UNIVERSITY OF OXFORD

Analysis of Epistasis in Human Complex Traits

LINO ANDRÉ FONSECA FERREIRA

Jesus College

DOCTOR OF PHILOSOPHY

2023

To my parents

Acknowledgements

The work presented in this thesis would not have been possible without the kindness and support of many people.

First, I am greatly indebted to my primary supervisor, Simon Myers, for his guidance throughout this research project and for the kindness he personally showed me at challenging moments during this degree. His enthusiasm and passion for science, and his commitment to scientific rigour, have been an inspiration. I am grateful to have been able to complete this defining period of my training under his wing.

I am very grateful also to Pier Palamara, my second supervisor, for having taken me as a rotation student in my first year and for being a supportive and friendly presence throughout my degree.

I thank Yang Luo for her help with my Confirmation of Status and for suggesting and co-supervising an exciting extension to my project which is included in this thesis as Chapter 7. On a similar note, I am grateful to Robbie Davies for his help with the major milestones of my degree and to Augustine Kong for being an assessor for my Transfer of Status.

My time as a DPhil student was enriched by being part of the Myers Group. I wish to thank, in particular, Ipsita Agarwal and Ana Ignatieva for their friendship and mentoring, and Sile Hu for his help in the beginning of my project.

I am fortunate to have received generous funding from the Wellcome Trust [222334/Z/21/Z] through the Genomic Medicine and Statistics DPhil programme. I am grateful to the team behind this programme, especially Julian Knight and Isabel Schmidt, for their support, and to my wonderful cohort (Árni Gunnarsson, Jenni Herrmann, Minnie Salmon and Samvida Venkatesh) for their friendship.

I thank the Oxford Biomedical Research Computing facility, which enabled the computational data analyses described in this thesis (these were supported by the Wellcome Trust [203141/Z/16/Z] with additional support from the NIHR Oxford BRC).

I am grateful to the Department of Statistics for giving me an excellent workspace and to my office colleagues, especially Leo Klarner, for their companionship.

As a last institutional acknowledgement, I would like to thank Jesus College for providing a friendly community and for the generous award of a Graduate Scholarship in my final year. I am grateful to my college advisors, especially Brieuc Lehmann, for their support.

On a more personal note, I am very grateful to St Ebbe's Church, a wonderful Christian community and a home away from home throughout my time in Oxford. I especially benefited from the ministries of Vaughan Roberts, Glenn Nesbitt and Nancy-Page Lowenfield, and from the friendship and support of Christopher Morton.

I would like to thank the friends with whom I was fortunate to walk alongside during these past four years. I am particularly grateful to Jeremiah Lau for his companionship and encouragement, as well as to several others including, in alphabetical order, Hugh Bamford, Ollie and Georgina Bartlett, David Bernabé, Grace Borchert, Pablo Brubeck, Joseph Chu, Steven Firmin, Kirolous Hanna, Lacey Helmuth, Eden Hildreth, Samuel Lam, Julia Loh, Elizabeth Nurse, Josh Smailes, Kay Song, Hosana Tagomori, Jeffrey Tse and Jenny Wang.

My parents, Ilda and Lino Ferreira, have been a constant source of love and support throughout my life, and have always encouraged me to pursue my dreams – even at personal cost, as this has meant that I have lived far from home for several years. I am deeply grateful for their kindness and love.

Lastly, during my time as a student at this university, I have remembered and valued its roots as a Christian institution – its official motto, 'Dominus illuminatio mea', is taken from Psalm 27, which begins, 'The LORD is my light and my salvation'. In my studies and throughout my life, I thank God for his gracious and abundant provision and wish to praise him in everything I do, including this research.

– LAFF
Oxford, October 2023

Contents

Abstract 1

1 Introduction 3
- 1.1 Definitions of epistasis 7
- 1.2 Motivation 10
- 1.3 Methods to detect statistical epistasis 12
- 1.4 A new approach to testing for epistasis 18
- 1.5 Overview of thesis 24

2 Data 25
- 2.1 Phenotype selection 25
- 2.2 Covariates 28
- 2.3 Sample selection 29
- 2.4 Variant quality control and filtering 31
- 2.5 Functional annotations 32

3 Methods 33
- 3.1 A linear model of a quantitative phenotype 33
- 3.2 The naive approach to interaction testing 35
- 3.3 Testing for SNP × PGS interactions 35
- 3.4 Bias and power considerations 37
- 3.5 Avoiding false positives while increasing power 52
- 3.6 Data analysis pipeline 65

4 Simulations 75
- 4.1 Null simulations 77
- 4.2 Simulations with interactions 90
- 4.3 Summary 95

5 Results: alkaline phosphatase 98
- 5.1 Background 98
- 5.2 SNP × PGS interaction GWAS 106

5.3 Pairwise interaction GWAS 114
5.4 Summary 123

6 Results: remaining traits 130
6.1 Distribution of LHS hits across traits 131
6.2 Comparison of associations in the original and interaction GWAS 131
6.3 LHS hits are enriched for functional relevance 133
6.4 SNP × PGS interactions account for little heritability 137
6.5 Distribution of pairwise hits across traits 138
6.6 Highlights of pairwise interactions 140
6.7 Summary 146

7 Interactions involving *HLA* genes 148
7.1 Background 148
7.2 Imputing *HLA* alleles for the UKB 154
7.3 Results of interaction testing 156
7.4 Summary and future work 159

8 Conclusion 162
8.1 Summary 162
8.2 Limitations 165
8.3 Discussion 167
8.4 Future work 169

References 171

Appendices 192

A Methods: proofs 192
A.1 Proof of Proposition 3.4.2 192
A.2 Proof of Proposition 3.4.3 194
A.3 Derivation of main result for Case 4 195

B Data: list of traits 198

C Simulations: supplementary results 201

D Alkaline phosphatase: supplementary results 213

E List of independent SNP × PGS interaction hits 215

F List of independent pairwise interaction hits 223

Abstract

Thousands of genetic mutations have been associated with many human complex traits and diseases, improving our understanding of the biological mechanisms underlying these phenotypes.

The great majority of genetic association studies have focused exclusively on the direct effects of single mutations, ignoring possible interactions (epistasis). However, since genes operate within complex networks, interactions are expected to exist. The modelling of epistasis could further biological understanding, but the detection of such effects is complicated by a vast search space.

In this thesis, we present a new statistical method to detect genetic interactions affecting quantitative traits in large-scale datasets. Our approach is based on testing for an interaction between a variant and a polygenic score (PGS) comprising a group of other mutations. We develop a new computational algorithm for PGS construction, and show through simulations that this method is robust to false-positives while retaining statistical power.

We apply our approach to 97 quantitative traits in the UK Biobank (UKB) and find 144 independent interactions with the PGS for 52 different traits, including important variants known to affect disease risk at the *APOE*, *FTO* and *LDLR* genes, for example.

We also develop a test to identify, for each variant interacting with the PGS, the variants driving that interaction. This recovers previously-known interactions and identifies several novel signals, primarily for biomarker traits. An example is a large network of genes (including *ABO*, *ASGR1*, *FUT2*, *FUT6*, *PIGC* and *TREH*) affecting alkaline phosphatase levels, or an interaction between *IL33* and *ALOX15* impacting eosinophil count, potentially implicated in asthma.

Lastly, we extend our analysis to a new dataset of imputed variation at *HLA* genes in the UKB and find, among others, a new interaction for glycated haemoglobin involving *HLA-DQA1*03:01*, an allele previously associated with diabetes.

Our results demonstrate the potential for detecting epistatic effects in presently-available genomic datasets. This can allow the uncovering of key 'core' genes modulating the impacts of other regions in the genome, as well as the identification of subgroups of interacting variants of likely functional relevance.

1 Introduction

Understanding the connection between genotype and phenotype is a central goal of genetics. In the past 25 years, rapid progress in DNA sequencing technology has enabled the generation of large datasets of human genetic sequences with associated phenotypes – a wide array of complex traits and diseases – which in turn have made possible the discovery of many genetic variants associated with these phenotypes.

A 2016 review of genome-wide association studies (GWASs) identified approximately 10,000 *independent*[1] associations between single-nucleotide polymorphisms (SNPs) and complex traits and diseases listed in the EMBL-EBI GWAS Catalog, a central repository of such associations [1, 2], as of September 2016. More recently, a separate article reported in 2019 that more than 3000 GWASs for over 1000 traits were represented in that repository [3]. As of 25 September 2023, 553,000 associations are listed in the Catalog. This great number of associations detected in little more than ten years[2] reflects the swift pace of sequencing technology development, and of data collection and analysis, in this period.

[1] Genetic variation in the human genome is organised into haplotype blocks. These are short regions – relatively to the size of the genome – within which genetic variants are highly correlated, a correlation termed linkage disequilibrium (LD). High LD within a region is caused by the absence of meiotic recombination between neighbouring variants, which means that these haplotypes tend to be transmitted as a single unit from parent to offspring. Because of this high correlation between variants in close physical proximity, human genetic associations generally appear as groups of neighbouring variants that are considered to constitute the same genetic signal. An important procedure in association testing is then the 'thinning' of these clusters into independent signals of association.

[2] The first GWAS, undertaken for age-related macular degeneration, the leading cause of blindness in the developed world, was published in 2005 [3–5].

Statistical studies of genetic association have contributed to our understanding of the genotype–phenotype map along three main dimensions. First, looking at the aggregate patterns of the associations discovered has provided insights into the *genetic architecture of complex traits*, i.e., the structure and statistical distribution of associated variants in the genome within and across traits. One robust finding in this regard is that complex traits are highly polygenic [1, 3], with the majority of traits having hundreds to thousands or tens of thousands of genetic associations [6]. A second observation is that the majority of variants identified through GWASs fall in non-coding regions of the genome [7], i.e., regions whose DNA is not directly translated into protein – this complicates the understanding of their biological function. Yet another conclusion is that many genetic variants are pleiotropic [1, 8], i.e., they have an effect on multiple phenotypes.

Beyond the statistical study of genetic architecture, and in spite of the challenges in discerning the function of most associated variants as mentioned in the preceding paragraph, GWASs have also enabled a *better understanding of specific functional mechanisms* underlying different complex traits and diseases [5, 9]. A prominent example of such a phenotype is Crohn's disease, whose connection with autophagy was only discovered after SNPs in the *ATG16L1* and *IRGM* genes (known to play a role in autophagy) were found to associate with this disease [5, 10]. For Crohn's disease but also for other autoimmune conditions such as psoriasis, ankylosing spondylitis and inflammatory bowel disease (IBD), the implication of the IL-23 pathway through statistical association led to the use of medications targeting this pathway in the treatment of these diseases [1, 5, 11].

Other examples of diseases whose mechanisms are now better understood because of genetic associations identified through GWASs include obesity and type 2 diabetes (T2D). For the former, the statistical association with the *FTO* gene – first reported

in 2007 [12–14] and widely replicated since then – led to the finding of a role for this gene in the regulation of food intake and less-established suggestions of an impact on processes related to the storage of fat [5, 14–16]. And for T2D, an association with *SLC16A11* led to further research that recognised its role in cellular fatty acid and lipid metabolism [5, 17] and to the observation that most disease-associated variants are linked to reduced insulin secretion rather than insulin action [1].

A third way in which statistical association testing has contributed to the general task of understanding the link from genotype to phenotype is the development of *risk prediction models for complex diseases* – as well as models to predict complex traits more generally – known as polygenic risk scores or simply polygenic scores (PGSs). PGSs are computed as weighted sums of relevant alleles at many sites along the genome, where the variants included and their weights can either be obtained from GWAS summary statistics or through specialised statistical methods for PGS construction.[3] The enthusiasm driving the development of PGSs is motivated by their potential for clinical application: an accurate model for predicting disease risk based on genetic factors could enable the prioritisation of patients at higher risk for clinical intervention [19–21], thus contributing to the realisation of one of the promises of 'personalised medicine', an idea which has received much attention in recent years.

The development in the last decade of several methods for building PGSs – some based only on summary statistics from GWASs, others taking advantage of individual-level genetic and phenotypic data[4] – has enabled the construction of reasonably well-performing scores for a variety of diseases. These include breast and prostate cancers [24, 25], coronary artery disease [26, 27], type 1 and 2 diabetes [20, 28, 29], Alzheimer's disease [30], atrial fibrillation or IBD [20]. PGSs have also been

[3] A recent article by Choi et al. [18] provides a protocol for building such scores, highlighting the different statistical issues involved.

[4] See Ni et al. [22] and Zhang et al. [23] for reviews of the different methods available.

developed to assist with discrimination between disease subtypes in diabetes [31, 32] and forecast the progression of Alzheimer's disease [33].

In spite of the rapid development of PGSs for different diseases, the adoption of these prediction models in clinical practice has been much slower.[5] Despite the popularity of PGSs in the academic literature, doubts have been raised regarding their clinical relevance [37, 38] and parts of the medical genetics community have urged caution in their application: for example, a very recent article by the American College of Medical Genetics and Genomics emphasises that limited evidence exists for the effective use of PGSs to improve clinical outcomes and advises against their general implementation [39]. It should also be noted that the predictive performance of PGSs for disease risk prediction is often limited: for instance, the best-performing scores in Mavaddat et al. [24] and Inouye et al. [27] explain only 45% and 27%, respectively, of the estimated SNP heritability of the phenotypes in question, indicating that most genetic factors underlying disease risk are still not covered by these models.[6]

Beyond the modest performance of most disease PGSs currently available, a major challenge to the future incorporation of these models into clinical care is their *lack of transferability across populations.* It has often been noted that the great majority of GWASs have been undertaken using data from European populations [41–44]: for example, Mills and Rahal [45] reported in 2020 that 88.5% of the individuals included in more than 4300 studies listed in the GWAS Catalog [2] were of European ancestry. This in turn has resulted in genetic associations and prediction models that generalise poorly when applied to individuals of non-European ancestries [46–

[5] One of the few examples of PGSs in use in the clinic is that of BOADICEA, a 313-variant PGS for breast cancer which is reported to have been accessed (via an online platform) in over 100 different countries [34–36].

[6] SNP heritability is defined as the proportion of phenotypic variation in a population that can in principle be explained by the combined *additive* effects of common SNPs [40], while broad-sense heritability is the proportion of variation that can be attributed to all the genetic factors impacting the phenotype (including non-linear effects, which we will discuss below).

48]. A pressing challenge preventing the clinical use of PGSs is therefore that this would exacerbate existing inequalities in access to healthcare both within and across countries. This problem has received much attention in the literature in recent years, and multiple methods have been developed that seek to improve the cross-population transferability of PGSs.[7] However, the improvement achieved has been limited and transferability therefore remains an open problem. We return to this issue below when motivating the project described in this thesis.

