638	https://github.com/insilico
SNPRuler	https://doi.org/10.1093/bioinformatics/btp622	https://www.mybiosoftware.com/snpruler-predictive-rule-inference-epistatic-in html
SNPTEST	https://doi.org/10.1038/nrg2796	https://mathgen.stats.ox.ac.uk/genetics_software/snptest/snptest.html#interactions

Continued on next page

Tool	DOI	Available
stability SCAD	https://doi.org/10.1186/1471-2105-15-62	-
SuRF	https://doi.org/10.1186/1756-0381-2-5	https://epistasislab.github.io/ReBATE/
Surv MDR	https://doi.org/10.1007/s00439-010-0905-5	-
TEAM	https://doi.org/10.1093/bioinformatics/btq186	http://www.csbio.unc.edu/epistasis/client-team2.php
TEPAA	https://doi.org/10.1089/cmb.2014.0163	http://genetics.cs.ucla.edu/tepaa/
TS-SIS	https://doi.org/10.1214/14-aoas771	-
Ttree	https://doi.org/10.1371/journal.pone.0093379	https://github.com/0asa/TTree-source
TwoFC	https://doi.org/10.1007/s10528-014-9656-7	-
UGMDR	https://doi.org/10.1038/hdy.2014.94	-
WISH-R	https://doi.org/10.1186/s12859-018-2291-2	https://github.com/QSG-group/wish
wtest	https://doi.org/10.1186/s12920-019-0638-9	https://CRAN.R-project.org/package=wtest

Continued on next page

Tool	DOI	Available
AFSBN	https://doi.org/10.1109/TCBB.2019.2949780	-
FAACOSE	https://doi.org/10.1155/2017/5024867	-
ETSACO	https://doi.org/10.1109/TETCI.2017.2699228	-
IEACO	https://doi.org/10.3390/genes10020114	-
HiSeeker	https://doi.org/10.3390/genes8060153	-

Continued on next page

References

Abecasis, Gonçalo R. et al. (Nov. 2005). "Handling Marker-Marker Linkage Disequilibrium: Pedigree Analysis with Clustered Markers". In: *American Journal of Human Genetics* 77.5, pp. 754–767.

Abo Alchamlat, Sinan et al. (Nov. 2018). "Aggregation of experts: an application in the field of "interactomics" (detection of interactions on the basis of genomic data)". In: *BMC Bioinformatics* 19.1, p. 445. DOI: 10.1186/s12859-018-2447-0.

Abou Ziki, Maen D. et al. (2021). "Epistatic interaction of PDE4DIP and DES mutations in familial atrial fibrillation with slow conduction". en. In: *Human Mutation* 42.10. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/humu.24265, pp. 1279–1293. DOI: 10.1002/humu.24265.

Abraham, Gad et al. (Sept. 2017). "FlashPCA2: principal component analysis of Biobank-scale genotype datasets". In: *Bioinformatics* 33.17, pp. 2776–2778. DOI: 10.1093/bioinformatics/btx299.

Adzhubei, Ivan et al. (Jan. 2013). "Predicting Functional Effect of Human Missense Mutations Using PolyPhen-2". In: *Current protocols in human genetics / editorial board, Jonathan L. Haines ... [et al.]* 0 7, Unit7.20. DOI: 10.1002/0471142905.hg0720s76.

Aguet, François et al. (Oct. 2017). "Genetic effects on gene expression across human tissues". en. In: *Nature* 550.7675. Number: 7675 Publisher: Nature Publishing Group, pp. 204–213. DOI: 10.1038/nature24277.

Allen, Naomi et al. (Sept. 2012). “UK Biobank: Current status and what it means for epidemiology”. en. In: *Health Policy and Technology* 1.3, pp. 123–126. DOI: 10.1016/j.hlpt.2012.07.003.

Altshuler, David et al. (Oct. 2005). “A haplotype map of the human genome”. en. In: *Nature* 437.7063. Number: 7063 Publisher: Nature Publishing Group, pp. 1299–1320. DOI: 10.1038/nature04226.

Amberger, Joanna S. et al. (June 2017). “Searching Online Mendelian Inheritance in Man (OMIM): A Knowledgebase of Human Genes and Genetic Phenotypes”. In: *Current protocols in bioinformatics* 58, pp. 1.2.1–1.2.12. DOI: 10.1002/cpbi.27.

Anderson, Carl A. et al. (Sept. 2010). “Data quality control in genetic case-control association studies”. In: *Nature protocols* 5.9, pp. 1564–1573. DOI: 10.1038/nprot.2010.116.

Ashburner, Michael et al. (May 2000). “Gene Ontology: tool for the unification of biology”. In: *Nature genetics* 25.1, pp. 25–29. DOI: 10.1038/75556.

Astle, William et al. (Nov. 2009). “Population Structure and Cryptic Relatedness in Genetic Association Studies”. In: *Statistical Science* 24.4. Publisher: Institute of Mathematical Statistics, pp. 451–471. DOI: 10.1214/09-STS307.

Auton, Adam et al. (Oct. 2015). “A global reference for human genetic variation”. en. In: *Nature* 526.7571. Number: 7571 Publisher: Nature Publishing Group, pp. 68–74. DOI: 10.1038/nature15393.

Balas, E. A. et al. (2000). “Managing Clinical Knowledge for Health Care Improvement”. en. In: *Yearbook of Medical Informatics* 09.1. Publisher: Georg Thieme Verlag KG, pp. 65–70. DOI: 10.1055/s-0038-1637943.

Baltzer, Fritz (May 1964). “Theodor Boveri”. In: *Science* 144.3620. Publisher: American Association for the Advancement of Science, pp. 809–815. DOI: 10.1126/science.144.3620.809.

Bandres-Ciga, Sara et al. (Apr. 2020). “Genetics of Parkinson’s disease: An introspection of its journey towards precision medicine”. en. In: *Neurobiology of Disease* 137, p. 104782. DOI: 10.1016/j.nbd.2020.104782.

Barnes, Deborah E et al. (Sept. 2011). “The projected effect of risk factor reduction on Alzheimer’s disease prevalence”. en. In: *The Lancet Neurology* 10.9, pp. 819–828. DOI: 10.1016/S1474-4422(11)70072-2.

Bateson, Patrick (Aug. 2002). “William Bateson: A biologist ahead of his time”. en. In: *Journal of Genetics* 81.2, pp. 49–58. DOI: 10.1007/BF02715900.

Bateson, William (1894). *Materials for the Study of Variation Treated with Especial Regard to Discontinuity in the Origin of Species.* en. Google-Books-ID: _HIZAAAAYAAJ. Macmillan and Company.

— (1905). “A suggestion as to the nature of the "walnut" comb in fowls”. In: *Proceedings of the Cambridge Philosophical Society* 13, pp. 65–168.

Bateson, William et al. (1905). “Experimental studies in the physiology of heredity”. In: vol. 2. Issue: 2 Number: 2. London: The Royal Society of London.

Bateson, William et al. (1909). *Mendel’s principles of heredity, by W. Bateson.* eng. Cambridge [Eng.] University Press.

Becker, Jessica et al. (Jan. 2012). “A systematic eQTL study of cis–trans epistasis in 210 HapMap individuals”. en. In: *European Journal of Human Genetics* 20.1. Number: 1 Publisher: Nature Publishing Group, pp. 97–101. DOI: 10.1038/ejhg.2011.156.

Belbin, Olivia et al. (2019). “The Epistasis Project: A Multi-Cohort Study of the Effects of BDNF, DBH, and SORT1 Epistasis on Alzheimer’s Disease Risk”. eng. In: *Journal of Alzheimer’s disease: JAD* 68.4, pp. 1535–1547. DOI: 10.3233/JAD-181116.

Bell, Graeme I et al. (Feb. 1984). “A Polymorphic Locus Near the Human Insulin Gene Is Associated with Insulin-dependent Diabetes Melliitus”. In: *Diabetes* 33.2, pp. 176–183. DOI: 10.2337/diab.33.2.176.

Benjamini, Yoav et al. (1995). “Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing”. In: *Journal of the Royal Statistical Society. Series B (Methodological)* 57.1. Publisher: [Royal Statistical Society, Wiley], pp. 289–300.

Bloem, Bastiaan R et al. (June 2021). "Parkinson's disease". en. In: *The Lancet* 397.10291, pp. 2284–2303. DOI: 10.1016/S0140-6736(21)00218-X.

Blumenthal, David B et al. (Dec. 2020). "EpiGEN: an epistasis simulation pipeline". In: *Bioinformatics* 36.19, pp. 4957–4959. DOI: 10.1093/bioinformatics/btaa245.

Bogatan, Simina et al. (June 2015). "Talin Is Required Continuously for Cardiomyocyte Remodeling during Heart Growth in Drosophila". en. In: *PLOS ONE* 10.6. Publisher: Public Library of Science, e0131238. DOI: 10.1371/journal.pone.0131238.

Boudellioua, Imane et al. (2018). "OligoPVP: Phenotype-driven analysis of individual genomic information to prioritize oligogenic disease variants." In: *Scientific reports* 8.1. DOI: 10.1038/s41598-018-32876-3.

— (Feb. 2019). "DeepPVP: phenotype-based prioritization of causative variants using deep learning". In: *BMC Bioinformatics* 20. DOI: 10.1186/s12859-019-2633-8.

Boyle, Evan A. et al. (June 2017). "An expanded view of complex traits: from polygenic to omnigenic". In: *Cell* 169.7, pp. 1177–1186. DOI: 10.1016/j.cell.2017.05.038.

Braude, M. C. et al. (Jan. 1986). "Genetic and biological markers in drug abuse and alcoholism". English. In: Publisher: National Institute on Drug Abuse,Rockville, MD.

Brown, Terence A. (2002). *Understanding a Genome Sequence*. en. Publication Title: Genomes. 2nd edition. Wiley-Liss.

Buniello, Annalisa et al. (Jan. 2019). "The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019". eng. In: *Nucleic Acids Research* 47.D1, pp. D1005–D1012. DOI: 10.1093/nar/gky1120.

Bycroft, Clare et al. (Oct. 2018). "The UK Biobank resource with deep phenotyping and genomic data". en. In: *Nature* 562.7726. Bandiera_abtest: a Cc_license_type: cc_by Cg_type: Nature Research Journals Number: 7726 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Genome;Genome-wide association studies;Genotype;Haplotypes;Popu
genetics Subject_term_id: genome;genome-wide-association-studies;genotype;haplotypes;populat
genetics, pp. 203–209. DOI: 10.1038/s41586-018-0579-z.

Carrasquillo, Minerva M. et al. (Oct. 2002). “Genome-wide association study and mouse model identify interaction between RET and EDNRB pathways in Hirschsprung disease”. en. In: *Nature Genetics* 32.2. Number: 2 Publisher: Nature Publishing Group, pp. 237–244. DOI: 10.1038/ng998.

CATTAERT, TOM et al. (Jan. 2011). “A detailed view on Model-Based Multifactor Dimensionality Reduction for detecting gene-gene interactions in case-control data in the absence and presence of noise”. In: *Annals of human genetics* 75.1, pp. 78–89. DOI: 10.1111/j.1469-1809.2010.00604.x.

Chang, Christopher C. et al. (Dec. 2015). “Second-generation PLINK: rising to the challenge of larger and richer datasets”. en. In: *GigaScience* 4.1. Publisher: Oxford Academic. DOI: 10.1186/s13742-015-0047-8.

Chang, Yu-Chuan et al. (Feb. 2020). “GenEpi: gene-based epistasis discovery using machine learning”. en. In: *BMC Bioinformatics* 21.1, p. 68. DOI: 10.1186/s12859-020-3368-2.

Chatelain, Clément et al. (June 2018). “Performance of epistasis detection methods in semi-simulated GWAS”. In: *BMC Bioinformatics* 19.1, p. 231. DOI: 10.1186/s12859-018-2229-8.

Chattopadhyay, Amrita et al. (Dec. 2019). “Gene-gene interaction: the curse of dimensionality”. In: *Annals of Translational Medicine* 7.24. DOI: 10.21037/atm.2019.12.87.

Chee, Mark et al. (Oct. 1996). “Accessing Genetic Information with High-Density DNA Arrays”. In: *Science* 274.5287. Publisher: American Association for the Advancement of Science, pp. 610–614. DOI: 10.1126/science.274.5287.610.

Chen, Zhifen et al. (2021). “Genetics of coronary artery disease in the post-GWAS era”. en. In: *Journal of Internal Medicine* 290.5. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/joim.1
pp. 980–992. DOI: 10.1111/joim.13362.

Chung, Yujin et al. (Jan. 2007). “Odds ratio based multifactor-dimensionality reduction method for detecting gene–gene interactions”. en. In: *Bioinformatics* 23.1. Publisher: Oxford Academic, pp. 71–76. DOI: 10.1093/bioinformatics/btl557.

Cisterna-Garcia, Alejandro et al. (Aug. 2022). *Genome-wide epistasis analysis in Parkinson's disease between populations with different genetic ancestry reveals significant variant-variant interactions*. en. Pages: 2022.07.29.22278162. DOI: 10.1101/2022.07.29.22278162.

Citterio, Cintia E. et al. (June 2019). "The role of thyroglobulin in thyroid hormonogenesis". en. In: *Nature Reviews Endocrinology* 15.6. Number: 6 Publisher: Nature Publishing Group, pp. 323–338. DOI: 10.1038/s41574-019-0184-8.

Clark, Andrew G. (2004). "The role of haplotypes in candidate gene studies". en. In: *Genetic Epidemiology* 27.4. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/gepi.20025, pp. 321–333. DOI: 10.1002/gepi.20025.

Coleman, Jonathan R. I. et al. (July 2016). "Quality control, imputation and analysis of genome-wide genotyping data from the Illumina HumanCoreExome microarray". In: *Briefings in Functional Genomics* 15.4, pp. 298–304. DOI: 10.1093/bfgp/elv037.

Combarros, Onofre et al. (Aug. 2009). "Replication by the Epistasis Project of the interaction between the genes for IL-6 and IL-10 in the risk of Alzheimer's disease". In: *Journal of Neuroinflammation* 6.1, p. 22. DOI: 10.1186/1742-2094-6-22.

Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing - Benjamini - 1995 - Journal of the Royal Statistical Society: Series B (Methodological) - Wiley Online Library (n.d.).

Cooke, Daniel P. et al. (July 2021). "A unified haplotype-based method for accurate and comprehensive variant calling". In: *Nature Biotechnology* 39.7. Number: 7 Publisher: Nature Publishing Group, pp. 885–892. DOI: 10.1038/s41587-021-00861-3.

Cooper, Thomas A. et al. (Feb. 2009). "RNA and disease". eng. In: *Cell* 136.4, pp. 777–793. DOI: 10.1016/j.cell.2009.02.011.

Cordell, Heather J. (Oct. 2002). "Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans". en. In: *Human Molecular Genetics* 11.20. Publisher: Oxford Academic, pp. 2463–2468. DOI: 10.1093/hmg/11.20.2463.

Correns, Carl Franz Joseph Erich (1900). “Mendel’s Regel uber das Verhalten der Nachkommenschaft der Rassenbastarde”. In: *Ber. dtsch. botanisch. Ges* 18, pp. 158–167.

Dai, Zhijun et al. (Jan. 2020). “Influence of Genetic Interactions on Polygenic Prediction”. In: *G3 Genes|Genomes|Genetics* 10.1, pp. 109–115. DOI: 10.1534/g3.119.400812.

Darden, Lindley (Mar. 1977). “William Bateson and the promise of Mendelism”. In: *Journal of the History of Biology* 10, pp. 87–106. DOI: 10.1007/BF00126096.

Desmet, François-Olivier et al. (May 2009). “Human Splicing Finder: an online bioinformatics tool to predict splicing signals”. In: *Nucleic Acids Research* 37.9, e67. DOI: 10.1093/nar/gkp215.

Donaldson, Peter et al. (Aug. 2015). *Genetics of Complex Disease*. en. Google-Books-ID: 5JVwCgAAQBAJ. Garland Science. ISBN: 978-1-317-33425-5.

Dorigo, Marco et al. (Nov. 2006). “Ant colony optimization”. In: *IEEE Computational Intelligence Magazine* 1.4. Conference Name: IEEE Computational Intelligence Magazine, pp. 28–39. DOI: 10.1109/MCI.2006.329691.

Drake, John W et al. (Apr. 1998). “Rates of Spontaneous Mutation”. In: *Genetics* 148.4, pp. 1667–1686. DOI: 10.1093/genetics/148.4.1667.

Edwards, A. W. F. (July 2008). “G. H. Hardy (1908) and Hardy–Weinberg Equilibrium”. en. In: *Genetics* 179.3. Publisher: Genetics Section: Perspectives, pp. 1143–1150. DOI: 10.1534/genetics.104.92940.

Edwards, J. H. (Dec. 1992). “Haldane and the mutation rate”. en. In: *Journal of Genetics* 71.3, p. 81. DOI: 10.1007/BF02927889.

Efron, Bradley (1998). “R. A. Fisher in the 21st Century”. In: *Statistical Science* 13.2. Publisher: Institute of Mathematical Statistics, pp. 95–114.

Elandt-johnson, R. C. (1971). “Probability models and statistical methods in genetics.” English. In: *Probability models and statistical methods in genetics.* Publisher: New York, London, Sydney, Toronto: John Wiley & Sons, Inc.

Ellinor, Patrick T. et al. (June 2012). “Meta-analysis identifies six new susceptibility loci for atrial fibrillation”. en. In: *Nature Genetics* 44.6. Bandiera_abtest: a Cg_type: Nature Research Journals Number: 6 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Cardiovascular diseases;Disease genetics;Genome-wide association studies Subject_term_id: cardiovascular-diseases;disease-genetics;genome-wide-association-studies, pp. 670–675. DOI: 10.1038/ng.2261.

Emigh, Ted H. (1980). “A Comparison of Tests for Hardy-Weinberg Equilibrium”. In: *Biometrics* 36.4. Publisher: [Wiley, International Biometric Society], pp. 627–642. DOI: 10.2307/2556115.

“Etymologia” (Feb. 2015). “Etymologia: Bonferroni Correction”. In: *Emerging Infectious Diseases* 21.2, p. 289. DOI: 10.3201/eid2102.ET2102.

Evans, David M. et al. (Sept. 2006). “Two-Stage Two-Locus Models in Genome-Wide Association”. en. In: *PLOS Genetics* 2.9. Publisher: Public Library of Science, e157. DOI: 10.1371/journal.pgen.0020157.

Fabregat, Antonio et al. (Mar. 2017). “Reactome pathway analysis: a high-performance in-memory approach”. In: *BMC Bioinformatics* 18.1, p. 142. DOI: 10.1186/s12859-017-1559-2.