The great progress in statistical association testing described above – as well as the risk prediction models based on such testing – has been achieved by focusing almost exclusively on estimating the direct or additive effects of genetic variants on phenotypes. A SNP is tested for association with a phenotype by comparing simple statistics for that phenotype – usually the average for a quantitative phenotype and the odds ratio for a disease – between individuals carrying zero, one or two copies of an allele at that locus.

In contrast, comparatively little attention has been paid to the effect of *combinations of* or *interactions between* two (or more) mutations on a phenotype, a phenomenon often referred to as *epistasis*. Different definitions of epistasis have been used in the literature since the term was first coined. Therefore, and before considering the question of why its study may be relevant, we first review its different possible meanings and define the sense in which it will be used in this thesis.

1.1 DEFINITIONS OF EPISTASIS

The term *epistasis*, which translates to 'standing upon', was first coined by William Bateson (1861–1926) in 1908 [50]. In his book, *Mendel's Principles of Heredity,*

[7] Kachuri et al. [49] provide a review of these methods and discuss the underlying causes of the lack of generalisability.

published in the following year [51], Bateson defines epistasis in the context of a dihybrid cross with two dominant alleles, i.e., a cross between two individuals both of which are heterozygous at the same two distinct sites and where one allele at each site is dominant over the alternative allele.[8]

It had been observed that some dihybrid crosses resulted in an unexpected set of phenotypes: instead of four different phenotypes being observed with a ratio of frequency of occurrence of 9:3:3:1,[9] some experiments had yielded only three observed phenotypes with ratio 9:4:3.[10] Bateson used the now classical example of mouse coat colours (grey, black and albino) and noted that the locus determining presence/absence of coat colour appeared to 'stand above' that determining colour: a mouse possessing two copies of the recessive allele responsible for albinism had no pigmentation regardless of their allele at the locus determining black or grey colour[11] – in other words, the allele determining black/grey colour can only be expressed if the individual is not homozygous for the recessive albinism allele.[12] The allele for colour is therefore said to be 'masked' (or blocked) by that for albinism, which suggests an ordering of the corresponding genes within a biological pathway; the blocked allele is called 'hypostatic' and the masking allele 'epistatic' [53].

A few years later, in 1918, R. A. Fisher (1890–1962) employed a similar term, *epistacy*, to refer to the more general phenomenon of a statistical deviation from the expected combined additive effect of two loci on a phenotype [53, 55].[13] The

[8] An allele is said to be dominant over another if the presence of a single copy of it leads to the same phenotypic effect as that of two copies of that allele.

[9] A dihybrid cross has 16 possible genotypes: if we denote the alleles at two loci by A/a and B/b, we obtain AABB, AABb, AAbB, AAbb, etc. Assuming A is dominant over a and B is dominant over b, we expect all nine genotypes with at least one A and B allele to have the same phenotype. Extending this reasoning leads to the ratio of 9:3:3:1.

[10] Bateson attributes the first remark on this difference to Cuénot in 1902 [52].

[11] See Phillips [53] for a modern biochemical explanation of this phenomenon.

[12] This type of epistasis has since been termed 'recessive epistasis' [54].

[13] Fisher presented this concept in the same seminal paper that reconciled Mendelian inheritance with the infinitesimal model and introduced the statistical concept of variance [50].

similarity between the two terms led to the original term, epistasis, being adopted by the population genetics community to refer to such statistical deviations, while it continued being used in classical genetics in its original sense [50].

Adding to this confusion, the term epistasis has also been used to refer simply to *functional* interactions between proteins in the same pathway, or between other biological components more generally [53]. We note that such 'interactions' are widespread in complex biological systems and that a functional interaction between the proteins produced from two different genetic loci need not imply an interaction of either type described above, and the converse is also true.

For clarity, we adopt here a comprehensive classification of the different concepts of epistasis proposed by Patrick C. Phillips in a widely-cited review article [53]. Phillips distinguishes *functional, compositional* and *statistical* forms of epistasis. *Functional epistasis* refers to protein-protein interactions and other such functional relationships between biological components. *Compositional epistasis* means a blocking of an allele's effect by another allele in the original sense of Bateson. We highlight that compositional epistasis assumes variation only at the two loci in question, with an invariant genetic background – because of this, this concept is most useful for the analysis of artificially-bred as opposed to natural populations. Finally, *statistical epistasis* is defined in the sense of Fisher and does not assume a fixed genetic background but is instead a statistical estimate that is averaged over the genetic backgrounds observed in the sample being examined.

This thesis concerns statistical epistasis: we will aim to detect pairs or larger groups of loci whose effect on a continuous phenotype requires the inclusion of an interaction term in a statistical model. Since estimation of such effects is generally done through linear regression, by construction these interactions are measured against an average of the genetic background at other loci in the genome.

1.2 MOTIVATION

Statistical epistasis, i.e., a deviation from pure additivity in the combined effect of two (or more) loci, is a concept with a rich history which can be seen – and indeed was originally introduced – as a natural component of a decomposition of quantitative phenotypic variation into additive, dominance and interactive effects. Yet despite its fundamental role in this canonical decomposition of heritability, such non-linear effects remain an under-explored component of genetic variation – at least when compared to their additive counterparts – and their role in the genetic architecture of complex traits has been an object of controversy [56].

Why has epistasis not received more attention in the many association studies undertaken in the past decade? A reason that has sometimes been proposed to justify the importance of research on epistasis in the past two decades is its potential to account for part of the currently-unexplained heritability of complex traits and diseases [56–59]. However, since the early days of GWAS, it has been shown in theory and subsequently verified empirically that the variance of complex traits is almost entirely additive [60–62].[14] Epistatic interactions are therefore unlikely to substantially increase the fraction of heritability explained by known genetic factors or improve the performance of PGSs.[15]

[14] This is partly because estimates of additive heritability will implicitly incorporate part of the variation driven by interactions if these exist [61].

[15] In contrast with results from human studies, research on epistasis in model organisms such as Drosophila or yeast has often produced evidence for an important role for interactions in the genetic architecture of a variety of traits [61, 63–65]. However, and as emphasised by Sackton and Hartl [61], the model organisms studied are part of artificial populations bred so as to achieve approximately-equal frequencies at all segregating loci, which is in stark contrast with natural populations in which most variants are rare [60]. Since the frequency of an interaction between two loci will be the product of their individual frequencies, the variance of an interaction will be largest for equally common variants [61] and therefore very small for pairs of mutations in which at least one is rare, as expected for humans. This discrepancy between frequency distributions in model organisms when compared to humans is thought to underlie the differences in the role of epistasis observed in model vs. natural populations [61].

A second motivation sometimes cited for the study of epistasis is the potential for the modelling of interactions to improve the cross-population transferability of PGSs [56], which as noted above is currently limited. The lack of PGS transferability is likely to be driven by multiple factors; some which have been proposed in the literature are differences in the allele frequencies and effect sizes of causal variants across populations as well as in patterns of linkage disequilibrium (LD) combined with difficulties in identifying the causal SNPs at many loci [49].

A difference in causal variant effect sizes between populations could be driven by epistasis: since an estimate of the additive effect of a genetic variant obtained through a linear regression model will also capture part of the epistatic effect from an interaction between this variant and another if such an effect exists, this estimated additive effect would depend on the allele frequency at the second interacting locus. And since differences in allele frequencies across populations are known to be common, such interactions could therefore lead to different estimated additive effect sizes in the two populations. However, recent research has found no evidence to support a widespread difference in effect sizes across populations [66, 67], making it unlikely that the modelling of interactions would directly lead to improved PGS transferability.

We have seen how two previously-proposed reasons for testing for and modelling epistasis are likely to be ill-founded. Is there then any reason to work on the study of interactions in the present GWAS era?

The primary motivation for continuing to investigate statistical epistasis is the same as that for research in statistical and population genetics more generally: to improve our understanding of human biology, regardless of whether this new knowledge can be directly applied to the prediction of disease risk.

While a *functional* epistatic relationship between proteins does not imply a statistical non-linearity between variation at their corresponding genes, the reverse

relationship is likely to often hold true: if the best model of the effect of two or more loci on a phenotype involves an interaction term, it seems probable that these loci play a role in a functional interaction in a pathway involved in the manifestation of the trait in question [53]. Detecting and examining epistatic interactions could therefore improve our understanding of various biological processes characterised by interdependencies between multiple components such as protein–protein [53, 68] and protein–DNA/RNA (physical) interactions [68], the functioning of regulatory pathways [53], genetic homeostasis [56], speciation [56, 69], genetic robustness [70] and the structure and evolution of complex biological systems more generally [53].

1.3 METHODS TO DETECT STATISTICAL EPISTASIS

Numerous methods have been proposed in the literature to identify epistatic interactions, and these have been the object of multiple reviews [71–74]. In this section, we do not attempt an exhaustive analysis of all existing methods but focus instead on providing a comprehensive overview of the different statistical strategies that have been used to address this question.

Our brief review below draws heavily on the excellent review article published by Wei et al. in 2014 [72] but extends it to include more recent methods as well as a few other approaches not originally covered in that article. Methods are grouped into five different categories corresponding to classes or types of statistical approaches; we note that these categories are not always mutually exclusive and that some methods could plausibly be listed under more than one category.

Before proceeding with our review, we would like to first highlight the central challenge to the detection of epistatic interactions. Consider a standard genomic dataset: this may include a few hundred thousand individuals with corresponding

DNA sequences which, after applying quality control procedures, may contain several million variants. Conceptually, the simplest approach to identifying pairwise interactions in such a dataset would be to exhaustively test for an epistatic effect on the phenotype for every possible pair of SNPs – if there are P SNPs, this would require $P(P-1)/2$ tests. For a modern dataset, however, this number of tests – in the order of P^2 – would likely be in the trillions, which reflects the enormously-vast search space of possible interactions.

The challenge presented by such a huge search space is two-fold. First, simply carrying out all these association tests is an expensive computational task – however, as we will see below, modern computing infrastructure and methods can allow for such an operation. An even greater obstacle that cannot be surmounted through more efficient computation is statistical in nature: setting aside the practical challenges of undertaking billions or trillions of tests, any approach that involves such a large number of statistical tests will be liable to generating a large number of false-positive findings; this would then require a strict procedure to keep the type-I error rate at a reasonable level, which would significantly diminish power to identify true interactions.

Methods for the detection of epistatic interactions must operate within this trade-off between the exhaustiveness with which they search the space of possible interactions and the power they have to identify true effects. The different categories of methods listed below take different approaches to this challenge, sometimes compromising power for exhaustiveness or vice-versa.

1.3.1 *Exhaustive search for pairwise interactions*

While exhaustive testing of pairwise interactions can be very expensive computationally, multiple methods exist that take advantage of modern computing infrastructure

and data structures to undertake a complete search for interacting pairs at reasonable cost.[16]

One strategy is simply to parallelise the testing over a cluster of CPUs, as these are currently available to most researchers working in computational biology [75]. A similar but more efficient approach is to instead parallelise operations over the cores of a GPU card, and several methods have been developed for this task [76–79]. Yet another alternative is to recode the genetic data into a data structure that allows for fast bitwise operations to be performed directly on it [80, 81]. Finally, one can use approximate test statistics that are simpler to compute when carrying out a genome-wide scan [80].

LD- and haplotype-based tests: The exhaustive approaches described above directly used the genotypes at a pair of loci to test for an interaction using standard statistical approaches – significance testing for the interaction term in a linear regression [75, 78], comparison of odds ratios for sample groups defined by combinations of genotypes [76], F-tests if interaction effects are allowed to vary between homozygous and heterozygous individuals at either site [77, 81] or likelihood-ratio tests and their approximation [79, 80]. An alternative is to make use of the following statistical implication of an interaction given the correlation structure of the genome: if two loci interact to increase the risk of a binary phenotype, then that interaction should be observed more frequently in cases than controls – this is equivalent to the correlation between the two loci (i.e., the LD between them) being higher in cases than controls, and at least two methods base their testing on this comparison (see refs. [82] and [83], implemented for large-scale analysis in ref. [84]).

[16] Upton et al. [74] provide a review of these approaches.

A related approach is to compare the joint occurrence of the two mutations on genomic segments inherited from the same parent, which therefore requires phasing of the data (i.e., the determination of the parent of origin for heterozygous mutations). Both kinds of approaches are reviewed in ref. [85].

1.3.2 *Two-stage approaches*

A possible solution to the statistical and computational challenges of exhaustive testing is to proceed in two stages: first, a filtering step is applied to select a moderately-sized subset of loci; then, pairwise interactions are tested only between loci in this set.[17] This approach can improve power by sacrificing exhaustiveness.

A widely-cited study on the genetics of psoriasis susceptibility that identified an interaction between a variant in the human leukocyte antigen (HLA) *HLA-C* gene and a mutation in *ERAP1*[18] employed this approach: first, a small set of distinct loci was selected based on previous literature and a new GWAS; then, interactions were tested only between SNPs in this set using a simple likelihood ratio test [87].

Two-stage methods have been compared with exhaustive approaches in simulated data [88, 89]. Perhaps surprisingly, these simulations showed that exhaustive testing can have greater power than a two-stage approach in some settings. It is therefore important that the initial filtering step used to select loci for interaction testing be optimised for finding loci for which epistatic effects are plausible.

While marginal association with the phenotype is a sensible criterium to be used in a first stage as in the psoriasis study above, some interacting SNPs may not show such direct signal. Alternatively, loci can be selected based on biological knowledge (for example, information retrieved from databases of gene pathways or protein–

[17] See ref. [86] for a review of some methods of this kind.
[18] We examine interactions involving *HLA* genes in Chapter 7.

protein interactions [72]) or through alternative computational approaches such as stochastic search [90].

Variance effect-based methods: Another way in which variants can be prioritised for downstream interaction testing is through statistical features other than their marginal association with the phenotype, and in particular by testing for an effect of individual markers on the phenotypic variance. It has been shown that an interaction between two SNPs leads to a difference in the phenotypic variance across the three groups defined by the allele count at either SNP [91, 92].[19] A possible strategy to prioritise SNPs for interaction testing is therefore to individually test them for an effect on the variance of the phenotype [93–95].

Testing for marginal epistasis: A related approach is to test for 'marginal epistasis', a concept introduced by Crawford et al. in 2017 [96] that refers to the combined effect of the interaction between one variant and *all* other variants in the genome. Crawford et al. developed a *variance-components model* for the estimation of marginal epistasis in quantitative traits, which was later extended to binary phenotypes [97]. More recently, a new algorithm for the estimation of a similar random effects model of marginal epistasis has been proposed by Fu et al. [98] that can scale to very large datasets, and which the authors apply to the UK Biobank (UKB). We consider these methods in more detail when comparing them with our approach below.

1.3.3 *Group-based approaches*

All the approaches discussed so far focus exclusively on individual SNPs or loci, which can either be exhaustively tested or pre-filtered in a two-stage approach. A natural

[19] The intuition behind this simple probabilistic result is that, if an interaction is present, there will be an additional term in the expression for the phenotypic variance for individuals with a mutation at one of the sites.

alternative given the size of the search space and the fact that individual variants can be grouped in various ways according to their biological properties is to use such groupings to reduce the number of interactions for which to test.