Fang, Yao-Hwei et al. (2012). “SVM-Based Generalized Multifactor Dimensionality Reduction Approaches for Detecting Gene-Gene Interactions in Family Studies”. en. In: *Genetic Epidemiology* 36.2. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/gepi.21602, pp. 88–98. DOI: 10.1002/gepi.21602.

Faqeih, Eissa A. et al. (2014). “Novel homozygous DEAF1 variant suspected in causing white matter disease, intellectual disability, and microcephaly”. en. In: *American Journal of Medical Genetics Part A* 164.6. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/ajmg.a.364
pp. 1565–1570. DOI: 10.1002/ajmg.a.36482.

Fernández-Santiago, Rubén et al. (Sept. 2019). “SNCA and mTOR Pathway Single Nucleotide Polymorphisms Interact to Modulate the Age at Onset of Parkinson’s Disease”. eng. In:

Movement Disorders: Official Journal of the Movement Disorder Society 34.9, pp. 1333–1344. DOI: 10.1002/mds.27770.

Firth, Helen V. et al. (Aug. 2011). “The Deciphering Developmental Disorders (DDD) study”. eng. In: *Developmental Medicine and Child Neurology* 53.8, pp. 702–703. DOI: 10.1111/j.1469-8749.2011.04032.x.

Fisher, R. A. (1919). “XV.—The Correlation between Relatives on the Supposition of Mendelian Inheritance.” en. In: *Earth and Environmental Science Transactions of The Royal Society of Edinburgh* 52.2. Publisher: Royal Society of Edinburgh Scotland Foundation, pp. 399–433. DOI: 10.1017/S0080456800012163.

Frankish, Adam et al. (Jan. 2021). “GENCODE 2021”. In: *Nucleic Acids Research* 49.D1, pp. D916–D923. DOI: 10.1093/nar/gkaa1087.

Frayling, TM (Aug. 2014). “Genome-wide association studies: the good, the bad and the ugly”. In: *Clinical Medicine* 14.4, pp. 428–431. DOI: 10.7861/clinmedicine.14-4-428.

GAMETES (July 2021). original-date: 2019-04-23T14:59:32Z.

Gan, Rongxing et al. (Feb. 2015). “Chinese Biobanking Initiatives”. In: *Biopreservation and Biobanking* 13.1. Publisher: Mary Ann Liebert, Inc., publishers, pp. 4–7. DOI: 10.1089/bio.2014.0096.

Gao, Hong et al. (Mar. 2010). “On the Classification of Epistatic Interactions”. en. In: *Genetics* 184.3. Publisher: Genetics Section: Investigations, pp. 827–837. DOI: 10.1534/genetics.109.111120.

Gibbs, Richard A. (Aug. 2020). “The Human Genome Project changed everything”. In: *Nature Reviews Genetics* 21, pp. 575–576. DOI: https://doi.org/10.1038/s41576-020-0275-3.

Gilbert, Walter (Dec. 1981). “DNA Sequencing and Gene Structure”. In: *Science* 214.4527. Publisher: American Association for the Advancement of Science, pp. 1305–1312. DOI: 10.1126/science.7313687.

Girirajan, Santhosh (May 2017). “Missing heritability and where to find it”. In: *Genome Biology* 18.1, p. 89. DOI: 10.1186/s13059-017-1227-x.

Goswami, Chayanika et al. (June 2021). “Rare variants: data types and analysis strategies”. In: *Annals of Translational Medicine* 9.12, p. 961. DOI: 10.21037/atm-21-1635.

Goudey, Benjamin et al. (May 2013). “GWIS - model-free, fast and exhaustive search for epistatic interactions in case-control GWAS”. en. In: *BMC Genomics* 14.3, S10. DOI: 10.1186/1471-2164-14-S3-S10.

GRCh38 (n.d.).

Guan, Boxin et al. (Dec. 2018). “Ant colony optimization with an automatic adjustment mechanism for detecting epistatic interactions”. en. In: *Computational Biology and Chemistry* 77, pp. 354–362. DOI: 10.1016/j.compbiolchem.2018.11.001.

Gudbjartsson, Daniel F. et al. (May 2008). “Many sequence variants affecting diversity of adult human height”. en. In: *Nature Genetics* 40.5. Number: 5 Publisher: Nature Publishing Group, pp. 609–615. DOI: 10.1038/ng.122.

Gui, Jiang et al. (Jan. 2011). “A novel survival multifactor dimensionality reduction method for detecting gene–gene interactions with application to bladder cancer prognosis”. en. In: *Human Genetics* 129.1, pp. 101–110. DOI: 10.1007/s00439-010-0905-5.

Guo, Yan et al. (Mar. 2017). “Improvements and impacts of GRCh38 human reference on high throughput sequencing data analysis”. en. In: *Genomics* 109.2, pp. 83–90. DOI: 10.1016/j.ygeno.2017.01.005.

Guo, Yang et al. (Aug. 2019). “Epi-GTBN: an approach of epistasis mining based on genetic Tabu algorithm and Bayesian network”. In: *BMC Bioinformatics* 20.1, p. 444. DOI: 10.1186/s12859-019-3022-z.

Hahn, Lance W. et al. (Feb. 2003). “Multifactor dimensionality reduction software for detecting gene–gene and gene–environment interactions”. In: *Bioinformatics* 19.3, pp. 376–382. DOI: 10.1093/bioinformatics/btf869.

Hahn, Lance W. et al. (2004). “Ideal discrimination of discrete clinical endpoints using multilocus genotypes”. eng. In: *In Silico Biology* 4.2, pp. 183–194.

Haldane, JBS (1919). “The combination of linkage values, and the calculation of distance between the loci of linked factors – ScienceOpen”. In: *Journal of Genetics* 8, pp. 299–309.

Hall, Jeff M. et al. (Dec. 1990). “Linkage of Early-Onset Familial Breast Cancer to Chromosome 17q21”. In: *Science* 250.4988. Publisher: American Association for the Advancement of Science, pp. 1684–1689. DOI: 10.1126/science.2270482.

Hallgrímsdóttir, Ingileif B et al. (Feb. 2008). “A complete classification of epistatic two-locus models”. In: *BMC Genetics* 9, p. 17. DOI: 10.1186/1471-2156-9-17.

Hammer, Andreas et al. (2021). “The impact of CD4+CD28null T lymphocytes on atrial fibrillation: a potential pathophysiological pathway”. In: *Inflammation Research* 70.10-12, pp. 1011–1014. DOI: 10.1007/s00011-021-01502-w.

Hardy, G. H. (July 1908). “MENDELIAN PROPORTIONS IN A MIXED POPULATION”. eng. In: *Science (New York, N.Y.)* 28.706, pp. 49–50. DOI: 10.1126/science.28.706.49.

Harrow, Jennifer et al. (Sept. 2012). “GENCODE: The reference human genome annotation for The ENCODE Project”. In: *Genome Research* 22.9, pp. 1760–1774. DOI: 10.1101/gr.135350.111.

Hassold, Terry et al. (Oct. 2007). “The origin of human aneuploidy: where we have been, where we are going”. In: *Human Molecular Genetics* 16.R2, R203–R208. DOI: 10.1093/hmg/ddm243.

Hill, W. G. et al. (June 1968). “Linkage disequilibrium in finite populations”. en. In: *Theoretical and Applied Genetics* 38.6, pp. 226–231. DOI: 10.1007/BF01245622.

Hill, William (Dec. 1990). “Sewall Wright, 21 December 1889 - 3 March 1988”. In: *Biographical Memoirs of Fellows of the Royal Society* 36. Publisher: Royal Society, pp. 567–579. DOI: 10.1098/rsbm.1990.0044.

Hirschhorn, Joel N. et al. (Mar. 2002). “A comprehensive review of genetic association studies”. en. In: *Genetics in Medicine* 4.2. Number: 2 Publisher: Nature Publishing Group, pp. 45–61. DOI: 10.1097/00125817-200203000-00002.

Hirschhorn, Joel N. et al. (Feb. 2005). "Genome-wide association studies for common diseases and complex traits". en. In: *Nature Reviews Genetics* 6.2. Number: 2 Publisher: Nature Publishing Group, pp. 95–108. DOI: 10.1038/nrg1521.

Hoehndorf, Robert et al. (Oct. 2011). "PhenomeNET: a whole-phenome approach to disease gene discovery". In: *Nucleic Acids Research* 39.18, e119. DOI: 10.1093/nar/gkr538.

Holliday, Gemma L. et al. (2017). "Evaluating Functional Annotations of Enzymes Using the Gene Ontology". In: *Methods in molecular biology (Clifton, N.J.)* 1446, pp. 111–132. DOI: 10.1007/978-1-4939-3743-1_9.

Hu, Wenchao et al. (June 2014). "Using Nonlinear Optical Networks for Optimization: Primer of the Ant Colony Algorithm". EN. In: *CLEO: 2014 (2014), paper FM1D.8.* Optical Society of America, FM1D.8. DOI: 10.1364/CLEO_QELS.2014.FM1D.8.

Human Genome Project Fact Sheet (n.d.). en.

Huppertz, Berthold et al. (July 2016). "Biobank Graz: The Hub for Innovative Biomedical Research". en. In: *Open Journal of Bioresources* 3.1. Number: 1 Publisher: Ubiquity Press, e3. DOI: 10.5334/ojb.20.

Ibrahim, Z. M. et al. (Apr. 2013). "Detecting epistasis in the presence of linkage disequilibrium: A focused comparison". In: *2013 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, pp. 96–103. DOI: 10.1109/CIBCB.2013.6595394.

Ikegawa, Shiro (Dec. 2012). "A Short History of the Genome-Wide Association Study: Where We Were and Where We Are Going". In: *Genomics & Informatics* 10.4, pp. 220–225. DOI: 10.5808/GI.2012.10.4.220.

"Improving databases for human variation" (Feb. 2016). en. In: *Nature Methods* 13.2. Number: 2 Publisher: Nature Publishing Group, pp. 103–103. DOI: 10.1038/nmeth.3762.

Input filtering - PLINK 1.9 (n.d.).

Ioannidis, Nilah M. et al. (Oct. 2016). “REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants”. en. In: *The American Journal of Human Genetics* 99.4, pp. 877–885. DOI: 10.1016/j.ajhg.2016.08.016.

Jing, Peng-Jie et al. (Mar. 2015). “MACOED: a multi-objective ant colony optimization algorithm for SNP epistasis detection in genome-wide association studies”. In: *Bioinformatics* 31.5, pp. 634–641. DOI: 10.1093/bioinformatics/btu702.

Joiret, Marc et al. (June 2019). “Confounding of linkage disequilibrium patterns in large scale DNA based gene-gene interaction studies”. In: *BioData Mining* 12, p. 11. DOI: 10.1186/s13040-019-0199-7.

Jun, Goo et al. (Nov. 2012). “Detecting and Estimating Contamination of Human DNA Samples in Sequencing and Array-Based Genotype Data”. In: *American Journal of Human Genetics* 91.5, pp. 839–848. DOI: 10.1016/j.ajhg.2012.09.004.

Kallioniemi, Olli-P (Jan. 2001). “Biochip technologies in cancer research”. In: *Annals of Medicine* 33.2, pp. 142–147. DOI: 10.3109/07853890109002069.

Kamat, Mihir A et al. (Nov. 2019). “PhenoScanner V2: an expanded tool for searching human genotype–phenotype associations”. In: *Bioinformatics* 35.22, pp. 4851–4853. DOI: 10.1093/bioinformatics/btz469.

Kanehisa, Minoru et al. (Jan. 2017). “KEGG: new perspectives on genomes, pathways, diseases and drugs”. In: *Nucleic Acids Research* 45.Database issue, pp. D353–D361. DOI: 10.1093/nar/gkw1092.

Keynes, Milo et al. (Mar. 2008). “William Bateson, the rediscoverer of Mendel”. In: *Journal of the Royal Society of Medicine* 101.3, p. 104. DOI: 10.1258/jrsm.2008.081011.

Klein, Robert J (Aug. 2007). “Power analysis for genome-wide association studies”. In: *BMC Genetics* 8, p. 58. DOI: 10.1186/1471-2156-8-58.

Klein, Robert J. et al. (Apr. 2005). “Complement Factor H Polymorphism in Age-Related Macular Degeneration”. In: *Science (New York, N.Y.)* 308.5720, pp. 385–389. DOI: 10.1126/science.1109557.

Köhler, Sebastian et al. (Jan. 2021). "The Human Phenotype Ontology in 2021". In: *Nucleic Acids Research* 49.D1, pp. D1207–D1217. DOI: 10.1093/nar/gkaa1043.

Kooperberg, Charles et al. (2008). "Increasing the power of identifying gene × gene interactions in genome-wide association studies". en. In: *Genetic Epidemiology* 32.3. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/gepi.20300, pp. 255–263. DOI: 10.1002/gepi.20300.

Laird, Nan M. et al. (2011). *The Fundamentals of Modern Statistical Genetics*. Statistics for Biology and Health. New York, NY: Springer. DOI: 10.1007/978-1-4419-7338-2.

Lambert, Christophe G. et al. (Apr. 2012). "Learning from our GWAS mistakes: from experimental design to scientific method". In: *Biostatistics (Oxford, England)* 13.2, pp. 195–203. DOI: 10.1093/biostatistics/kxr055.

Landegren, Ulf et al. (Jan. 1998). "Reading Bits of Genetic Information: Methods for Single-Nucleotide Polymorphism Analysis". en. In: *Genome Research* 8.8. Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab, pp. 769–776. DOI: 10.1101/gr.8.8.769.

Lander, Eric S. et al. (Feb. 2001). "Initial sequencing and analysis of the human genome". en. In: *Nature* 409.6822. Number: 6822 Publisher: Nature Publishing Group, pp. 860–921. DOI: 10.1038/35057062.

Landrum, Melissa J et al. (Jan. 2018). "ClinVar: improving access to variant interpretations and supporting evidence". In: *Nucleic Acids Research* 46, pp. D1062–D1067. DOI: 10.1093/nar/gkx1153.

Lange, Kenneth (1997). *Mathematical and Statistical Methods for Genetic Analysis*. Ed. by K. Dietz et al. Statistics for Biology and Health. New York, NY: Springer. DOI: 10.1007/978-1-4757-2739-5.

Lappalainen, Ilkka et al. (Jan. 2013). "dbVar and DGVa: public archives for genomic structural variation". In: *Nucleic Acids Research* 41.D1, pp. D936–D941. DOI: 10.1093/nar/gks1213.

Lareau, Caleb A. et al. (2015). “Network Theory for Data-Driven Epistasis Networks”. en. In: *Epistasis: Methods and Protocols*. Ed. by Jason H. Moore et al. Methods in Molecular Biology. New York, NY: Springer, pp. 285–300. ISBN: 978-1-4939-2155-3. DOI: 10.1007/978-1-4939-2155-3_15.

Lau, Alexandria et al. (Jan. 2020). “Turning genome-wide association study findings into opportunities for drug repositioning”. In: *Computational and Structural Biotechnology Journal* 18, pp. 1639–1650. DOI: 10.1016/j.csbj.2020.06.015.

Lee, C. (2007). “Coimmunoprecipitation assay”. In: *Methods Mol Biol.* 362. DOI: 10.1007/978-1-59745-257-1_31.

Leem, Sangseob et al. (Mar. 2017). “An empirical fuzzy multifactor dimensionality reduction method for detecting gene-gene interactions”. In: *BMC Genomics* 18.Suppl 2. DOI: 10.1186/s12864-017-3496-x.

Lehmann, Donald J. et al. (Jan. 2012). “Transferrin and HFE genes interact in Alzheimer’s disease risk: the Epistasis Project”. en. In: *Neurobiology of Aging* 33.1, 202.e1–202.e13. DOI: 10.1016/j.neurobiolaging.2010.07.018.

Lehnen, Harald et al. (July 2013). “Epigenetics of gestational diabetes mellitus and offspring health: the time for action is in early stages of life”. eng. In: *Molecular Human Reproduction* 19.7, pp. 415–422. DOI: 10.1093/molehr/gat020.

Lewontin, R. C. (Jan. 1964). “The Interaction of Selection and Linkage. I. General Considerations; Heterotic Models”. In: *Genetics* 49.1, pp. 49–67.

Lewontin, R. C. et al. (1960). “The Evolutionary Dynamics of Complex Polymorphisms”. In: *Evolution* 14.4. Publisher: [Society for the Study of Evolution, Wiley], pp. 458–472. DOI: 10.2307/2405995.

Li, Yun et al. (2009). “Genotype Imputation”. In: *Annual review of genomics and human genetics* 10, pp. 387–406. DOI: 10.1146/annurev.genom.9.081307.164242.

Linz, Benedikt et al. (June 2020). “Pharmacological inhibition of sodium-proton-exchanger subtype 3-mediated sodium absorption in the gut reduces atrial fibrillation susceptibility in obese spontaneously hypertensive rats”. eng. In: *International Journal of Cardiology. Heart & Vasculature* 28, p. 100534. DOI: 10.1016/j.ijcha.2020.100534.

Lishout, François Van et al. (Nov. 2015). “gammaMAXT: a fast multiple-testing correction algorithm”. In: *BioData Mining* 8.1, p. 36. DOI: 10.1186/s13040-015-0069-x.

Louche A. Salcedo, S. et al. (2017). “Protein-Protein Interactions: Pull-Down Assays”. In: *Methods Mol Biol.* 1615. DOI: 10.1007/978-1-4939-7033-9_20.

Lygirou, Vasiliki et al. (Apr. 2018). “Plasma proteomic analysis reveals altered protein abundances in cardiovascular disease”. In: *Journal of Translational Medicine* 16.1, p. 104. DOI: 10.1186/s12967-018-1476-9.

MacArthur, Jacqueline et al. (Jan. 2017). “The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog)”. In: *Nucleic Acids Research* 45.D1, pp. D896–D901. DOI: 10.1093/nar/gkw1133.

MacDonald, Marcy E. et al. (May 1992). “The Huntington’s disease candidate region exhibits many different haplotypes”. en. In: *Nature Genetics* 1.2. Number: 2 Publisher: Nature Publishing Group, pp. 99–103. DOI: 10.1038/ng0592-99.

Macgregor, Stuart et al. (Apr. 2006). “GAIA: An easy-to-use web-based application for interaction analysis of case-control data”. In: *BMC Medical Genetics* 7, p. 34. DOI: 10.1186/1471-2350-7-34.

Maher, Brendan (Nov. 2008). “Personal genomes: The case of the missing heritability”. en. In: *Nature* 456.7218. Number: 7218 Publisher: Nature Publishing Group, pp. 18–21. DOI: 10.1038/456018a.