More specifically, variants can be grouped at the level of the genes in which they are located or that they are believed to affect – although not all variants can be linked to a unique gene in this way – and a test can then be carried out for an interaction between different genes [99]. Since there are many fewer genes than variants, this is a much more tractable problem. Beyond gene-associations, variants can be grouped based on other genetic features such as whether they fall into evolutionarily-conserved regions or haplotype blocks, or even in simple moving windows [100].

1.3.4 *A non-parametric statistical method: multifactor-dimensionality reduction*

A popular non-parametric method that can be used to identify potentially-interacting loci is multifactor-dimensionality reduction (MDR) [101]. This method does not directly test for an interaction but instead finds *combinations of factors* (genetic loci) – these can be pairs or of higher order – that influence a *binary trait,* where their effect on the trait need not be additive and can therefore include an interaction component [71]. It is then possible to investigate such factors for interactions using traditional approaches.

For pairwise interactions (i.e., no higher order terms), MDR works as follows: given a set of bi-allelic variants, enumerate all the possible two-variant combinations and identify the nine groups of samples resulting from the nine possible combinations of genotypes at each pair of loci. For each group, compute the odds ratio and use this information to define high- and low-risk groups that partition the sample in two. Through cross validation, find the best-performing pair of loci. These can then be tested for an interaction.

One of the appealing properties of this method is that it can non-parametrically test for higher order interactions by considering, e.g., all sets of three SNPs. A disadvantage is that the definition of high- and low-risk groups is arbitrary. And the major shortcoming of this approach is that it becomes very expensive for large numbers of SNPs, as it again requires exhaustive testing; to address this difficulty, implementations of MDR for GPUs have been developed [102, 103].

1.3.5 *Machine learning approaches*

Perhaps unsurprisingly given the high dimensionality of the space of possible interactions, multiple machine learning methods have been developed that attempt to navigate this search space efficiently to identify interactions. These include classical algorithms such as support vector machines [104], random forests [105–108] or ensemble learning/boosting [109], as well as Gaussian process regression [110] and kernel-based methods [111]. Several methods based on information-theoretic metrics have also been proposed for this problem [112–117]. Finally, other approaches such as reconstructability analysis [118] and ant colony optimisation [119, 120] have also been used.

1.4 A NEW APPROACH TO TESTING FOR EPISTASIS

In this thesis, we propose a new approach to the statistical testing of epistatic effects. This approach consists of testing, within a linear regression framework, for an interaction between each SNP and a PGS.

The PGS is constructed so as to approximately capture all the additive genetic signal for a quantitative phenotype. By this we mean that, if this score is regressed out from the phenotype – together with the usual covariates measuring demographics and ancestry – and standard testing of association is carried out at each of the SNPs

included in the PGS, none of these SNPs remain significantly associated with the phenotype. As we will explain in detail in Chapter 3, guaranteeing that the additive signal is exhausted in this way improves statistical power and reduces the likelihood of false-positive findings.

Once we have obtained a PGS in this manner, we carry out a GWAS for the phenotype in question which includes, for each SNP, a first term for that SNP's direct effect on the trait, a second term for the effect of the PGS and, crucially, a statistical interaction term between the SNP and the PGS (we regress covariates out from the phenotype in advance so as to minimise computational cost). We then analyse the results of significance testing for this third term to identify loci that interact with the PGS in this setting. Finally, the set of independent SNPs that interact with the PGS can be tested for pairwise interaction with other SNPs either genome-wide or only with those included in the PGS – in this way, we are able to detect some pairwise interactions without the computational and multiple-testing burden of an exhaustive test.

Relation to existing approaches: The method we propose is related to a few of the existing strategies for the testing of epistasis that we described in the previous section. First, it can be regarded as a two-stage approach, since the loci found to interact with the PGS can be subject to downstream pairwise testing; however, unlike other two-stage approaches whose first stage serves only as a basis for the final, second stage, interactions between an individual variant and a PGS can be meaningful and interesting in themselves, as we explain below.

Among the different two-stage strategies covered in Section 1.3.2, our approach is most similar to the recently-developed concept of testing for marginal epistasis, but differs from it in significant ways. First, marginal epistasis refers to the effect of a 'lead'

variant on *all other variants* in the genome; by contrast, we test for interactions only with a specific set of variants with a *direct effect on the trait* – reflecting functional importance – which are combined into a PGS. Second, the random effects models estimated by Crawford et al. [96] and Fu et al. [98] specify an interaction of a more general form than a SNP × PGS term, allowing each of the variants interacting with the 'lead' variant to have a different interaction effect coefficient, possibly with a different sign; we instead estimate a *directional* epistatic effect in which the 'lead' SNP modulates the additive effect of all the variants in the PGS in the *same direction* and with the *same magnitude.* Since our model is less parsimonious, it can be specified as a simple linear model with fixed effects and estimated with standard statistical techniques and software. While this is not necessarily an advantage *a priori,* it will allow us to find many more instances of epistasis (which also correspond to more specific genetic effects) than those found in the analysis of Fu et al., as we will argue in our discussion in Chapter 8.

Finally, in testing for an interaction with a score comprised of multiple variants, our method is reminiscent of group-based testing (Section 1.3.3), where the 'group' is in this case a PGS.

Motivation: The motivation for our approach is two-fold. First, testing for an interaction between each SNP and the PGS is a one-dimensional procedure as each SNP is tested only once. Compared to the exhaustive testing of every possible pairwise combination of SNPs, this approach is therefore much simpler computationally and requires only the usual multiple-testing correction employed in a standard GWASs.

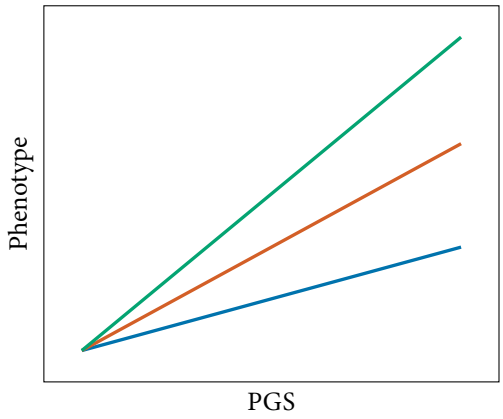
However, the most important advantage of this approach is not statistical but rather driven by biology. We noted above that biological systems are complex and driven by functional interactions between their constituent parts. While, as we saw,

these functional interdependencies do not generally equate to statistical interactions between individual genetic variants, the grouping of genes into pathways that underly the manifestation of complex traits nevertheless suggests a possible role for statistical interactions involving groups of SNPs. More specifically, we hypothesise that a disruptive mutation at a site of central importance in the architecture of a phenotype may have downstream effects on several other mutations at genes in downstream pathways – for example, a mutation affecting the functioning of a transcription factor may interact with mutations at genes whose expression is controlled by this transcription factor. In general, it seems plausible that the functional interdependency between regions of the genome could be better captured by a statistical interaction involving several mutations representative of one or more such regions as opposed to the pairwise interactions on which the epistasis literature has focused so far. We now consider illustrative examples of biological phenomena that may be well-modelled by an interaction between a single SNP and a PGS.

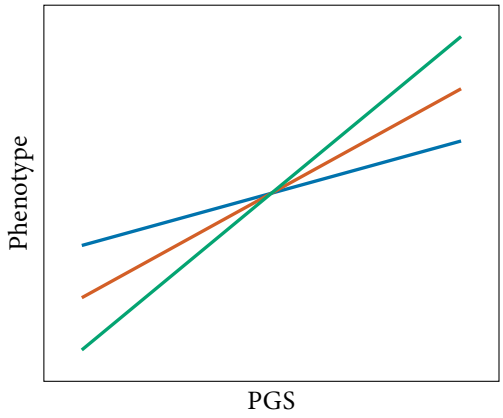
1.4.1 *Biological interpretation of SNP × PGS interactions*

In Figure 1.1, we plot three idealised examples of the relationship between the effects of an individual variant and of the PGS on a quantitative phenotype. The horizontal axis represents the range of values of the PGS: lower values of the PGS result in a lower value of the phenotype, displayed in the vertical axis, which may be a true quantitative trait such as height or instead a probability of developing or displaying a binary phenotype, such as a disease.[20] A statistical interaction between an individual SNP and the PGS corresponds to a change in the slope of the line linking the PGS to the phenotype depending on the genotype at this SNP (an individual may have 0, 1 or 2 copies of the mutated allele).

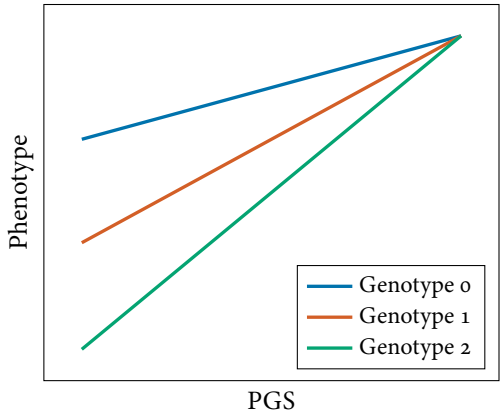
[20] In the latter case, the vertical axis corresponds to a 'liability scale' for this phenotype [121].



(a) Effect in upper range of PGS only.



(b) No direct effect on the phenotype.



(c) Effect in lower range of PGS only.

Figure 1.1: Examples of possible relationships between the effects of a SNP and of the PGS on a quantitative phenotype.

Considering the first plot, we see that the single variant increases the value of the phenotype with an effect that is much more pronounced for individuals in the upper range of the PGS distribution – indeed, there is no effect at all at the very bottom of the PGS range. Since the effect of the interaction on the phenotype operates in the same direction as that of the variant, we may detect an interaction of this form in real data by identifying a similar pattern in the signs of the estimated coefficients. This type of effect could represent that of a damaging mutation in a key gene for a particular disease: it is often hypothesised that biological systems exhibit a degree of redundancy such that it is only if multiple components of the system fail to perform adequately that a function is disrupted; a high value of the PGS for a disease could represent a 'system at risk' and whose function can be compromised by a final important mutation.

The middle plot represents a hypothetical SNP that has no direct effect on the phenotype – its average is unchanged regardless of the genotype at this locus – but which does show an interaction with the PGS. While we expect such a case to be unusual, we do find several examples of SNPs whose interaction with the PGS is significant but which have no detectable direct effect on the phenotype. The interpretation of an effect of this kind is difficult – it suggests that a SNP decreases the phenotype for individuals at the lower end of its distribution but increases it at the upper end – but deserves further investigation given the multiple examples of this behaviour that we identify.

Finally, the example displayed in the third figure is complementary to the first: here a SNP has an effect on the phenotype that is marked at the lower end of the PGS distribution but non-existent at the top – the interaction has an opposite sign to the marginal effect of the variant. This mutation could be thought to disrupt homeostasis: we may conceive of a scenario where, at the lower end of the distribution, a homeostasis mechanism is preventing the phenotype from becoming harmfully low;

a mutation then disrupts this mechanism, leading to a sharp fall in the phenotype. On the upper end of the PGS distribution, on the other hand, the fundamental mechanisms driving this phenotype have not been compromised and the homeostatic mechanism is not necessary, so that the mutation in question has no effect.

1.5 Overview of thesis

The remainder of this thesis is organised as follows. In Chapter 2, we describe the dataset to which we will apply our interaction testing approach. This is followed by a detailed description of this method in Chapter 3, which includes a discussion of relevant statistical challenges and how we address them.

Having explained our approach, we proceed to examine its properties – focusing on propensity for false-positive findings and statistical power – in a simulated dataset; the generation and analysis of this artificial data is reported in Chapter 4.

We proceed to apply our method to a real-world dataset (the UK Biobank) and examine the interactions it identifies. This analysis is divided into two chapters: in Chapter 5, we focus on a single trait (serum levels of alkaline phosphatase), whose interaction results we explore in detail; then in Chapter 6, we give an overview of the results for all the remaining traits analysed in this project, highlighting specific traits and interactions of particular relevance.

In the final chapter before concluding, we present an on-going project in which we apply our approach to a dataset of genetic variation in *HLA* genes, also in the UK Biobank.

Finally, in Chapter 8, we summarise our work, discuss its limitations and impact, and point to future research.

2 Data

We use phenotypic and genetic data from the UK Biobank (UKB) study [122].

2.1 PHENOTYPE SELECTION

We begin by selecting the quantitative traits for which we will test for the presence of genetic interactions. Our aim is to have as large and diverse a selection of traits as possible. To this end, we begin by selecting the majority[1] of traits from the following categories in the UKB data:

- Blood assays:
  - Blood biochemistry;
  - Blood count;
- Physical measures:
  - Blood pressure;
  - Body-size measures;
  - Hand-grip strength;
  - Impedance measures;
- Urine assays.

[1] When multiple closely-related traits were available (for example, measurements from the left and right side of the body), we keep only one.

In addition to the above, we also select a smaller list of traits of general interest (excluding technical measurements related to sample collection, for example) from the following more specialised categories:

- Bone densitometry of heel;
- Fluid intelligence;
- Numeric memory;
- Memory tests (pairs matching card game);
- Prospective memory (recollection of shapes and colours);
- Reaction time;
- Spirometry;

adding up to a total of 103 quantitative phenotypes.

We then filter this combined list as described below to obtain a final set of phenotypes.

2.1.1 *Genetic heritability*

Given the central role played by polygenic scores (PGSs) in our method, as will be explained below, we choose traits for which we expect to be able to build a PGS with some predictive power. To assess this, we use estimates of each phenotype's genetic heritability (specifically, additive common single-nucleotide polymorphism (SNP) heritability[2] estimated through LD Score Regression [123] for many phenotypes in the UKB by Benjamin M. Neale and collaborators [124]) and filter out traits whose heritability is less than 15%.

[2]We define heritability more precisely in Section 3.6.6 of the next chapter.

2.1.2 *Sample size*

In addition to choosing traits that are moderately heritable, we also restrict our analysis to traits available for a sizeable number of individuals, with the aim of ensuring sufficient power to detect both direct associations (which will be used to build PGSs) and interactions.

We retrieve the sample size available for each phenotype from the heritability dataset mentioned above[3] and apply a threshold of 100,000 individuals.

In selecting the final list of traits, we allow a lower heritability value to be compensated for by a larger sample size and vice-versa (e.g., if a trait has a heritability estimate that is 5% greater than 0.15 but a sample size 5% smaller than 100,000, we consider it to (exactly) satisfy our threshold). Specifically, we require the following condition to hold for each trait:

$$h^2_{\text{SNP}} \times N \geq 0.15 \times 100{,}000,$$

where h^2_{SNP} is the SNP heritability and N the sample size for a particular trait.

This selection procedure yields 84 quantitative traits from the following six main categories:

- Blood biochemistry;
- Blood count;
- Blood pressure;
- Body size measures;
- Impedance measures;
- Urine assays;

[3]The full set of samples considered in the Neale Lab's heritability analysis is comprised of approximately 361,000 people of European ancestry. This is similar to, but slightly larger than, the set of unrelated White British individuals that we use for our analysis.

and eight traits from these six more specialised categories which we group into an 'Other traits' category in our results:

- Bone-densitometry of heel;
- Fluid intelligence score;
- Hand grip strength;
- Pairs matching;
- Reaction time;
- Spirometry.

To these 92 traits, we add five others that we compute based on existing traits:

- Waist-to-hip ratio;
- Leg length (computed as the difference between standing and sitting heights);
- Relative leg length and sitting height; and
- FEV1/FVC ratio (forced expiratory volume in 1 second divided by forced vital capacity).

In total, therefore, we analyse 97 traits. These are listed in Table B.1 in the appendix.

2.2 COVARIATES

We account for the following individual covariates in our statistical analyses (how these are included in our models is described in Chapter 3):

- Age (at recruitment) and its square;
- Sex;
- Age × Sex and Age^2 × Sex;
- Genotype measurement batch (encoded as a set of 105 binary variables);
- Assessment centre attended (encoded as a set of 21 binary variables);

- 'Ancestry Components'.

Ancestry Components (ACs) were designed as an alternative to Principal Components (PCs) to account for genetic ancestry in genomic analyses [67]. They are a set of 127 variables corresponding to different regions within the UK, Europe and the world which form a partition of each individual's ancestry.

2.3 SAMPLE SELECTION

We begin by identifying all individuals in the UKB who satisfy the following conditions:

- They had not withdrawn from the study as of 14 June 2021;
- Imputed SNP dosage data is available for them;
- All covariates above are also available.

There are 487,255 such individuals. We then filter these into three subsets as described below.

2.3.1 *Main set of unrelated White British individuals*

All our primary analyses are performed on a set of 343,969 unrelated individuals of 'White British' (WB) ancestry. To identify these individuals, we rely on a previously-defined set used in the statistical analyses in the main UKB publication [122] and which we obtained directly from the authors. To define this subsample, Bycroft et al. identify a maximal set of individuals who satisfy *all* the following conditions:

- Report their ethnic background as 'British';
- Have 'very similar ancestral backgrounds' according to Principal Component Analysis (using PCs 1 to 6);
- Have imputed genotype dosages available;

- Have an inferred sex that matches their self-reported sex;
- Are unrelated (i.e., with no 3rd-degree relative or closer as inferred through estimation of kinship coefficients).

We take the intersection of this set and the full set above to obtain 343,969 WB individuals.

2.3.2 *Validation and test sets*

While our main analyses are undertaken on the set of 343,969 unrelated WB individuals discussed above, for parameter tuning and evaluation of PGS performance it is important to have separate sets of observations who are similar to those in the main sample but which are not used for building the PGSs.

For this purpose, and following standard practice in the literature, we define *validation* and *test* sets of 10,000 and 9071 individuals, respectively.

To define these two subsets of samples, we first identify all individuals in the full sample whose ancestry (as measured through ACs) is at least 90% from Great Britain (we exclude Ireland) but who are *not* part of the subset of unrelated WB individuals above (52,729 individuals).

Then, using the estimates of kinship coefficients[4] between pairs of individuals related to the 3rd degree or closer provided by the UKB, we filter out individuals related to the 2nd degree or closer[5] with someone in the WB set above (this gives us 32,148 individuals).

Finally, we remove individuals from this set to ensure that they are also unrelated to one another while aiming to keep as large a number of individuals as possible for the validation and test sets. For this, we begin by removing individuals with the

[4]These were computed with the software package KING [125].

[5]This corresponds to a kinship coefficient of at least 0.0884 according to the KING software manual.

highest number of relatives until each person appears only once in the list of related pairs, and then exclude half of these at random.

This procedure gives us 19,071 individuals from which we randomly sample 10,000 to make the validation set while the remaining 9071 are assigned to the test set.

2.4 VARIANT QUALITY CONTROL AND FILTERING

In this analysis, we consider only autosomal genetic variation. The UKB dataset includes imputed genotype calls (unphased dosages) for 93,095,623 variants in the 22 autosomes.[6] To this set of variants, we apply the following quality control filters:

- Minor allele frequency (MAF) ≥ 0.001;
- Genotype missing rate ≤ 0.05;
- Hardy-Weinberg equilibrium p-value $\geq 10^{-10}$;
- 'INFO' imputation score ≥ 0.8.

The four filtering variables above are standard in the literature (see, e.g., [122, 126, 127]). The first filter ensures that the SNPs used in our analyses are not very rare; this is important as rare SNPs can lead to false positive associations as will be discussed in the next chapter. With the second and third filters, we aim to remove SNPs for which genotyping errors are plausible. Finally, the imputation score filter corresponds to a filter on effective sample size.[7]

[6] Note that multi-allelic sites are recorded as separate variants.

[7] Bycroft et al. [122] report that an information score of 0.3 corresponds to an effective sample size of approximately 150,000. Our imputation score filter is therefore stringent in requiring the imputation quality to equate to a large effective sample size.

Applying the filters above to the set of imputed variants[8] yields a total of 12,690,793 variants that we use in our analysis.

2.5 FUNCTIONAL ANNOTATIONS

When reporting association results, we include information on the likely function of each variant and the gene within which it lies (or which it is predicted to affect) obtained from the Ensembl Variant Effect Predictor (VEP) [128] database.

[8]We compute MAF, genotype missing rate and Hardy-Weinberg equilibrium p-values for our main set of WB individuals only rather than relying on estimates for the whole UKB sample. The INFO imputation score is taken directly from the variant quality-control information supplied with the dataset.

3 Methods

Our aim is to identify statistical interactions between a variant and one or more other variants that are significantly associated with a quantitative phenotype after taking into account possible confounders. We want to detect interactions that reflect true biological interdependencies between genetic mutations rather than spurious interactions driven by statistical artefacts, and we will describe below which confounders we believe are most likely to be problematic when testing for interactions and how we have designed our method to be robust to them.

3.1 A LINEAR MODEL OF A QUANTITATIVE PHENOTYPE

The framework underlying our approach is a linear statistical model for a quantitative trait which is decomposed as the sum of additive and possible epistatic and dominant/recessive effects, the contributions of multiple covariates (these are thought to be primarily demographic – age, sex and some measure of genetic ancestry) and a zero-mean stochastic error term.

Consider a sample of N individuals indexed by $i = 1, \ldots, N$ and a set of K positions along the genome indexed by $k = 1, \ldots, K$, and denote by x_{ik} the genotype (minor allele count) of individual i at position k. Denoting the (mean-centred) phenotype

by y_i, we can write this model as follows:

$$y_i = \sum_{k=1}^{K} \beta_k x_{ik} + \sum_{k=1}^{K} \gamma_k \mathbb{I}(x_{ik} = 1) + \sum_{k<k'} \delta_{k,k'} x_{ik} x_{ik'} + \\ + \eta a_i + \theta b_i + \sum_{m=1}^{M} \iota_m c_{im} + \varepsilon_i, \tag{3.1}$$

where:

- β_k is the 'additive' or 'direct' effect of SNP k;
- $\mathbb{I}(\cdot)$ is the indicator function and γ_k a possible dominant/recessive effect of SNP k;
- $\delta_{k,k'}$ is the epistatic effect – which may be zero – of the interaction between SNPs k and k';
- η and θ represent the effects of age and sex, respectively, on the phenotype;
- the set of M variables indexed by $m = 1, \ldots, M$ with effect ι_m represents additional covariates (including a measure of ancestry); and
- ε_i is a zero-mean stochastic error term which is independent and identically distributed (i.i.d.) across individuals.

We note that a trait may need to be scaled or transformed (e.g., through a logarithmic function) for this model to be appropriate – this will be important as we consider how to test for interactions below.

We also note that it is possible for a SNP to affect the phenotype only by modulating the effect of another SNP (e.g., it is possible for $\delta_{k,k'} \neq 0$ while either $\beta_k = 0$ or $\beta_{k'} = 0$), and we do find examples of such SNPs in our analysis of real data.

Lastly, while we have not included gene–environment interactions in the model above as we will not examine such non-linear effects in this thesis, we note that these are also of scientific interest and that multiple interactions of this kind have been identified for a variety of human phenotypes [129–131].

3.2 THE NAIVE APPROACH TO INTERACTION TESTING

The simplest approach to testing for interactions between P SNPs is to test for a possible interaction between each pair of SNPs. The computational cost of such a method is quadratic in the number of SNPs considered, as with P SNPs one would need to perform $\binom{P}{2} = P(P-1)/2$ tests.

There are two important drawbacks to exhaustively testing for all possible pairwise interactions in this way. First, given the large number of SNPs which are routinely available in genetic datasets, carrying out $P(P-1)/2$ tests can be prohibitively expensive computationally. However, this challenge can be partly overcome through the use of modern computing infrastructure – by parallelising operations using graphical processing units (GPUs) or a cluster of central processing units (CPUs) – or computational methods that leverage high-efficiency data structures and operations, such as bitwise computing [72].

The second drawback is the low statistical power that methods that rely on such a large number of statistical tests will necessarily have. This challenge is more severe than that posed by their computational burden, because it is inherent to any statistical approach that undertakes a large number of independent statistical tests and so cannot be solved with more computational resources or greater efficiency in computation.

3.3 TESTING FOR SNP × PGS INTERACTIONS

We propose a new approach to detecting interactions that does not require exhaustive pairwise testing and is computationally tractable.

Consider a PGS for a trait y denoted by $\hat{S}_i$ for individual i and defined as a weighted sum of mutation counts at a set of L positions along the genome indexed by $l = 1, \ldots L$, i.e., $\hat{S}_i = \sum_{l=1}^{L} \hat{\beta}_l x_{il}$. Ideally, this PGS would correspond exactly to the

set of additive effects in the 'true model' given in equation (3.1), i.e., $\sum_{k=1}^{K} \beta_k x_{ik}$. In practice, however, we will need to rely on an empirically-derived PGS for which both the SNPs included and their coefficients may be different from the set of SNPs and coefficients in the true PGS.[1] We will discuss below how this will affect our inference and the steps we take to minimise its impact.

Given an inferred PGS $\hat{S}_i$ for each individual, we propose to test for statistical interactions genome-wide between each SNP x_{ij} under consideration and the PGS. More specifically, we fit the following model (containing an interaction term between the SNP x_{ij} and the estimated PGS $\hat{S}_i$) once for each SNP in the genome present in the data:

$$\begin{aligned} y_i = \beta_j x_{ij} + \kappa \hat{S}_i + \lambda_j x_{ij} \hat{S}_i + \\ + \eta a_i + \theta b_i + \sum_{m=1}^{M} \iota_m c_{im} + \varepsilon_i, . \end{aligned} \tag{3.2}$$

Our motivation in looking for interactions between single genetic variants and a PGS is two-fold. First, this approach requires only P tests and is therefore a possible solution to the computational challenges and low power of exhaustive approaches.

More significantly, however, this approach has greater power than standard methods to *find genetic variants that interact with multiple other variants.* We consider this to be an important advantage as we expect such interactions to exist: given the importance of gene regulatory networks – where a SNP that impacts the expression of one gene can have downstream effects on many others – we expect some SNPs to interact statistically with mutations in multiple loci along the genome. To our knowledge, this is the first method to specifically target such interactions.

We will study the performance of our interaction-testing strategy both theoretically and in simulations, and will show that this approach performs well in the realistic

[1] We denote the 'true' PGS by S_i and the estimated PGS by $\hat{S}_i$; similarly, β represents true coefficients and $\hat{\beta}$ their estimated counterparts.

case where a variant interacts with only a subset (i.e., a proportion) of the variants in the PGS.

3.4 BIAS AND POWER CONSIDERATIONS

We now proceed to consider our method's propensity to lead to inferential errors of either type I or type II.

Our first concern is to *avoid false positives*, i.e., type I errors: we aim to find interactions between groups of SNPs which reflect real biological relationships rather than being the result of statistical artefacts.

We will consider potential sources of false positives by deriving the bias of the ordinary least-squares (OLS) estimator of coefficient λ_j in equation (3.2) in a simplified setting and examining the different components making up this bias term. In the process, we will also gain insight into factors that can reduce the method's statistical power to identify true effects. Then, in Section 3.5, we will examine how we may prevent false positives while increasing statistical power whenever possible.

3.4.1 *Asymptotic properties of the OLS estimator of the SNP × PGS interaction term*

We again consider a sample of N individuals indexed by i for whom we observe a quantitative phenotype y_i. We assume the following simple model for y_i which contains both additive and interaction genetic effects:

$$y_i = S_i + \lambda_j x_{ij} R_i + \varepsilon_i, \tag{3.3}$$

where S_i is the (true) PGS, x_{ij} is a SNP, R_i is part of the PGS S_i as explained next, and ε_i is a zero-mean error term with finite variance which is i.i.d. across individuals. For simplicity and without loss of generality, all variables are assumed to have mean zero and x_{ij} is also assumed to have variance one.

We further assume that the PGS S_i can be decomposed as follows:

$$S_i = Q_i + R_i, \tag{3.4}$$

where

$$Q_i \coloneqq \sum_{o=1}^{O} \beta_o x_{io} \tag{3.5}$$

$$R_i \coloneqq \sum_{p=1}^{P} \beta_p x_{ip}. \tag{3.6}$$

Under this model, therefore, y_i is driven by an additive component S_i and also by an *interaction* between a single SNP x_{ij} and a component R_i of the PGS.

As noted previously, in real applications we do not observe the true PGS S_i and instead have to rely on an approximation or estimate $\hat{S}_i$. Here, we will assume that this is observed with an *additive and independent* measurement error relative to the true score, i.e.:

$$\hat{S}_i = S_i + \zeta_i, \tag{3.7}$$

with ζ_i a zero-mean, finite-variance error which is i.i.d. across individuals.

Equations (3.4) and (3.7) imply the following expression for R_i, the part of the PGS that interacts with SNP x_{ij}:

$$\begin{aligned} R_i &= S_i - Q_i \\ &= \hat{S}_i - \zeta_i - Q_i. \end{aligned} \tag{3.8}$$

Finally, we make the following assumptions regarding independence relationships and correlations between variables. First, the general error term ε_i is assumed to be *independent* of the explanatory variables x_{ij} and S_i, as well as of ζ_i, which implies independence from $\hat{S}_i$. Second, the variant x_{ij} is assumed to be *uncorrelated* with all the variables related to the PGS and its components, i.e., $\mathbb{E}(x_{ij}Q_i) = \mathbb{E}(x_{ij}R_i) =$

$\mathbb{E}(x_{ij}S_i) = \mathbb{E}(x_{ij}S_i^2) = \mathbb{E}(x_{ij}\zeta_i) = \mathbb{E}(x_{ij}\hat{S}_i) = \mathbb{E}(x_{ij}\hat{S}_i^2) = 0$, while x_{ij}^2 is assumed to be uncorrelated with ζ_i, $\hat{S}_i$ and $\hat{S}_i^2$.