Manichaikul, Ani et al. (Nov. 2010). “Robust relationship inference in genome-wide association studies”. In: *Bioinformatics* 26.22, pp. 2867–2873. DOI: 10.1093/bioinformatics/btq559.

Mann, Stefan A. et al. (Mar. 2012). “Epistatic Effects of Potassium Channel Variation on Cardiac Repolarization and Atrial Fibrillation Risk”. In: *Journal of the American College of Cardiology* 59.11. Publisher: American College of Cardiology Foundation, pp. 1017–1025. DOI: 10.1016/j.jacc.2011.11.039.

Manolio, Teri A. et al. (Oct. 2009). “Finding the missing heritability of complex diseases”. en. In: *Nature* 461.7265. Number: 7265 Publisher: Nature Publishing Group, pp. 747–753. DOI: 10.1038/nature08494.

Marees, Andries T. et al. (Feb. 2018). “A tutorial on conducting genome-wide association studies: Quality control and statistical analysis”. In: *International Journal of Methods in Psychiatric Research* 27.2, e1608. DOI: 10.1002/mpr.1608.

Marie, De Vries Hugo (1900). “Das Spaltungsgesetz der Bastarde”. In: *Ber. dtsch. hot. Ges* 18, pp. 83–90.

Martin, Antonio Rueda et al. (Nov. 2019). “PanelApp crowdsources expert knowledge to establish consensus diagnostic gene panels”. en. In: *Nature Genetics* 51.11. Number: 11 Publisher: Nature Publishing Group, pp. 1560–1565. DOI: 10.1038/s41588-019-0528-2.

Martin, E. R. et al. (2006). “A novel method to identify gene–gene effects in nuclear families: the MDR-PDT”. fr. In: *Genetic Epidemiology* 30.2. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1
pp. 111–123. DOI: https://doi.org/10.1002/gepi.20128.

Mathé, Catherine et al. (Oct. 2002). “Current methods of gene prediction, their strengths and weaknesses”. In: *Nucleic Acids Research* 30.19, pp. 4103–4117. DOI: 10.1093/nar/gkf543.

Mbatchou, Joelle et al. (July 2021). “Computationally efficient whole-genome regression for quantitative and binary traits”. en. In: *Nature Genetics* 53.7. Number: 7 Publisher: Nature Publishing Group, pp. 1097–1103. DOI: 10.1038/s41588-021-00870-7.

McCarthy, Mark I. (May 2017). “Painting a new picture of personalised medicine for diabetes”. en. In: *Diabetologia* 60.5, pp. 793–799. DOI: 10.1007/s00125-017-4210-x.

McCarthy, Mark I. et al. (May 2008). “Genome-wide association studies for complex traits: consensus, uncertainty and challenges”. en. In: *Nature Reviews Genetics* 9.5. Number: 5 Publisher: Nature Publishing Group, pp. 356–369. DOI: 10.1038/nrg2344.

McDonald, Michael J. et al. (June 2011). “Clusters of Nucleotide Substitutions and Insertion/Deletion Mutations Are Associated with Repeat Sequences”. In: *PLoS Biology* 9.6, e1000622. DOI: 10.1371/journal.pbio.1000622.

McGall, Glenn H. et al. (2002). “High-Density GeneChip Oligonucleotide Probe Arrays”. en. In: *Chip Technology*. Ed. by Jörg Hoheisel et al. Advances in Biochemical Engineering/Biotechnology. Berlin, Heidelberg: Springer, pp. 21–42. ISBN: 978-3-540-45713-8. DOI: 10.1007/3-540-45713-5_2.

McKinney, Brett A. et al. (2006). “Machine Learning for Detecting Gene-Gene Interactions”. In: *Applied bioinformatics* 5.2, pp. 77–88.

McLaren, William et al. (June 2016). “The Ensembl Variant Effect Predictor”. In: *Genome Biology* 17.1, p. 122. DOI: 10.1186/s13059-016-0974-4.

Mega, J. L. et al. (June 2015). “Genetic risk, coronary heart disease events, and the clinical benefit of statin therapy: an analysis of primary and secondary prevention trials”. eng. In: *Lancet (London, England)* 385.9984, pp. 2264–2271. DOI: 10.1016/S0140-6736(14)61730-X.

Menozzi, P. et al. (Sept. 1978). “Synthetic Maps of Human Gene Frequencies in Europeans”. In: *Science* 201.4358. Publisher: American Association for the Advancement of Science, pp. 786–792. DOI: 10.1126/science.356262.

Milligan, Brook G (Mar. 2003). “Maximum-likelihood estimation of relatedness.” In: *Genetics* 163.3, pp. 1153–1167.

Millstein, Joshua et al. (Dec. 2005). “Identifying susceptibility genes by using joint tests of association and linkage and accounting for epistasis”. In: *BMC Genetics* 6.Suppl 1, S147. DOI: 10.1186/1471-2156-6-S1-S147.

Miranda, Magdalena et al. (2019). “Brain-Derived Neurotrophic Factor: A Key Molecule for Memory in the Healthy and the Pathological Brain”. In: *Frontiers in Cellular Neuroscience* 13.

Mn, Weedon et al. (Feb. 2021). “Use of SNP chips to detect rare pathogenic variants: retrospective, population based diagnostic evaluation”. eng. In: *BMJ (Clinical research ed.)* 372, n214. DOI: 10.1136/bmj.n214.

Moore, Jason H. et al. (July 2006). “A flexible computational framework for detecting, characterizing, and interpreting statistical patterns of epistasis in genetic studies of human disease susceptibility”. en. In: *Journal of Theoretical Biology* 241.2, pp. 252–261. DOI: 10.1016/j.jtbi.2005.11.036.

Morgan, T.H. et al. (1915). *The Mechanism of Mendelian heredity*. The Mechanism of Mendelian heredity. Pages: Pp. xiii+, 262. Oxford, England: Holt. DOI: 10.5962/bhl.title.6001.

Morris, A. P. et al. (2007). “Whole Genome Association”. en. In: *Handbook of Statistical Genetics*. Section: 37 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/9780470061619.ch37. John Wiley & Sons, Ltd, pp. 1238–1263. ISBN: 978-0-470-06161-9. DOI: 10.1002/9780470061619.ch37.

Morton, Newton E. (Sept. 1955). “Sequential tests for the detection of linkage”. In: *American Journal of Human Genetics* 7.3, pp. 277–318.

Muzzey D. Evans, E. A. et al. (Sept. 2015). “Understanding the Basics of NGS: From Mechanism to Variant Calling”. In: *Current Genetic Medicine Reports* 3. DOI: https://doi.org/10.1007/s40142-015-0076-8.

Myers, Timothy A. et al. (2020). “LDlinkR: An R Package for Rapidly Calculating Linkage Disequilibrium Statistics in Diverse Populations”. In: *Frontiers in Genetics* 11.

Natarajan, Pradeep et al. (May 2017). “Polygenic Risk Score Identifies Subgroup With Higher Burden of Atherosclerosis and Greater Relative Benefit From Statin Therapy in the Primary Prevention Setting”. In: *Circulation* 135.22, pp. 2091–2101. DOI: 10.1161/CIRCULATIONAHA.116.024436.

Ng, Pauline C. et al. (July 2003). “SIFT: predicting amino acid changes that affect protein function”. In: *Nucleic Acids Research* 31.13, pp. 3812–3814.

Nimwegen, K. J. van et al. (Nov. 2016). “Is the $1000 Genome as Near as We Think? A Cost Analysis of Next-Generation Sequencing”. In: *Clinical Chemistry* 62. DOI: https://doi.org/10.1373/clinchem.2016.258632.

Noor, Nurulamin M et al. (Jan. 2020). “Personalised medicine in Crohn’s disease”. en. In: *The Lancet Gastroenterology & Hepatology* 5.1, pp. 80–92. DOI: 10.1016/S2468-1253(19)30340-1.

Norton, Bernard et al. (July 1976). “A note on the background to, and refereeing of, R. A. Fisher’s 1918 paper ‘On the correlation between relatives on the supposition of Mendelian inheritance’”. In: *Notes and Records of the Royal Society of London* 31.1. Publisher: Royal Society, pp. 151–162. DOI: 10.1098/rsnr.1976.0005.

Nurk, Sergey et al. (Apr. 2022). “The complete sequence of a human genome”. In: *Science* 376.6588. Publisher: American Association for the Advancement of Science, pp. 44–53. DOI: 10.1126/science.abj6987.

Orntoft, Torben F et al. (2006). “DNA Chips and Microarrays”. en. In: *eLS*. _eprint: https://onlinelibrary.wiley John Wiley & Sons, Ltd. ISBN: 978-0-470-01590-2. DOI: 10.1038/npg.els.0005675.

Otani, Takahiro et al. (Jan. 2019). “Exploring predictive biomarkers from clinical genome-wide association studies via multidimensional hierarchical mixture models”. In: *European Journal of Human Genetics* 27.1, pp. 140–149. DOI: 10.1038/s41431-018-0251-y.

Otwinowski, Jakub et al. (Aug. 2018). “Inferring the shape of global epistasis”. en. In: *Proceedings of the National Academy of Sciences* 115.32. Publisher: National Academy of Sciences Section: PNAS Plus, E7550–E7558. DOI: 10.1073/pnas.1804015115.

Oughtred, Rose et al. (Jan. 2021). “The BioGRID database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions”. In: *Protein Science : A Publication of the Protein Society* 30, pp. 187–200. DOI: 10.1002/pro.3978.