The first assumption is required for the OLS estimator to be unbiased, as we will see in more detail below.[2] The second assumption is more interesting as it has biological meaning: assuming that the SNP and the PGS are uncorrelated is equivalent to assuming that they are not in LD. One way to achieve this is by making sure that the PGS does not contain any variant present in the same chromosome as the SNP being tested for an interaction, i.e., building a 'leave-one-chromosome-out' (LOCO) PGS, as we will discuss in Section 3.5.3.

Our goal is to recover the interaction coefficient λ_j. To this end, we fit the following linear regression model,[3] estimating its coefficients through OLS:

$$y_i = \alpha_0 + \alpha_1\hat{S}_i + \alpha_2 x_{ij}\hat{S}_i + u_i.$$

We now proceed by formally deriving basic properties of our estimator under different scenarios. We will focus on large-sample properties (in particular, consistency) as these are the most relevant to our application in a setting with many thousands of samples. The aim of this theoretical analysis is to formalise some of our arguments as we consider how to best build a PGS and perform interaction testing in our context.

We will begin with a trivial example (Case 1) in which we assume that the interaction in the true model involves the *full PGS* which we *observe perfectly* – this will serve as a baseline for subsequent derivations.

We will then relax these two assumptions in turn: first, we will consider a scenario in which the interaction is with only part of the PGS as above (Case 2); then, we

[2] For consistency, it is sufficient to assume no correlation between the error term and the explanatory variables.

[3] Not including a main term for x_{ij} in the regression model will greatly simplify the derivations below. When running regressions in practice, however, we will always include main terms for all variables.

will assess the consequences of additive measurement error in the PGS when the interaction again involves the full PGS for simplicity (Case 3).

Finally, in Case 4, we will consider what happens when both issues (interaction with only part of the PGS, which is measured with additive error) are present simultaneously, and will examine the consequences of relaxing the assumptions made above of no correlation between the variant and the PGS.

Case 1: interaction with the full PGS which is perfectly observed: We begin by considering the OLS estimator of λ_j under the simplest possible setting: we assume that the SNP x_{ij} interacts with the full PGS and that we observe the true value of this PGS without error, i.e., $R_i = S_i = \hat{S}_i$.

The true model is therefore:

$$y_i = S_i + \lambda_j x_{ij} S_i + \varepsilon_i \tag{3.9}$$

and we fit the regression model through OLS:

$$y_i = \alpha_0 + \alpha_1 S_i + \alpha_2 x_{ij} S_i + u_i. \tag{3.10}$$

Proposition 3.4.1. *The OLS estimator $\hat{\alpha}_2$ of the coefficient λ_j in the true model given in equation* (3.9) *obtained from the regression model in equation* (3.10) *is unbiased, efficient*[4] *and consistent.*

Proof. These properties follow from classic results for least-squares estimators of linear models under standard assumptions (see, e.g., refs. [132, Sections 1.3 and 2.3]). We now briefly list the assumptions required for our three claims (unbiasedness, efficiency and consistency) and explain why they are met in this toy model.

Unbiasedness requires (i) the 'strict exogeneity' assumption ($\mathbb{E}(\varepsilon_i | S_i, x_{ij} S_i) = 0$ for all i); and (ii) the 'no multicollinearity' assumption ($x_{ij} S_i$ is not a linear function

[4] Efficiency means that it achieves the lowest variance within the class of linear unbiased estimators.

of S_i). We assumed strict exogeneity in the general model setup at the start of this session and no multicollinearity is met whenever x_{ij} segregates in the sample.

Efficiency is guaranteed by the Gauss-Markov Theorem. In addition to the two assumptions required for consistency, this theorem requires (iii) the 'spherical error variance' assumption, which comprises homoscedasticity (constant variance across observations, i.e., $\mathbb{E}(\varepsilon_i^2|S_i, x_{ij}S_i) = \sigma_2 > 0$ for all i) and no correlation of errors between observations (i.e., $\mathbb{E}(\varepsilon_i\varepsilon_j|S_i, x_{ij}S_i) = 0$ for all i, j with $i \neq j$), conditions which are implied by our assumption of i.i.d. errors with finite variance.

Finally, and moving on from the two finite-sample properties above to asymptotic results, *consistency* (i.e., convergence in probability[5] of the least-squares estimator to the true value of the parameter as the sample size tends to infinity[6]) requires (i) i.i.d. samples,[7] (ii) no correlation between explanatory variables and the errors (i.e., $\mathbb{E}(S_i\varepsilon_i) = \mathbb{E}(x_{ij}S_i\varepsilon_i) = 0$ for all i)[8] and, finally, (iii) no multicollinearity, as before. All these assumptions are met by construction in this toy example. □

Case 2: interaction with only part of the PGS which is perfectly observed: We now return to the setting introduced at the start of this section (equation (3.3)), assuming that the interaction is with only part of the PGS, R_i. However, for simplicity we assume that we still observe the full PGS perfectly.

The true model is now:

$$y_i = S_i + \lambda_j x_{ij} R_i + \varepsilon_i \tag{3.11}$$

[5] Consider a sequence of estimators $\hat{X}_1, \hat{X}_2, \ldots, \hat{X}_n$, where $\hat{X}_j$ is computed using the first j observations. Convergence in probability of this sequence of estimators to a value a means that, for any $\varepsilon > 0$, $\lim_{n\to\infty} \Pr(|\hat{X}_n - a| < \varepsilon) = 1$, which we denote more concisely by $\hat{X}_n \xrightarrow{\text{p}} a$. See, e.g., ref. [133, Section 5.2].

[6] See, e.g., ref. [133, Section 6.3].

[7] Strictly speaking, it requires 'ergodic stationarity', of which the i.i.d. samples setting is a special case.

[8] Note that this is a weaker assumption than independence – which we had to assume for unbiasedness and efficiency – and is implied by that stronger assumption.

but as we do not know R_i – indeed, our goal is to identify the SNPs which it contains – we fit the following regression:

$$y_i = \alpha_0 + \alpha_1 S_i + \alpha_2 x_{ij} S_i + u_i. \tag{3.12}$$

We begin by noting that the least-squares estimator of α_2 in the simple two variable linear model $y_i = \alpha_0 + \alpha_1 x_{1i} + \alpha_2 x_{2i} + u_i$ can be shown to be [134]:

$$\hat{\alpha}_2 = \frac{\sum_i x_{1i}^2 \sum_i x_{2i} y_i - \sum_i x_{1i} x_{2i} \sum_i x_{1i} y_i}{\sum_i x_{1i}^2 \sum_i x_{2i}^2 - \left(\sum_i x_{1i} x_{2i}\right)^2}.$$

The estimator for the model in equation (3.12) is therefore:

$$\hat{\alpha}_2 = \frac{\sum_i S_i^2 \sum_i x_{ij} S_i y_i - \sum_i S_i^2 x_{ij} \sum_i S_i y_i}{\sum_i S_i^2 \sum_i x_{ij}^2 S_i^2 - \left(\sum_i S_i^2 x_{ij}\right)^2}. \tag{3.13}$$

Proposition 3.4.2. *The OLS estimator $\hat{\alpha}_2$ of the coefficient λ_j in equation* (3.11) *converges in probability according to the following expression:*

$$\hat{\alpha}_2 \xrightarrow{\text{p}} \lambda_j \frac{Cov(R_i, S_i)}{Var(S_i)}.$$

Proof. See Appendix A.1. □

Proposition 3.4.2 shows formally that our ability to detect an interaction between a SNP and the PGS will depend on the *correlation between the fraction of the PGS with which the SNP interacts and the full score:* the estimated coefficient will be *attenuated* (i.e., biased towards zero) with a degree of attenuation inversely proportional to the fraction of the PGS affected.

It is worth noting that R_i and Q_i represent the *weighted* contribution of the SNPs contained in them to the PGS, with weights given by each variant's β coefficient. Therefore, if a SNP interacts with a small fraction of the SNPs that comprise the PGS but the SNPs with which it interacts have large effect sizes, this will reduce the degree of dilution.

In summary, *our method will have greater power to detect associations that involve either a large fraction of the SNPs that are part of the PGS, or especially impactful SNPs within that PGS.*

Case 3: interaction with the full PGS which is estimated with additive error: We proceed by considering the slightly more complex setup under which the PGS is estimated with additive error (i.e., we observe $\hat{S}_i = S_i + \zeta_i$) while, for simplicity, the SNP x_{ij} again interacts with the full PGS as in Case 1.

The true model is:

$$y_i = S_i + \lambda_j x_{ij} S_i + \varepsilon_i \tag{3.14}$$

but we now fit a regression model with an estimated PGS $\hat{S}_i$:

$$y_i = \alpha_0 + \alpha_1 \hat{S}_i + \alpha_2 x_{ij} \hat{S}_i + u_i. \tag{3.15}$$

This setting is reminiscent of the classical errors-in-variables model in the statistics and econometrics literatures (see, e.g., [135, Section 9.14], [136, Section 1.1] or [137, Section 8.8.1]). However, unlike in those models, in which the error is purely additive for all the variables observed with error, here the error in our variable of interest, $x_{ij}\hat{S}_i$, is multiplied by the allele count of the SNP under consideration.

We will now derive the impact of this error on the estimator of λ_j and show that it will again be attenuated, with the strength of attenuation increasing as the signal-to-noise ratio in the PGS decreases. This is again similar to the basic errors-in-variables model.

Adapting equation (3.13), we see that the estimator for the model in equation (3.15) is:

$$\hat{\alpha}_2 = \frac{\sum_i \hat{S}_i^2 \sum_i x_{ij} \hat{S}_i y_i - \sum_i \hat{S}_i^2 x_{ij} \sum_i \hat{S}_i y_i}{\sum_i \hat{S}_i^2 \sum_i x_{ij}^2 \hat{S}_i^2 - \left(\sum_i \hat{S}_i^2 x_{ij}\right)^2}. \tag{3.16}$$

Proposition 3.4.3. *The OLS estimator $\hat{\alpha}_2$ of the coefficient λ_j in equation* (3.14) *converges in probability according to the following expression:*

$$\hat{\alpha}_2 \xrightarrow{p} \lambda_j \frac{Cov(\hat{S}_i, S_i)}{Var(\hat{S}_i)}.$$

Proof. See Appendix A.2. □

We see first from Proposition 3.4.3 that, in the case of no noise in the estimated PGS (i.e., $\zeta_i = 0$ for all i), the OLS estimator converges to the true parameter value λ_j as we had seen previously in Case 1. What is new in this setting is that, as the expected noise fraction in the PGS increases, the estimate of the interaction coefficient is attenuated towards zero. This implies that *the less accurate our PGS, the less power we will have to detect interaction signals.* We will discuss below how we may define 'accuracy' of a PGS in this context.

Case 4: interaction with only part of the PGS which is estimated with additive error: We conclude this brief theoretical analysis of the OLS estimator of the toy model given in equation (3.3) by considering a case in which both conditions examined separately above are present simultaneously. We now assume:

- That the SNP x_{ij} interacts with only part of the PGS, denoted again by R_i; and
- That the PGS is estimated with additive error, i.e., we observe $\hat{S}_i = S_i + \zeta_i$.

The true model is:

$$y_i = S_i + \lambda_j x_{ij} R_i + \varepsilon_i \tag{3.17}$$

and we fit:

$$y_i = \alpha_0 + \alpha_1 \hat{S}_i + \alpha_2 x_{ij} \hat{S}_i + u_i. \tag{3.18}$$

The OLS estimator for α_2 is given in equation (3.13). Proceeding similarly to our previous derivations *but without assuming no correlation between x_{ij}, x_{ij}^2 and the different variables related to the PGS as before,* we arrive at the following expression for the probability limit of $\hat{\alpha}_2$ (see Appendix A.3):

$$\hat{\alpha}_2 \xrightarrow{\text{p}} \lambda_j \Bigg[1 - \frac{\mathbb{E}(\hat{S}_i^2)\mathbb{E}(x_{ij}^2 \hat{S}_i Q_i)}{\mathbb{E}(\hat{S}_i^2)\mathbb{E}(x_{ij}^2 \hat{S}_i^2) - \left[\mathbb{E}(\hat{S}_i^2 x_{ij})\right]^2} \tag{3.19}$$

$$+ \frac{\mathbb{E}(\hat{S}_i^2 x_{ij})\mathbb{E}(\hat{S}_i x_{ij} Q_i)}{\mathbb{E}(\hat{S}_i^2)\mathbb{E}(x_{ij}^2 \hat{S}_i^2) - \left[\mathbb{E}(\hat{S}_i^2 x_{ij})\right]^2} \tag{3.20}$$

$$- \frac{\mathbb{E}(\hat{S}_i^2)\mathbb{E}(x_{ij}^2 \hat{S}_i \zeta_i)}{\mathbb{E}(\hat{S}_i^2)\mathbb{E}(x_{ij}^2 \hat{S}_i^2) - \left[\mathbb{E}(\hat{S}_i^2 x_{ij})\right]^2} \tag{3.21}$$

$$+ \frac{\mathbb{E}(\hat{S}_i^2 x_{ij})\mathbb{E}(\hat{S}_i x_{ij} \zeta_i)}{\mathbb{E}(\hat{S}_i^2)\mathbb{E}(x_{ij}^2 \hat{S}_i^2) - \left[\mathbb{E}(\hat{S}_i^2 x_{ij})\right]^2} \Bigg] \tag{3.22}$$

$$- \frac{\mathbb{E}(\hat{S}_i^2)\mathbb{E}(x_{ij} \hat{S}_i \zeta_i) - \mathbb{E}(\hat{S}_i^2 x_{ij})\mathbb{E}(\hat{S}_i \zeta_i)}{\mathbb{E}(\hat{S}_i^2)\mathbb{E}(x_{ij}^2 \hat{S}_i^2) - \left[\mathbb{E}(\hat{S}_i^2 x_{ij})\right]^2}. \tag{3.23}$$

The limit above allows us to see at a glance three sources of bias in the estimator $\hat{\alpha}_2$ of λ_j:

1. In lines (3.19) and (3.20), we see the *attenuation bias* caused by the fact that we are testing for an interaction with the full PGS while the SNP x_{ij} interacts only with a fraction R_i of it, similarly to what we saw in Case 2. Note that, if we assumed that x_{ij} is uncorrelated with $\hat{S}_i^2$, and that x_{ij}^2 is uncorrelated with Q_i, $\hat{S}_i$ and $\hat{S}_i^2$, the first fraction would simplify to $\text{Cov}(Q_i, \hat{S}_i)/\text{Var}(\hat{S}_i)$ while the second would be zero;

2. In lines (3.21) and (3.22), we again see an *attenuation bias*, this time caused by the additive error ζ_i in the inferred PGS. If we assumed that x_{ij} is uncorrelated with $\hat{S}_i^2$, and that and x_{ij}^2 is uncorrelated with ζ_i, $\hat{S}_i$ and $\hat{S}_i^2$, we would obtain a similar result to that in Case 3;

3. Lastly, in line (3.23), we see a *general bias term* driven by ζ_i, the additive error in the PGS. Importantly, this bias is present even if there is no interaction involving this locus, i.e., $\lambda_j = 0$. This is significant as it is *liable to cause false positives:* our estimate of the interaction term may be non-zero in cases where there is no interaction. However, assuming that x_{ij} is uncorrelated with ζ_i, $\hat{S}_i$ and $\hat{S}_i^2$ makes this term disappear. In other words, *there being no correlation (or LD) between the focal SNP and the PGS can reduce the potential for false positives;* this source of bias can also be minimised by having an 'accurate' PGS with as little error as possible.

Of these three sources of bias, the first is unavoidable while the second and third can in principle be reduced through careful design of the methods used for interaction testing. More specifically, the attenuation bias caused by the fact that we test for an interaction with the full PGS while many interactions may involve only a small subset of this PGS is a natural and necessary result of our testing strategy – it implies that this approach will be more likely to detect interactions with a substantial proportion of the PGS, as we noted above in Case 2. By contrast, the attenuation bias caused by the additive error in the inferred PGS can be reduced by using a more 'accurate' PGS – we discuss how we may define accuracy in this context below.

Finally, the general bias term that results from additive error in the PGS can be eliminated by ensuring that the variant being tested for an interaction is not in LD with any of the variants in the PGS, or reduced by minimising the additive error in the inferred score. We will make use of these observations below when describing how we will construct our PGS and perform interaction testing to avoid false positives. For now, we take a second look from a different perspective at the potential consequences of LD between genetic variants being tested for an interaction.

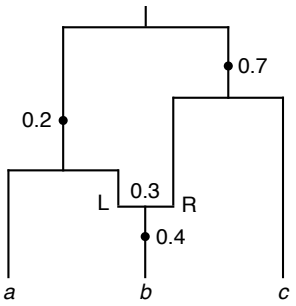
3.4.2 *LD as a source of false positives in interaction testing*

Consider the toy example of genetic variation for a sample of three individuals a, b and c given in Figure 3.1. We assume:

- Three segregating sites in the sample at relative positions 0.2, 0.4 and 0.7 in the genome;
- The genealogical history given in Figure 3.1(b), which includes a recombination event at position 0.3.[9]

	Position		
Ind.	0.2	0.4	0.7
a	1	0	0
b	1	1	1
c	0	0	1

(a) Genotype table



(b) Ancestral recombination graph

Figure 3.1: Example of genetic variation at three sites with genealogical history.

The recombination event means that different parts of individual b's genome coalesce first with either individual a (the left part) or individual c (the right part), and leads to b having all three mutations as shown in the genotype table.

We define three variables x_{1i}, x_{2i} and x_{3i} for individuals indexed by $i \in \{a, b, c\}$ corresponding to mutations at positions 0.2, 0.4 and 0.7, respectively. These variables take the value 1 if a mutation is present for an individual at that site and the value 0 otherwise. Note that, in this sample, we have the following relationship between the

[9] We are grateful to Anastasia Ignatieva for conceiving of the ancestral recombination graph for this example.

three genotypes:

$$x_{2i} = x_{1i}x_{i3} \quad \text{for all } i. \tag{3.24}$$

We now consider a quantitative phenotype y and assume the following model:

$$y_i = \alpha_1 x_{1i} + \alpha_2 x_{2i} + \varepsilon_i, \tag{3.25}$$

where α_1, α_2 are the coefficients of SNPs x_{1i}, x_{2i} and ε_i is an error term as before.

Using equation (3.24), the model in equation (3.25) can be equivalently rewritten as:

$$y_i = \alpha_1 x_{1i} + \alpha_2 x_{1i}x_{3i} + \varepsilon_i.$$

From this we see that, even though there is no interaction in the original model given in equation (3.25), a test for an interaction between x_{1i} and x_{3i} could nevertheless suggest an interaction between these two variants. However, this interaction is not 'real' in the sense that it can be equivalently accounted for by a single variant, x_{2i}, and results from the particular genealogical history of this sample.

We note in conclusion that this is caused by the close pattern of LD between the three variants, each of which is correlated with the other two. We will return to this example when discussing how to address potential false positives caused by LD below.

3.4.3 *Dominant variants*

The concepts of dominance and recessiveness were introduced by Gregor Mendel in reference to discrete phenotypic traits when reporting on his experiments with peas [138], and were later generalised to refer to alleles rather than phenotypes and to apply to quantitative traits as well as discrete ones [139].

In modern usage, an allele is considered dominant if the effect of a single copy of this allele being present is larger in magnitude than the additional effect of a second

copy of the allele on the phenotype [140]. Indeed, the second copy of the allele may have no effect at all on the phenotype beyond that caused by the first copy.

We denote by x_{ij} a variant whose value is 0, 1 or 2 depending on whether 0, 1 or 2 copies of the allele in question are present, respectively, for an individual, and write the following linear model of a quantitative phenotype:

$$y_i = \alpha\, \mathbb{I}(x_{ij} = 1) + \beta\, \mathbb{I}(x_{ij} = 2) + \boldsymbol{\theta}^\top \boldsymbol{x}_{i,-j}, \tag{3.26}$$

where $\boldsymbol{x}_{i,-j}$ is a vector of genotypes at locations other than j that have an effect on the phenotype and $\boldsymbol{\theta}^\top$ is a transposed vector of their effects. Based on this simple model, we can define the allele whose number of copies the variable x_j counts as dominant (which makes the alternative allele at that position recessive) if and only if $|\alpha| > |\beta|/2$.

We note two relevant facts about dominant effects in the context of modelling epistasis. First, representing dominance for a genetic variant requires *two parameters*. This is because the typical assumption that the second copy of an allele has the same effect as the first, which is implicit in linear genetic models that do not consider dominance, does not hold. Moreover, the model in equation (3.26) could equivalently be written as:

$$y_i = \gamma x_{ij} + \delta\, \mathbb{I}(x_{ij} = 1) + \boldsymbol{\theta}^\top \boldsymbol{x}_{i,-j}$$

or as:

$$y_i = \zeta x_{ij} + \eta x_{ij}^2 + \boldsymbol{\theta}^\top \boldsymbol{x}_{i,-j}. \tag{3.27}$$

As long as there are two free parameters available to represent the distinct effects of the first and second copy of the mutated alleles, dominance effects can be fully captured by these different parameterisations of the model.

Second, we see from equation (3.27) that dominance may be thought of and represented as a *self-interaction*. This is important in the context of our work because

it suggests that dominance could potentially be detected by our method to test for interactions. While dominance is certainly worthy of further study, it is not the focus of this work and in fact has already been explored recently and in depth for over 1000 phenotypes in the UKB [140]. Because of this, we will treat dominance as a possible confounder in our search for interactions and seek to prevent our method from detecting these non-linearities as potential interactions, as described in Section 3.5.

3.4.4 *Non-linear functional form of the genetics–trait mapping*

When introducing the general statistical model in equation (3.1), we noted that it may be necessary to transform the trait in question for this simple linear model to be appropriate. By this we mean that the most appropriate model for a particular trait may take the following form:

$$y_i = f\left(\sum_{k=1}^{K} \beta_k x_{ik} + \sum_{k=1}^{K} \gamma_k \mathbb{I}(x_{ik} = 1) + \sum_{k<k'} \delta_{k,k'} x_{ik} x_{ik'} + \eta a_i + \theta b_i + \sum_{m=1}^{M} \iota_m c_{im} \right) + \varepsilon_i, \tag{3.28}$$

where $f : \mathbb{R} \to \mathbb{R}$ is a function that non-linearly transforms the linear genetic and environmental effects (e.g., f can be a quadratic, exponential or sigmoid function) to generate the trait on the scale in which it is observed.

In general, there is no biological or mathematical reason to expect genetic variation, even if it can be additively combined, to translate linearly onto the scale of the trait. This is especially true for traits whose scale is in some sense 'artificial' by construction – for example, the body mass index (BMI), which does not measure a directly observable phenotype but is rather the non-linear transformation of two different phenotypes.

Such 'global' non-linearities involving all SNPs relevant for a trait are interesting in themselves and have been studied before (see, e.g., Otwinowski et al. [141], who refer

to this non-linear mapping as 'global epistasis'). However, they are not our object of focus and, importantly, may produce spurious statistical interactions between smaller sets of SNPs if unaccounted for. For example, if f is quadratic, then even in the absence of dominance effects and pairwise interactions (i.e., $\gamma_k = 0$ and $\delta_{k,k'} = 0$ for all $k : k < k'$), and ignoring the effects of covariates, by the Binomial Theorem the expression $\left(\sum_{k=1}^{K} \beta_k x_{ik}\right)^2$ will result in pairwise interaction terms for every possible pair of mutations.

3.4.5 *Phenotypic outliers*

We conclude this sub-section on bias and power by considering the potential effects of phenotypic outliers.

In its simplest form, association testing using genotyping array data (where allele counts take values $0, 1$ or 2 rather than being an imputed dosage taking values in $[0, 2]$) is a test for the differences in the mean phenotypic value across the three discrete groups of samples defined by the allele count. Since means are notoriously sensitive to outlying values, association tests for the effect of a variant can be affected by observations with extreme phenotypes, which may lead to false-positive associations.

Consider for example a rare variant (with MAF < 0.001, for example), as these are usually the most likely to yield false-positive associations. Since the frequency of this mutation is very low, the group of samples who are homozygous for this mutation will also be very small. Because of this, if a few observations in that group happen to have a very low or very high value for their phenotype, this can sufficiently decrease or increase the mean phenotypic value in that group such that the null hypothesis of no association is rejected – while it would not be if the observations in question were removed from the sample. These associations may be spurious because they are driven by a small number of observations, and so their significance is sensitive

to departures from modelling assumptions such as normality. This problem is even more severe when testing for the effects of interactions, since the relevant frequency will then be the product of the frequencies of the two individual variants, and this will be very small for two rare variants.

Due to the potential for rare variants to be incorrectly classified as being associated with a phenotype, it is common to filter out variants with very low frequency before undertaking association testing – we also perform such filtering, as discussed in Section 2.4 of the preceding chapter. Moreover, it is also standard practice to identify and pre-process observations with extreme phenotypic values to ensure that these do not carry a disproportionate weight in the analysis, a procedure which we will also undertake, as described below.

3.5 Avoiding false positives while increasing power

In Section 3.4, we identified five factors that may lead to false-positive associations and potentially also reduce our power to identify true interactions. We list them here in the order in which each will be addressed in our data analysis pipeline, which we describe in detail in Section 3.6:

1. Observations with extreme phenotypic values;
2. Errors in the inferred PGS;
3. LD between variants being tested for an interaction;
4. Global non-linearities in the genetics–trait map; and
5. Genetic dominance.

We now consider each of these five factors in order, describing our strategies to pre-empt their potential for causing false positives and, where possible, increase the statistical power of our approach.

3.5.1 *Truncation and quantile-normalisation of the phenotype*

We begin with extreme phenotypic values or outliers. Outliers are common in real-world data and as such different methods have been developed in the statistics literature to correct for their potentially-harmful effects. We use two standard approaches, Winsorisation and rank-based inverse Normal transformation (INT), to obtain two alternative versions of each phenotype, and then run the remaining of our data analysis pipeline in parallel for each version.

Employing two alternative approaches in parallel gives us the opportunity of comparing our findings under common transformations of the phenotype and in this way assess their robustness.

Winsorisation (named after statistician Charles P. Winsor, 1895–1951) consists of truncating the values of all outliers below and above two given percentiles in the empirical distribution of a variable to the values at those percentiles. For example, if the percentiles chosen are the 1st and the 99th, observations whose value is lower than the 1st percentile of the distribution have their value reassigned to that of the 1st percentile, and similarly for the 99th percentile (see, e.g., [142] or [143, Section 9.1]).

In our analysis, we first regress out the covariates (listed in the previous chapter in Section 2.2) from the phenotypes and then apply Winsorisation at the 1st and 99th percentiles to the residuals of this regression.

INT consists of ranking the values of the variable in question and mapping those to the quantiles of a standard Normal distribution, thus ensuring that the transformed variable is Normally distributed. More precisely, if r_i is the rank of subject i of N, we can compute a transformed value y_i as follows [144, 145]:

$$y_i = \Phi^{-1}\left(\frac{r_i - c}{N - 2c + 1}\right), \quad c \in [0, 1/2],$$

where $\Phi(\cdot)$ is the cumulative density function of the standard Normal distribution.

Common values for c include 3/8, 1/2 and 0, with different values considered unlikely to have an impact on the results of the analysis [144]. In our pipeline, we will use the value 0 as this is the default adopted in the `R` package `bestNormalise` [146, 147] which we employ for this purpose.

3.5.2 *An iterative linear regression algorithm for PGS construction*

Our approach to testing for interactions relies on first obtaining an inferred PGS against which interactions involving individual mutations will be tested. We now consider the question of how to build a suitable PGS for this application.

While several methods for PGS construction have been proposed in the literature in recent years (see, e.g., Ni et al. [22] for a comparison of ten popular methods and Zhang et al. [23] for a helpful graphical summary of eight alternative Bayesian approaches), none of these methods has the properties that we believe are ideal in this setting. We therefore propose a new PGS construction algorithm designed to perform well in our context.

In Section 3.4, we explored the consequences of *additive error* in the PGS on the asymptotic behaviour of the OLS estimator of the coefficient of the SNP × PGS interaction term. While assuming a purely additive error term is a simplification, it nevertheless allowed us to derive two insightful results: additive error in the inferred PGS (i) reduces power to detect true interactions and (ii) can lead to biased estimates – in an unpredictable direction – of the interaction coefficient.

This second result is particularly important as it suggests that a poorly constructed PGS could lead to false-positive evidence of interactions. We will therefore seek to build a score that makes false positives less likely while retaining power to discover true interactions.

We begin by considering the following question: in what ways can a PGS be 'wrong' or poorly estimated? Our goal is to estimate the additive genetic component of the trait given by the term $\sum_{k=1}^{K} \beta_k x_{ik}$ in equation (3.1) as accurately as possible.

We note first that a PGS is defined and computed based on two vectors:

- A vector of the SNPs included in the PGS; and
- A vector of the same length with the coefficients of each of these SNPs.

Each of these vectors is a source of potential error: the SNPs that we choose to include in the PGS will not always be the true causal SNPs (they may only imperfectly tag a causal variant, for example, or even miss it completely) and their effect sizes will be imperfectly estimated. Therefore, when inferring a PGS for interaction testing, our aim will be to minimise both sources of error.