Paci, P. et al. (2021). “Gene coexpression in the interactome moving from correlation toward causation via an integrated approach to disease module discovery”. In: *npj Syst Biol Appl* 7. DOI: https://doi.org/10.1038/s41540-020-00168-0.

Park, Leeyoung (Aug. 2019). “Population-specific long-range linkage disequilibrium in the human genome and its influence on identifying common disease variants”. en. In: *Scientific Reports* 9.1. Number: 1 Publisher: Nature Publishing Group, p. 11380. DOI: 10.1038/s41598-019-47832-y.

Paththinige, Chamara S. et al. (2019). "The Frequency and Spectrum of Chromosomal Translocations in a Cohort of Sri Lankans". en. In: *BioMed Research International* 2019. Publisher: Hindawi Limited. DOI: 10.1155/2019/9797104.

Patterson, Nick et al. (Dec. 2006). "Population Structure and Eigenanalysis". In: *PLoS Genetics* 2.12, e190. DOI: 10.1371/journal.pgen.0020190.

Piegorsch, Walter W. (1990). "Fisher's Contributions to Genetics and Heredity, with Special Emphasis on the Gregor Mendel Controversy". In: *Biometrics* 46.4. Publisher: [Wiley, International Biometric Society], pp. 915–924. DOI: 10.2307/2532437.

Ponte-Fernández, Christian et al. (Jan. 2020). "Fast search of third-order epistatic interactions on CPU and GPU clusters". en. In: *The International Journal of High Performance Computing Applications* 34.1. Publisher: SAGE Publications Ltd STM, pp. 20–29. DOI: 10.1177/1094342019852128.

Popejoy, A. et al. (Oct. 2016). "Genomics is failing on diversity". In: *Nature* 538. DOI: https://doi.org/10.1038/538161a.

Price, Alkes L. et al. (Aug. 2006). "Principal components analysis corrects for stratification in genome-wide association studies". en. In: *Nature Genetics* 38.8. Number: 8 Publisher: Nature Publishing Group, pp. 904–909. DOI: 10.1038/ng1847.

Punnett, Reginald Crundall (May 2009). *MendelismThird Edition*. English.

Purcell, Shaun et al. (Sept. 2007). "PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses". In: *American Journal of Human Genetics* 81.3, pp. 559–575.

Quang, Daniel et al. (Mar. 2015). "DANN: a deep learning approach for annotating the pathogenicity of genetic variants". In: *Bioinformatics* 31.5, pp. 761–763. DOI: 10.1093/bioinformatics/btu703.

Raghavan, Neha et al. (Aug. 2017). "Genetics of Alzheimer's Disease: the Importance of Polygenic and Epistatic Components". en. In: *Current Neurology and Neuroscience Reports* 17.10, p. 78. DOI: 10.1007/s11910-017-0787-1.

Rastogi, Maynika V. et al. (June 2010). “Congenital hypothyroidism”. In: *Orphanet Journal of Rare Diseases* 5.1, p. 17. DOI: 10.1186/1750-1172-5-17.

Reay, William R. et al. (Oct. 2021). “Advancing the use of genome-wide association studies for drug repurposing”. en. In: *Nature Reviews Genetics* 22.10. Number: 10 Publisher: Nature Publishing Group, pp. 658–671. DOI: 10.1038/s41576-021-00387-z.

Reid, Steven (Nov. 2010). “What is so normal about the normal distribution?” en. In: *Evidence-Based Mental Health* 13.4. Publisher: Royal College of Psychiatrists Section: EBMH Notebook, pp. 100–100. DOI: 10.1136/ebmh.13.4.100.

Reimand, Jüri et al. (July 2007). “g:Profiler—a web-based toolset for functional profiling of gene lists from large-scale experiments”. In: *Nucleic Acids Research* 35, W193–W200. DOI: 10.1093/nar/gkm226.

Rentzsch, Philipp et al. (Jan. 2019). “CADD: predicting the deleteriousness of variants throughout the human genome”. In: *Nucleic Acids Research* 47.D1, pp. D886–D894. DOI: 10.1093/nar/gky1016.

Resnik, Philip (Nov. 1995). *Using Information Content to Evaluate Semantic Similarity in a Taxonomy*. en. arXiv:cmp-lg/9511007.

Risch, Neil et al. (Sept. 1996). “The Future of Genetic Studies of Complex Human Diseases”. In: *Science* 273.5281. Publisher: American Association for the Advancement of Science, pp. 1516–1517. DOI: 10.1126/science.273.5281.1516.

Ritchie, Marylyn D. (2015). “Finding the Epistasis Needles in the Genome-Wide Haystack”. en. In: *Epistasis: Methods and Protocols.* Ed. by Jason H. Moore et al. Methods in Molecular Biology. New York, NY: Springer, pp. 19–33. ISBN: 978-1-4939-2155-3. DOI: 10.1007/978-1-4939-2155-3_2.

Ritchie, Marylyn D. et al. (July 2001). “Multifactor-Dimensionality Reduction Reveals High-Order Interactions among Estrogen-Metabolism Genes in Sporadic Breast Cancer”. en. In: *The American Journal of Human Genetics* 69.1, pp. 138–147. DOI: 10.1086/321276.

Rohlfs, Rori V. et al. (May 2010). “Detecting Coevolution through Allelic Association between Physically Unlinked Loci”. en. In: *The American Journal of Human Genetics* 86.5, pp. 674–685. DOI: 10.1016/j.ajhg.2010.03.001.

Roselli, Carolina et al. (June 2020). “Genetics of Atrial Fibrillation in 2020”. In: *Circulation Research* 127.1. Publisher: American Heart Association, pp. 21–33. DOI: 10.1161/CIRCRESAHA.120.316575.

Rubinacci, Simone et al. (Nov. 2020). “Genotype imputation using the Positional Burrows Wheeler Transform”. en. In: *PLOS Genetics* 16.11. Publisher: Public Library of Science, e1009049. DOI: 10.1371/journal.pgen.1009049.

Ruiz-Narváez, Edward A. et al. (Feb. 2004). “Transmission disequilibrium test (TDT) for case–control studies”. en. In: *European Journal of Human Genetics* 12.2. Number: 2 Publisher: Nature Publishing Group, pp. 105–114. DOI: 10.1038/sj.ejhg.5201099.

Russ, Dominic et al. (Feb. 2022). “Evaluating the detection ability of a range of epistasis detection methods on simulated data for pure and impure epistatic models”. en. In: *PLOS ONE* 17.2. Publisher: Public Library of Science, e0263390. DOI: 10.1371/journal.pone.0263390.

Russell, Peter J. (1998). *Genetics*. en. Google-Books-ID: yaDuAAAAMAAJ. Benjamin/Cummings. ISBN: 978-0-321-00038-5.

Saiki, Randall K. et al. (Dec. 1985). “Enzymatic Amplification of $\beta-GlobinGenomicSequencesandRestri$

In: *Science* 230.4732. Publisher: American Association for the Advancement of Science, pp. 1350–1354. DOI: 10.1126/science.2999980.

Salanti, Georgia et al. (July 2005). “Hardy–Weinberg equilibrium in genetic association studies: an empirical evaluation of reporting, deviations, and power”. In: *European Journal of Human Genetics* 13.7. Number: 7 Publisher: Nature Publishing Group, pp. 840–848. DOI: 10.1038/sj.ejhg.5201410.

Samaras, Patroklos et al. (Jan. 2020). “ProteomicsDB: a multi-omics and multi-organism resource for life science research”. In: *Nucleic Acids Research* 48.D1, pp. D1153–D1163. DOI: 10.1093/nar/gkz974.

Sanderson, Eleanor et al. (Feb. 2022). "Mendelian randomization". en. In: *Nature Reviews Methods Primers* 2.1. Number: 1 Publisher: Nature Publishing Group, pp. 1–21. DOI: 10.1038/s43586-021-00092-5.

Schaid, Daniel J. et al. (Aug. 2018). "From genome-wide associations to candidate causal variants by statistical fine-mapping". In: *Nature reviews. Genetics* 19.8, pp. 491–504. DOI: 10.1038/s41576-018-0016-z.

Schork, N. J. (Oct. 1997). "Genetics of complex disease: approaches, problems, and solutions". eng. In: *American Journal of Respiratory and Critical Care Medicine* 156.4 Pt 2, S103–109. DOI: 10.1164/ajrccm.156.4.12-tac-5.

Schriml, Lynn Marie et al. (Jan. 2012). "Disease Ontology: a backbone for disease semantic integration". In: *Nucleic Acids Research* 40.D1, pp. D940–D946. DOI: 10.1093/nar/gkr972.

Schuurman, Nadine et al. (Mar. 2008). "Ontologies for Bioinformatics". In: *Bioinformatics and Biology Insights* 2, pp. 187–200.

Sham, Pak C. et al. (May 2014). "Statistical power and significance testing in large-scale genetic studies". en. In: *Nature Reviews Genetics* 15.5. Number: 5 Publisher: Nature Publishing Group, pp. 335–346. DOI: 10.1038/nrg3706.

Shang, J. et al. (2019). "A Review of Ant Colony Optimization Based Methods for Detecting Epistatic Interactions". In: *IEEE Access* 7. Conference Name: IEEE Access, pp. 13497–13509. DOI: 10.1109/ACCESS.2019.2894676.

Shang, Junliang et al. (May 2016). "CINOEDV: a co-information based method for detecting and visualizing n-order epistatic interactions". In: *BMC Bioinformatics* 17.1, p. 214. DOI: 10.1186/s12859-016-1076-8.