Our approach involves two steps which we now summarise and then explain in greater detail below:

1. We will first build an *initial PGS* which is designed to account for as great a fraction of the SNPs with a direct (i.e., additive) effect on the trait as possible.

   To achieve this, we include in our PGS a large selection of SNPs found to have an effect of some significance on the phenotype in a simple GWAS. In this first stage, the SNPs' effect sizes are those obtained in that initial GWAS;

2. The inclusion of a large number of SNPs in the PGS – many of which will be in close physical proximity with one another – with effects obtained through a (univariate) GWAS will lead to a 'double counting' of the effects of neighbouring variants.

   To address this, in a second stage we employ an *iterative algorithm* that calibrates the effect sizes of SNPs in close proximity to account for the LD between them.

> The goal of this algorithm – which is also its stopping condition – is that none of the SNPs in the PGS remain significant after running it. This means that, if we regress out the final PGS from the trait and perform a GWAS on the resulting residuals, none of the PGS SNPs should be 'genome-wide significant' (GWS; i.e., have a p-value lower than or equal to 5×10^{-8}).

Through this two-step approach, we are able to obtain a PGS which includes most of the relevant SNPs for the trait in question while also exhausting the additive signal for these SNPs. In the context of our list of three potential sources of bias in the estimator of the SNP–PGS interaction term, exhausting the additive signal can be seen as minimising the additive error ζ_i – this can reduce both the *attenuation bias* caused by this error (the second item in this list) as well as potential *false-positives* (the third item in the list).

Recent results in the epistasis literature highlight the importance of accounting for all additive signal for a trait when testing for interactions: it has been shown, both in theory and in practice, that imperfectly-tagged additive causal effects in the vicinity of SNPs being tested for an interaction can lead to the incorrect detection of spurious epistatic associations (termed 'phantom epistasis') [148–150]. In accounting for as many relevant variants as possible and exhausting their additive signals, our method seeks to pre-empt this potential source of false positives.

Training, validation and test sets: Our PGS construction method requires, in addition to a main dataset which we will term the 'training set', two other smaller datasets for calibrating parameters (the 'validation set') and testing the final performance of the scores (the 'test set'). We described how we constructed these three separate datasets in our data in the previous chapter (Sections 2.3.1 and 2.3.2).

Step 1: Building an initial PGS with lax Clumping and Thresholding: To build an initial PGS which includes as many variants with a direct effect on the trait as possible, we use Clumping and Thresholding (C+T) with a very *loose threshold for the correlation between nearby SNPs.* An advantage of employing C+T is that it is an easily-interpretable approach: we can justify precisely why certain variants were or were not included in the PGS. The motivation for using a loose LD threshold is that – after adjusting for the 'double counting' of the signal of highly-correlated SNPs as we explain below – it allows us to achieve a higher performance than that obtained by using a low threshold which does not require such an adjustment.[10]

C+T is based on the LD Clumping procedure, which we now define. Given a set of GWAS summary statistics, begin by ordering SNPs in decreasing order of significance. Then, iteratively select the most significant SNP and discard all SNPs within a window of size 2ϖ around it if their LD with the variant selected is greater than a threshold ρ (see, e.g., [151]). We provide a more formal definition in Algorithm 3.1.

For each trait, therefore, we first run a GWAS in the *training set* to obtain estimated effect sizes and p-values for each SNP. We then run C+T (Algorithm 3.2) with a threshold $\rho = 0.9$ for the squared correlation between nearby SNPs, a window size of 1 Mb (i.e., $\varpi = 500$ kb) and 1000 different p-value thresholds τ equally spaced on the logarithmic scale between 10^{-9} and 0.05, computing estimated phenotypes in the *validation set.*

We evaluate the performance (measured through the coefficient of determination, R^2, from a linear regression of the trait – after regressing out covariates as we will

[10] Comparing the performance on the held-out test set of our final PGS (beginning with C+T with a loose LD threshold of 0.9 and then applying our iterative algorithm as described below) with that of a standard PGS obtained through C+T with a low LD threshold of 0.1 shows that the former achieves an average improvement of approximately 26% (standard deviation of 25.2%) compared to the latter across our set of 97 phenotypes with quantile-normalised residuals.

Algorithm 3.1: LD clumping.

Input:

For each SNP j in array $\mathcal{J} = (1, \ldots, J)$:

- A (log-transformed) p-value p_j for association with the trait;[a]
- The SNP's physical position q_j.

For each pair of distinct SNPs j, j' such that $|q_j - q_{j'}| \leq \varpi$:

- A measure of the LD between j and j', denoted $r^2_{j,j'}$.

Parameters:

- ϖ: threshold for distance between SNPs;
- ρ: threshold for LD (squared correlation) between SNPs.

Output: Array of filtered SNPs, denoted by $\mathcal{J}'$.

Initialisation

Sort array $\mathcal{J}$ by p_j *in decreasing order.*
Initialise empty array $\mathcal{J}'$.

Iterative procedure

while *array* $\mathcal{J}$ *is not empty* **do**
    Denote the first element of $\mathcal{J}$ by α.
    Remove α from $\mathcal{J}$ and add it to $\mathcal{J}'$.
    if $\mathcal{J}$ *is empty* **then**
        break
    end
    Generate an array $\mathcal{J}_\alpha$ containing all SNPs $j \in \mathcal{J} : |q_j - q_\alpha| \leq \varpi$.
    for $j \in \mathcal{J}_\alpha : j \neq \alpha$ **do**
        if $r^2_{j,\alpha} \leq \rho$ **then**
            Remove j from $\mathcal{J}$ and add it to $\mathcal{J}'$.
        else
            Remove j from $\mathcal{J}$.
        end
    end
end
return $\mathcal{J}'$

[a] One may use any other statistic representing the importance of a SNP – this could be derived from functional annotations, for example. All that is needed is an ordering of SNPs so that more important variants can be prioritised.

Algorithm 3.2: Clumping and Thresholding.

Input:

- An array $\mathcal{J} = (1, \ldots, J)$ of SNPs with effect sizes $\hat{\beta}_j$;
- All inputs required for LD Clumping (Algorithm 3.1) on array $\mathcal{J}$;
- Genotypes x_{ij} for each SNP j and each observation i.

Parameters:

- The same parameters as in Algorithm 3.1;
- A p-value threshold τ.

Output:

- Filtered array $\mathcal{L}$ of independent SNPs for which $p_j \geq \tau$;
- Array of estimated phenotypic values $\hat{y}_i$ for each individual i.

Initialisation

Discard all SNPs j for which $p_j < \tau$.

Main procedure

Run LD Clumping (Algorithm 3.1) on this filtered array of SNPs and obtain array $\mathcal{L}$ of (approximately) independent SNPs.

Compute estimated phenotype $\hat{y}_i$ for each i using $\mathcal{L}$, the array containing the subset of SNPs that survived LD Clumping, and effect sizes $\hat{\beta}_j$:

$$\hat{y}_i = \sum_{j \in \mathcal{L}} \hat{\beta}_j x_{ij}.$$

return

- $\mathcal{L}$;
- *Array of estimated phenotypes* $(\hat{y}_1, \ldots, \hat{y}_N)$.

describe in Section 3.6.1 – on the PGS in the *validation set*) of the PGS resulting from each of the 1000 p-value thresholds and select the best-performing threshold.

Finally, we compute a PGS from this set of SNPs using the coefficients from the original GWAS for the full set of samples (training, validation and test sets). This

gives us a starting set of SNPs and associated coefficients which we will improve upon through the iterative approach described below.

Step 2: Calibrating the coefficients of neighbouring SNPs with an iterative algorithm: C+T achieves our first goal of obtaining a PGS with a large selection of relevant SNPs. However, since it includes many SNPs in close proximity and high LD with one another (due to the loose R^2 threshold used for LD) and whose coefficients – obtained through univariate linear regressions – do not account for these correlation patterns, it suffers from significant bias due to double counting of neighbouring effects.

To correct for this, we propose an iterative algorithm that adjusts the effect sizes of each SNP in the PGS *until their additive signal is fully accounted for.* By this we mean that, after running the algorithm, if we regress the PGS out from the trait and perform a GWAS, none of the PGS SNPs should be genome-wide significant.

The iterative algorithm is formally described below (Algorithm 3.3) and we now summarise it more informally. We begin by regressing out the initial PGS from the trait and obtaining the resulting residuals (intuitively, these represent the signal in the trait that remains to be explained). We then take these residuals, run a GWAS *only on the SNPs included in the PGS* and build a new score through C+T as above (but now with a much more stringent LD threshold of $\rho = 0.1$ so as to only update approximately-independent SNPs). We add this new PGS to the existing one (their relative weights are allowed to differ), regress out this combined PGS and continue iterating in this manner until no significant SNPs are left. (Note that the set of SNPs included in the PGS is fixed from the start by the C+T algorithm with which we build the initial PGS – only the coefficients of the included SNPs are updated by the iterative algorithm.)

Algorithm 3.3: Iterative PGS algorithm (setup only).

Input:

- Initial PGS:
  - Array $\mathcal{L} = (1, \ldots, L)$ of SNPs that comprise the PGS;
  - Effect sizes $\hat{\beta}_l^0$ and physical positions q_l for each SNP;
  - A measure $r^2_{l,l'}$ of LD between pairs $(l, l') : |q_l - q_{l'}| \leq \varpi$;
  - Genotypes x_{il} for each SNP and each observation $i \in 1, \ldots, N$;
  - Array of estimated phenotypic values $\hat{S}_i^0$ for each individual i;
- Phenotype y_i for each observation (from which covariates have been regressed out);
- A partition of observations $i \in 1, \ldots, N$ into separate *training* and *validation* sets.

Parameters:

- LD Clumping parameters ϖ and ρ (for Algorithm 3.1);
- An ordered array $\mathcal{T} = (\tau_1, \ldots, \tau_T)$ of p-value thresholds for C+T (Algorithm 3.2);
- A threshold of significance υ for termination of the algorithm.

Output:

- Updated effect sizes $\hat{\beta}_l^R$ for each SNP after R iterations;
- Array of final estimated phenotypic values $\hat{S}_i^R$ for each individual i.

This algorithm is a slightly modified version of the simple boosting algorithm [152, 153] known as Least-Squares Boosting or LS-Boost [154], which is equivalent to gradient descent boosting on a squared-error (or L_2) loss function with a linear model as the 'base learner'. LS-Boost differs from the algorithm above in two main ways: it begins with all coefficients equal to zero (while we take the initial PGS as the starting point); and it updates only the most significant variable (strictly speaking, the variable that achieves the greatest reduction in squared error) at each step rather than several variables simultaneously as we do.

Algorithm 3.3: Iterative PGS algorithm.

Initialisation

$r \leftarrow 0$; `// start iteration counter`

Regress original trait y_i on the initial PGS $\hat{S}_i^0$ in the *training set* and obtain the resulting residuals $\hat{u}_i^0$:

$$y_i = \hat{\alpha} + \hat{\psi}^0 \hat{S}_i^0 + \hat{u}_i^0;$$

Run a GWAS of the residuals $\hat{u}_i^0$ in the *training set* for each SNP l and obtain coefficients $\hat{\delta}_l^0$ and corresponding (log-transformed) p-values p_l^0:

$$\hat{u}_i^0 = \hat{\alpha}' + \hat{\delta}_l^0 x_{il} + \hat{\varepsilon}_i^0.$$

Iterative procedure

while $\max_{l \in \mathcal{L}} p_l^r \geq \upsilon$ **do** `// while there are GWS SNPs in GWAS`

$r \leftarrow r + 1$; `// update iteration counter`

for $\tau_t : \tau_t \leq \tau^{r-1}$ **do** `// build intermediate PGS`

Build *temporary PGS* $\hat{S}_i'^r$ for all samples using C+T(ϖ, ρ);

Run the following regression in the *training set:*

$$y_i = \hat{\alpha} + \hat{\phi}^r \hat{S}_i^{r-1} + \hat{\psi}^r \hat{S}_i'^r + \hat{\varepsilon}_i^0;$$

Compute an updated PGS for the *validation set:*

$$\hat{S}_i^{r,\tau} = \hat{\phi}^r \hat{S}_i^{r-1} + \hat{\psi}^r \hat{S}_i'^r;$$

Evaluate the updated PGS's performance on the *validation set.*

end

The *new PGS* is $\hat{S}_i^r = \hat{S}_i^{r,\tau}$ for the best-performing threshold (τ^r);

Regress out new PGS from the trait in the *training set* $(y_i = \hat{\alpha} + \hat{\psi}^r \hat{S}_i^r + \hat{u}_i^r)$ and obtain the resulting residuals $\hat{u}_i^r$;

Run GWAS of residuals $\hat{u}_i^r$ in the *training set* $(\hat{u}_i^r = \hat{\alpha}' + \hat{\delta}_l^r x_{il} + \hat{\varepsilon}_i^r)$ and obtain coefficients $\hat{\delta}_l^r$ and p-values p_l^r.

end

return

- *Updated effect sizes $\hat{\beta}_l^R$ for each SNP where R is the final iteration;*
- *Array of final estimated phenotypic values $\hat{S}_i^R$ for each individual i.*

Despite its simplicity, LS-Boost has appealing theoretical properties: Freund et al. [155] have recently shown that this algorithm converges linearly to the multiple regression OLS solution. This property would be appealing in our context as it would guarantee that our algorithm would converge in a finite number of steps and that it would indeed exhaust all additive signal (since this is true by construction for multiple regression solutions as they ensure $\mathbb{E}(x_i\hat{\varepsilon}_i) = 0$ for every regressor x, where $\hat{\varepsilon}_i$ are the residuals).

Since our algorithm is slightly different, this theoretical property may not hold. However, our algorithm does converge in practice for all traits (real and simulated) for which we have run it and, since we only update variants that are not in high LD with one another simultaneously, our multiple updating seems unlikely to disrupt convergence (these modifications increase speed considerably in our context with a very large number of SNPs).

This iterative algorithm achieves our goal of exhausting all the additive GWS signal for a trait after the PGS is regressed out from it, which in turn allows us to focus on detecting *non-additive* GWS signal for the trait by testing for interactions with the PGS.

3.5.3 *Leave-one-chromosome-out PGS to counter LD-driven confounding*

To avoid potential confounding caused by LD between the SNP being tested for an interaction with the PGS and neighbouring SNPs in that PGS, after building a PGS we also compute 22 LOCO versions of that score that exclude SNPs from each chromosome in turn. Then, when fitting models like the one in equation (3.2), for each SNP we run both a model with the regular (i.e., full) PGS and the corresponding LOCO PGS (i.e., the PGS without any SNPs in the same chromosome as the SNP being tested for an interaction).

By testing for an interaction with the LOCO PGS, we eliminate LD as a source of possible confounding since none of the SNPs in the LOCO PGS can be in LD with the SNP being tested. Under our simplified theoretical scenario above, this would eliminate the general bias term (the third item in our list of potential sources of bias), thus preventing false-positive findings.