Sherry, S. T. et al. (Jan. 2001). "dbSNP: the NCBI database of genetic variation". In: *Nucleic Acids Research* 29.1, pp. 308–311. DOI: 10.1093/nar/29.1.308.

Shu, Le et al. (May 2018). "Translating GWAS Findings to Novel Therapeutic Targets for Coronary Artery Disease". In: *Frontiers in Cardiovascular Medicine* 5, p. 56. DOI: 10.3389/fcvm.2018.00056.

Sillanpää, M. J. (Apr. 2011). “Overview of techniques to account for confounding due to population stratification and cryptic relatedness in genomic data association analyses”. en. In: *Heredity* 106.4. Number: 4 Publisher: Nature Publishing Group, pp. 511–519. DOI: 10.1038/hdy.2010.91.

Skipper, Robert A. (2009). “Revisiting the Fisher-Wright Controversy”. In: *Transactions of the American Philosophical Society* 99.1. Publisher: [American Philosophical Society, American Philosophical Association], pp. 299–322.

Slatko B. E. Gardner, A. F. et al. (Apr. 2018). “Overview of Next-Generation Sequencing Technologies”. In: *Current Protocols in Molecular Biology* 122. DOI: 10.1002/cpmb.59.

Smedley, Damian et al. (Jan. 2009). “BioMart – biological queries made easy”. In: *BMC Genomics* 10.1, p. 22. DOI: 10.1186/1471-2164-10-22.

Solomon, Joel S. et al. (Feb. 2000). “Evaluation and Treatment of BRCA-Positive Patients”. en-US. In: *Plastic and Reconstructive Surgery* 105.2, pp. 714–719.

Sonesson, Anna K. et al. (2005). “Kinship, Relationship and Inbreeding”. en. In: *Selection and Breeding Programs in Aquaculture*. Ed. by Trygve Gjedrem. Dordrecht: Springer Netherlands, pp. 73–87. ISBN: 978-1-4020-3342-1. DOI: 10.1007/1-4020-3342-7_6.

Stanislas, Virginie et al. (Jan. 2017). “Eigen-Epistasis for detecting gene-gene interactions”. In: *BMC Bioinformatics* 18.1, p. 54. DOI: 10.1186/s12859-017-1488-0.

Stoupa, Athanasia et al. (Mar. 2021). “New genetics in congenital hypothyroidism”. en. In: *Endocrine* 71.3, pp. 696–705. DOI: 10.1007/s12020-021-02646-9.

Sturtevant, A. H. (1913). “The linear arrangement of six sex-linked factors in Drosophila, as shown by their mode of association”. en. In: *Journal of Experimental Zoology* 14.1. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/jez.1400140104, pp. 43–59. DOI: 10.1002/jez.1400140104.

Subramanian, Aravind et al. (Oct. 2005). “Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles”. EN. In: *Proceedings of the National Academy of Sciences* 102.43. Company: National Academy of Sciences Distributor: National Academy of Sciences Institution: National Academy of Sciences Label: National Academy of Sciences Publisher: Proceedings of the National Academy of Sciences, pp. 15545–15550. DOI: 10.1073/pnas.0506580102.

Sun, Rui et al. (Dec. 2019). “wtest: an integrated R package for genetic epistasis testing”. In: *BMC Medical Genomics* 12.9, p. 180. DOI: 10.1186/s12920-019-0638-9.

Sun, Yingxia et al. (July 2017). “epiACO - a method for identifying epistasis based on ant Colony optimization algorithm”. en. In: *BioData Mining* 10.1, p. 23. DOI: 10.1186/s13040-017-0143-7.

Sunkin, Susan M. et al. (Jan. 2013). “Allen Brain Atlas: an integrated spatio-temporal portal for exploring the central nervous system”. In: *Nucleic Acids Research* 41.D1, pp. D996–D1008. DOI: 10.1093/nar/gks1042.

Sutton, Walter S. (1903). “The Chromosomes in Heredity”. In: *Biological Bulletin* 4.5. Publisher: Marine Biological Laboratory, pp. 231–251. DOI: 10.2307/1535741.

Svejgaard, Arne et al. (1994). “HLA and disease associations: Detecting the strongest association”. en. In: *Tissue Antigens* 43.1. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1399-0039.1994.tb02291.x, pp. 18–27. DOI: 10.1111/j.1399-0039.1994.tb02291.x.

Szathmáry, Eörs et al. (May 2001). “Can Genes Explain Biological Complexity?” In: *Science* 292.5520. Publisher: American Association for the Advancement of Science, pp. 1315–1316. DOI: 10.1126/science.1060852.

Szklarczyk, Damian et al. (Jan. 2019). “STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets”. In: *Nucleic Acids Research* 47.D1, pp. D607–D613. DOI: 10.1093/nar/gky1131.

Szklarczyk, Damian et al. (Jan. 2021). “The STRING database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets”. In: *Nucleic Acids Research* 49.D1, pp. D605–D612. DOI: 10.1093/nar/gkaa1074.

Theis, Jeanne L et al. (Oct. 2020). “Patient-specific genomics and cross-species functional analysis implicate LRP2 in hypoplastic left heart syndrome”. In: *eLife* 9. Ed. by Didier YR

Stainier et al. Publisher: eLife Sciences Publications, Ltd, e59554. DOI: 10.7554/eLife.59554.

Theunissen, Frances et al. (2020). “Structural Variants May Be a Source of Missing Heritability in sALS”. In: *Frontiers in Neuroscience* 14.

Thomas, Duncan C. et al. (Sept. 2005). “Recent Developments in Genomewide Association Scans: A Workshop Summary and Review”. In: *American Journal of Human Genetics* 77.3, pp. 337–345.

Thompson, E. A. (1974). “Gene Identities and Multiple Relationships”. In: *Biometrics* 30.4. Publisher: [Wiley, International Biometric Society], pp. 667–680. DOI: 10.2307/2529231.

— (1975). “The estimation of pairwise relationships”. en. In: *Annals of Human Genetics* 39.2. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1469-1809.1975.tb00120.x, pp. 173–188. DOI: 10.1111/j.1469-1809.1975.tb00120.x.

— (May 2008). “The IBD process along four chromosomes”. In: *Theoretical population biology* 73.3, pp. 369–373. DOI: 10.1016/j.tpb.2007.11.011.

Torkamani, Ali et al. (Sept. 2018). “The personal and clinical utility of polygenic risk scores”. en. In: *Nature Reviews Genetics* 19.9. Number: 9 Publisher: Nature Publishing Group, pp. 581–590. DOI: 10.1038/s41576-018-0018-x.

Trerotola, Marco et al. (July 2015). “Epigenetic inheritance and the missing heritability”. In: *Human Genomics* 9.1, p. 17. DOI: 10.1186/s40246-015-0041-3.

Tschermak, Erich (1900). *Ueber künstliche Kreuzung bei Pisum sativum*. de. Google-Books-ID: XC8bAAAAYAAJ. E. Tschermak.

Uchida, Katsuya et al. (2021). “Congenital Hypothyroidism and Brain Development: Association With Other Psychiatric Disorders”. In: *Frontiers in Neuroscience* 15.

Ueki, Masao et al. (Apr. 2012). “Improved Statistics for Genome-Wide Interaction Analysis”. en. In: *PLOS Genetics* 8.4. Publisher: Public Library of Science, e1002625. DOI: 10.1371/journal.pgen.1002625.

Uffelmann, Emil et al. (Aug. 2021). "Genome-wide association studies". en. In: *Nature Reviews Methods Primers* 1.1. Number: 1 Publisher: Nature Publishing Group, pp. 1–21. DOI: 10.1038/s43586-021-00056-9.

Urbanowicz, Ryan J (n.d.). *UrbsLab/GAMETES*. en.

Urbanowicz, Ryan J et al. (Sept. 2012a). "Predicting the difficulty of pure, strict, epistatic models: metrics for simulated model selection". In: *BioData Mining* 5, p. 15. DOI: 10.1186/1756-0381-5-15.

Urbanowicz, Ryan J. et al. (Oct. 2012b). "GAMETES: a fast, direct algorithm for generating pure, strict, epistatic models with random architectures". In: *BioData Mining* 5.1, p. 16. DOI: 10.1186/1756-0381-5-16.

Urbanowicz, Ryan J. et al. (Sept. 2018). "Benchmarking relief-based feature selection methods for bioinformatics data mining". en. In: *Journal of Biomedical Informatics* 85, pp. 168–188. DOI: 10.1016/j.jbi.2018.07.015.

Van Hout, Cristopher V. et al. (Oct. 2020). "Exome sequencing and characterization of 49,960 individuals in the UK Biobank". en. In: *Nature* 586.7831. Number: 7831 Publisher: Nature Publishing Group, pp. 749–756. DOI: 10.1038/s41586-020-2853-0.

Vasilevska, M et al. (Dec. 2013). "The Incidence and Type of Chromosomal Translocations from Prenatal Diagnosis of 3800 Patients in the Republic of Macedonia". In: *Balkan Journal of Medical Genetics : BJMG* 16.2, pp. 23–28. DOI: 10.2478/bjmg-2013-0027.

Venter, J. Craig et al. (Feb. 2001). "The Sequence of the Human Genome". In: *Science* 291.5507. Publisher: American Association for the Advancement of Science, pp. 1304–1351. DOI: 10.1126/science.1058040.

Voight, Benjamin F et al. (Sept. 2005). "Confounding from Cryptic Relatedness in Case-Control Association Studies". In: *PLoS Genetics* 1.3, e32. DOI: 10.1371/journal.pgen.0010032.

Vuillaumier-Barrot, Sandrine et al. (Dec. 2012). "Identification of Mutations in TMEM5 and ISPD as a Cause of Severe Cobblestone Lissencephaly". In: *American Journal of Human Genetics* 91.6, pp. 1135–1143. DOI: 10.1016/j.ajhg.2012.10.009.

Wallace, R. B. et al. (Aug. 1979). “Hybridization of synthetic oligodeoxyribonucleotides to phi chi 174 DNA: the effect of single base pair mismatch”. eng. In: *Nucleic Acids Research* 6.11, pp. 3543–3557. DOI: 10.1093/nar/6.11.3543.

Wan, Xiang et al. (Sept. 2010a). “BOOST: A Fast Approach to Detecting Gene-Gene Interactions in Genome-wide Case-Control Studies”. en. In: *The American Journal of Human Genetics* 87.3, pp. 325–340. DOI: 10.1016/j.ajhg.2010.07.021.

Wan, Xiang et al. (Jan. 2010b). “Predictive rule inference for epistatic interaction detection in genome-wide association studies”. eng. In: *Bioinformatics (Oxford, England)* 26.1, pp. 30–37. DOI: 10.1093/bioinformatics/btp622.

Wang, James Z. et al. (May 2007). “A new method to measure the semantic similarity of GO terms”. In: *Bioinformatics* 23.10, pp. 1274–1281. DOI: 10.1093/bioinformatics/btm087.

Wang, Kai et al. (Sept. 2010). “ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data”. In: *Nucleic Acids Research* 38.16, e164. DOI: 10.1093/nar/gkq603.

Wang, William Y. S. et al. (Feb. 2005). “Genome-wide association studies: theoretical and practical concerns”. en. In: *Nature Reviews Genetics* 6.2. Number: 2 Publisher: Nature Publishing Group, pp. 109–118. DOI: 10.1038/nrg1522.

Wang, Yupeng et al. (Apr. 2010). “AntEpiSeeker: detecting epistatic interactions for case-control studies using a two-stage ant colony optimization algorithm”. en. In: *BMC Research Notes* 3.1, p. 117. DOI: 10.1186/1756-0500-3-117.

Wang, Zhen et al. (2001). “SNPs, protein structure, and disease”. en. In: *Human Mutation* 17.4. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/humu.22, pp. 263–270. DOI: 10.1002/humu.22.

Wang, Zhuo et al. (Nov. 2004). “A Brief Review of Computational Gene Prediction Methods”. en. In: *Genomics, Proteomics & Bioinformatics* 2.4, pp. 216–221. DOI: 10.1016/S1672-0229(04)02028-5.

Weale, Michael E. (2010). “Quality Control for Genome-Wide Association Studies”. en. In: *Genetic Variation: Methods and Protocols*. Ed. by Michael R. Barnes et al. Methods in Molecular Biology. Totowa, NJ: Humana Press, pp. 341–372. ISBN: 978-1-60327-367-1. DOI: 10.1007/978-1-60327-367-1_19.

Weiling, Franz (1991). “Historical study: Johann Gregor Mendel 1822–1884”. de. In: *American Journal of Medical Genetics* 40.1. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/ajmg.132 pp. 1–25. DOI: 10.1002/ajmg.1320400103.

Weinberg, Wilhelm (1908). *Über den Nachweis der Vererbung beim Menschen*. ger. [Place of publication not identified] : [publisher not identified].

Weir, B. S. et al. (1984). “Estimating F-Statistics for the Analysis of Population Structure”. In: *Evolution* 38.6. Publisher: [Society for the Study of Evolution, Wiley], pp. 1358–1370. DOI: 10.2307/2408641.

WHO (2003). *ICD-10 : international statistical classification of diseases and related health problems : tenth revision, 2nd ed.* eng. WHO.

Wigginton, Janis E. et al. (May 2005). “A Note on Exact Tests of Hardy-Weinberg Equilibrium”. In: *American Journal of Human Genetics* 76.5, pp. 887–893.

Winzer, T. et al. (2006). “P66 The UK Blood Service/Wellcome Trust Control Collection: a unique public resource of control samples for disease association studies”. en. In: *Transfusion Medicine* 16.s1. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1365-3148.2006.00694_6 pp. 53–53. DOI: 10.1111/j.1365-3148.2006.00694_66.x.

Wood, Andrew R. et al. (Nov. 2014). “Defining the role of common variation in the genomic and biological architecture of adult human height”. en. In: *Nature Genetics* 46.11. Number: 11 Publisher: Nature Publishing Group, pp. 1173–1186. DOI: 10.1038/ng.3097.

Wray, Naomi R. et al. (Jan. 2021). “From Basic Science to Clinical Application of Polygenic Risk Scores: A Primer”. In: *JAMA Psychiatry* 78.1, pp. 101–109. DOI: 10.1001/jamapsychiatry.2020.3049.

Wright, Sewall (June 1920). “The Relative Importance of Heredity and Environment in Determining the Piebald Pattern of Guinea-Pigs”. In: *Proceedings of the National Academy of Sciences* 6.6. Publisher: Proceedings of the National Academy of Sciences, pp. 320–332. DOI: 10.1073/pnas.6.6.320.

— (1922). “Coefficients of Inbreeding and Relationship”. In: *The American Naturalist* 56.645. Publisher: [University of Chicago Press, American Society of Naturalists], pp. 330–338.

— (1932). “The roles of mutation, inbreeding, crossbreeding, and selection in evolution”. In: *Proceedings of the Sixth International Congress of Genetics*, pp. 356–366.

— (Aug. 1950). “Genetical Structure of Populations”. en. In: *Nature* 166.4215. Number: 4215 Publisher: Nature Publishing Group, pp. 247–249. DOI: 10.1038/166247a0.

Wu, P. et al. (Oct. 2019). “Mapping ICD10 and ICD10CM Codes to Phecodes Workflow Development and Initial Evaluation”. In: *JMIR Medical Informatics* 7. DOI: https://doi.org/10.2196/14325.

Wu, Y. et al. (2007). “Detecting proteinprotein interactions by far western blotting”. In: *Nature Protocols* 2. DOI: https://doi.org/10.1038/nprot.2007.459.

Wu, Yongxin et al. (Aug. 2020). “Interleukin-22 is elevated in the atrium and plasma of patients with atrial fibrillation and increases collagen synthesis in transforming growth factor-1-treated cardiac fibroblasts via the JNK pathway”. In: *Experimental and Therapeutic Medicine* 20.2, pp. 1012–1020. DOI: 10.3892/etm.2020.8778.

Xu, Jing et al. (Mar. 2013). “Intracellular lactate signaling cascade in atrial remodeling of mitral valvular patients with atrial fibrillation”. In: *Journal of Cardiothoracic Surgery* 8, p. 34. DOI: 10.1186/1749-8090-8-34.

Yang, Cheng-Hong et al. (Jan. 2020). “An improved fuzzy set-based multifactor dimensionality reduction for detecting epistasis”. en. In: *Artificial Intelligence in Medicine* 102, p. 101768. DOI: 10.1016/j.artmed.2019.101768.

Yang, Jian et al. (July 2010). “Common SNPs explain a large proportion of the heritability for human height”. en. In: *Nature Genetics* 42.7. Number: 7 Publisher: Nature Publishing Group, pp. 565–569. DOI: 10.1038/ng.608.

Yin, Wen et al. (May 2018). “Learning Opportunities for Drug Repositioning via GWAS and PheWAS Findings”. In: *AMIA Summits on Translational Science Proceedings* 2018, pp. 237–246.

You, Qian et al. (2018). “Development and Applications of a High Throughput Genotyping Tool for Polyploid Crops: Single Nucleotide Polymorphism (SNP) Array”. In: *Frontiers in Plant Science* 9.

Young, Alexander I. (June 2019). “Solving the missing heritability problem”. en. In: *PLOS Genetics* 15.6. Publisher: Public Library of Science, e1008222. DOI: 10.1371/journal.pgen.1008222.

Zeggini, Eleftheria et al. (Feb. 2009). “Meta-analysis in genome-wide association studies”. In: *Pharmacogenomics* 10.2, pp. 191–201. DOI: 10.2217/14622416.10.2.191.

Zhou, X. et al. (May 2012). “Genome-wide Efficient Mixed Model Analysis for Association Studies”. In: *Nature Genetics* 44. DOI: https://doi.org/10.1038/ng.2310.

Zhou, Xiang et al. (June 2012). “Genome-wide Efficient Mixed Model Analysis for Association Studies”. In: *Nature genetics* 44.7, pp. 821–824. DOI: 10.1038/ng.2310.

Zhou, Xiaoxu et al. (2020). “Evidence for Inflammation as a Driver of Atrial Fibrillation”. In: *Frontiers in Cardiovascular Medicine* 7.

Zoni-Berisso, Massimo et al. (June 2014). “Epidemiology of atrial fibrillation: European perspective”. In: *Clinical Epidemiology* 6, pp. 213–220. DOI: 10.2147/CLEP.S47385.

Zuk, Or et al. (Jan. 2012). “The mystery of missing heritability: Genetic interactions create phantom heritability”. EN. In: *Proceedings of the National Academy of Sciences* 109.4. Company: National Academy of Sciences Distributor: National Academy of Sciences Institution: National Academy of Sciences Label: National Academy of Sciences Publisher:

Proceedings of the National Academy of Sciences, pp. 1193–1198. DOI: 10.1073/pnas.1119675109.