3.5.4 *Data-driven adjustment of the functional form of the PGS*

We noted previously that our goal is to detect interactions between specific pairs or larger groups of SNPs rather than global non-linearities where the additive effect of all relevant SNPs is transformed by some function f to yield the final phenotype (equation (3.28)).

To account for a possible global transformation of this kind, after obtaining an estimate of the PGS we will transform it by a function $\hat{f}$ with which we aim to approximate f. We perform this transformation as follows:

- We begin by sorting all individuals in our training set by their PGS in increasing order;
- We then divide them into bins of 1000 individuals. (To increase granularity at the ends of the PGS distribution, the first and last 1000 individuals are allocated into 20 bins of 50 samples;)
- For each bin, we compute the mean of the true phenotype (after regressing out covariates) for the individuals in that bin;
- Finally, we perform linear interpolation of these values (with the x-values given by the median PGS in each bin) and replace each individual's PGS with this 'mean-corrected' version.

The aim of applying this transformation is to account for global non-linear genetic effects on the phenotype through the PGS, and so reduce the likelihood of detecting false-positive interactions driven by such effects, as discussed in Section 3.4.4. In the simulation study that we describe in the next chapter, we will generate phenotypes with global non-linear effects (by applying sigmoidal transformations) and show that our method is robust to this potential source of spurious interactions.

3.5.5 *Testing whether dominance explains interaction signals*

Our primary interaction testing model, given in equation (3.2), does not include a dominance term for the SNP being tested. While in principle we could have included such a term, in practice this would increase the computational cost of fitting the model and, more importantly, for rare variants it would lead to a nearly-perfect correlation between the allele count x_{ij} and its square, further complicating estimation.

As an alternative, we test whether an interaction can be explained by a dominance term after we have obtained initial summary statistics for the SNP × PGS interaction model. We will include dominance terms as we filter these first results down to an independent set of loci as described below in Section 3.6.5.

3.6 DATA ANALYSIS PIPELINE

In this final section, we describe in more detail the computational implementation of our approach, including the main software packages used. We implemented our data analysis pipeline as a Snakemake workflow [156], with simple statistical operations undertaken in R [157] and most analyses of genetic data run with PLINK 2.0 [158, 159]. All analyses were performed on the Oxford Biomedical Research Computing Cluster (BMRC).

We provide a diagram of the pipeline in Figure 3.2 and now proceed to briefly describe the different steps involved.

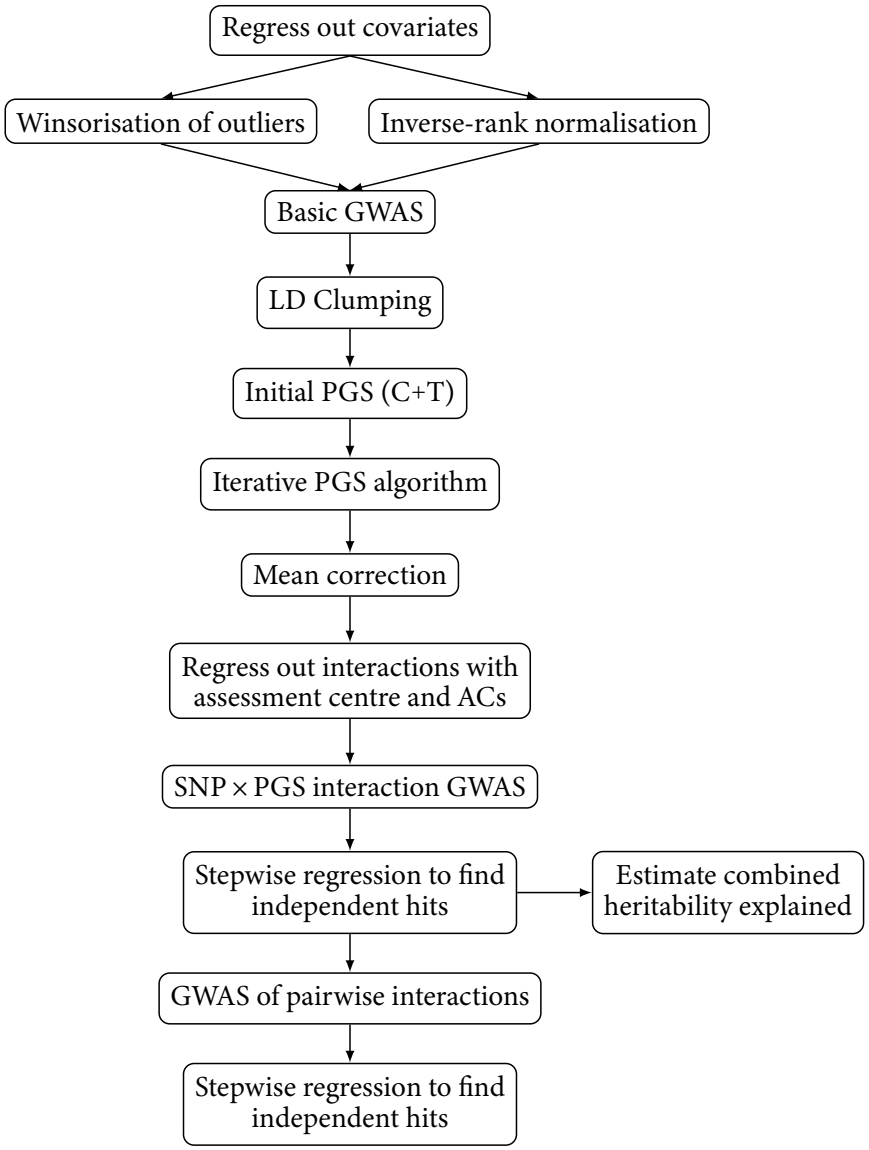


Figure 3.2: Diagram of data analysis pipeline.

3.6.1 *Pre-processing of traits*

Our main regression model will include a large number of covariates that we listed in Section 2.2. Because of this, fitting the model exactly as described in equation (3.2) would substantially increase the computational cost of the interaction GWAS. Therefore, and as is common practice in the field, we begin by first regressing these out from the trait and then proceed to analyse the resulting residuals.[11]

After regressing out the covariates from each trait, we then process the resulting residuals in two alternative ways: as described in Section 3.5.1, we either perform Winsorisation at the 1st and 99th percentiles of their empirical distribution; or we simply transform them into a Normal distribution by applying a rank-based inverse Normal transformation. These operations are undertaken in R, where the `orderNorm` function of the `bestNormalise` [146, 147] package is used for INT.

3.6.2 *GWAS*

We continue by running a basic GWAS, i.e., we regress the residuals obtained in the previous step on each SNP in turn, without including any interaction terms or PGSs in these regressions. This first GWAS provides the summary statistics used to build the initial PGS in the next step.

The UKB imputed genetic data is provided in BGEN format [160]. We chose to use PLINK 2.0 [158, 159] for operations involving genetic data whenever possible due to its versatility, stability and speed. Since PLINK cannot directly operate on BGEN files, we first convert the imputed genetic data from the original BGEN format

[11] By the Frisch-Waugh-Lovell Theorem (see, e.g., [137, Section 3.3]), this procedure is equivalent to fitting the larger model with all covariates as long as these are uncorrelated with the explanatory variables of interest – in our case, the allele count of the SNP, the PGS and their product.

into PGEN, PLINK 2.0's native format. The basic GWAS is then run using PLINK 2.0's `--glm` function.

3.6.3 *Building the PGS*

To build the PGS, we begin by running C+T on the results from the basic GWAS. Since an LD Clumping function had not yet been implemented in PLINK 2.0 at the time of analysis, we instead use the `bed_clumping` function in the R package `bigsnpr` [161], which can work directly on the original BGEN files.

We run LD with two alternative thresholds $\rho \in \{0.1, 0.9\}$ for the squared correlation between neighbouring SNPs. The results for $\rho = 0.9$ are used as the starting point for the iterative step as described in Section 3.5.2 while the results for $\rho = 0.1$ give us a set of approximately independent SNPs within each PGS.

Having obtained the LD Clumping results, we then build a series of PGSs on the *validation set* for 1000 different p-value thresholds τ equally spaced on the logarithmic scale between 10^{-9} and 0.05 using PLINK's `--score` function, assess their performance on this sample set and, finally, extend the best-performing score to all samples as described above.

We proceeded by running the second (iterative) step in our PGS construction algorithm. Since this consists of multiple iterative steps each involving GWAS, C+T and intermediate PGS construction, we found PLINK was not fast enough and use the `bigsnpr` and `bigstatsr` R packages [161] instead.

3.6.4 *Testing for interactions*

To test for interactions between each SNP and a trait, we again use PLINK's `--glm` function to fit linear regressions of the residuals from regressing out covariates from the trait on an intercept, the minor allele count for the SNP, the PGS and an interaction

term between the SNP and the PGS. We fit each regression twice, once with the LOCO version of the PGS (which is robust to LD as a confounder) and once with the full PGS.

The residuals used in this stage are different from those used for the basic GWAS: from those first residuals we also regress out interaction terms between the (full) PGS and each of the assessment centre and ancestry component (AC) variables described in the previous chapter. This is to account for any potential 'generalised interaction effects' between an individual's ancestry (measured directly through the ACs and also correlated with the assessment centre where each individual had their data collected, since these tend to be located in the same region where they reside) and their genome-wide additive effects for a trait as aggregated in the PGS.

3.6.5 *Identifying independent interaction hits*

GWASs identify association peaks consisting of tens of closely-linked SNPs in a single region of the genome that is not usually split by recombination. It is generally accepted that only one or a few of these SNPs will have a 'causal' effect on the trait (meaning that there is a biological mechanism – such as disruption of a protein or regulation of gene expression, for example – linking the mutation to the phenotype, rather than the SNP simply being highly correlated with another such SNP) and many methods have been developed that attempt to distil or 'fine-map' these sets of associations into smaller groups of likely causal variants (see [162, 163] for recent reviews).

After obtaining summary statistics from our interaction GWAS, we filter these down to a set of independent associations for each trait. A first use of this filtering is to provide us with an estimate of the number of independent associated loci which is not upwardly biased by LD as described in the previous paragraph. A second

advantage of this approach in our setting is that it prioritises specific SNPs for further investigation: in the section below, we describe how we proceed to search for specific SNPs within the PGS that drive each observed interaction.

For fine-mapping in a standard GWAS, it can be helpful to use functional information on associated SNPs to prioritise those which are likely to be causal, possibly integrating these formally into a statistical fine-mapping approach. However, in our context, the priority is to increase power to detect loci in the PGS driving the interaction in a subsequent step. Because of this, we did not use functional information or sophisticated fine-mapping methods to filter loci with statistical evidence of an interaction but rather chose the most-significantly-associated SNP at each locus.

In more detail, we employ a stepwise regression procedure of variable selection to identify independent interacting SNPs. This consists of the following steps (Algorithm 3.4):

1. Sort all GWS SNPs in decreasing order of significance;
2. Beginning with the most significant SNP, run a regression of the residuals on an intercept, the SNP's minor allele count, *its square* (to account for a possible dominance effect), the PGS and an interaction between the SNP and the PGS. Keep the SNP as an independent association if the interaction term is genome-wide significant;
3. Iteratively add the remaining SNPs one-by-one together with their squares and interactions with the PGS, keeping those with significant interactions after more significant SNPs have already been accounted for and discarding the rest.

We begin by running this procedure for the SNPs that interact with the LOCO PGS, using this PGS in the regression at each step. We then continue with the same procedure for the SNPs interacting with the full PGS *while including the independent*

Algorithm 3.4: Stepwise regression.

Input:
For each SNP j in array $\mathcal{J} = (1, \ldots, J)$:
- A (log-transformed) p-value p_j for the interaction term between the SNP and the PGS in a regression with the trait;
- Genotypes x_{il} for each observation $i \in 1, \ldots, N$.

Phenotypic residuals y_i for each observation.

Parameters: A p-value threshold of significance τ.

Output: Array of independent interacting SNPs.

Initialisation
Remove from $\mathcal{J}$ all SNPs j for which $p_j < \tau$.
Sort array $\mathcal{J}$ by p_j *in decreasing order.*
Initialise empty array $\mathcal{J}'$.

Main procedure
while *array* $\mathcal{J}$ *is not empty* **do**
  Denote the first element of $\mathcal{J}$ by α and remove it from $\mathcal{J}$.
  Run a regression of y_i on:
    - An intercept term;
    - All SNPs j in $\mathcal{J}'$, their squares and their products with the PGS;
    - The SNP α, its square and its product with the PGS.
  if *the p-value of the interaction term between* α *and the PGS is* $\leq \tau$ **then**
    Add α to $\mathcal{J}'$.
  end
end
return $\mathcal{J}'$

LOCO hits and their interactions with the full *PGS in each regression.* This produces a set of independent SNPs interacting with the full PGS that is disjoint from the set of SNPs interacting with the LOCO PGS: since LOCO associations are preferable (as they are not confounded by LD), we prioritise these and only include interactions with the full PGS that cannot be accounted for by a LOCO interaction.

3.6.6 *Estimating heritability explained*

A natural question when new genetic associations are identified is what is the fraction of heritability for which they account. *Broad-sense* heritability is normally defined as the proportion of the variation of a trait in the population that can be attributed to genetic factors [164], while *narrow-sense or additive* heritability can be defined as the proportion of variation that can in principle be explained by additive genetic effects [40].

One motivation to study genetic interactions is that these may explain additional (broad-sense) heritability (although it is thought that they are unlikely to account for a significant share of population variation [60]). We therefore would like to provide an approximate estimate of how much heritability is explained by the independent interaction hits we find.

Traditional approaches to (broad-sense) heritability estimation are based on correlations of phenotypes across family members [165], with perhaps the most common method being a comparison between monozygotic and dizygotic twins [40]. More recent methods[12] focus on narrow-sense heritability, and a prominent example among them is LD Score Regression [123]. However, neither classic family-based study designs nor modern methods for (additive) heritability estimation are applicable in our setting in which our sample is comprised of unrelated individuals and we wish to estimate the heritability accounted for by *interactions* between mutations rather than additive genetic effects.

We propose a simple method for obtaining an approximate estimate of the heritability explained by the *combined* set of independent interaction hits identified for

[12] See Evans et al. [166] and Barry et al. [167] for two recent reviews.

each trait that is based on measuring improvement in predictive performance on the held-out test set.

We first regress a trait on its PGS, main terms for the independent interaction hits and their squares (to account for possible dominance effects), and the SNP × PGS interactions in the *training* set, storing the coefficients of the different variables. We then evaluate the predictive performance (measured through the coefficient of determination) in the *test* set of the PGS and these additional terms *while keeping the coefficients fixed* to their value in the regression in the training set (so as to avoid overfitting through additional degrees of freedom). Finally, this performance is then compared to that of only the PGS in the held-out test set, and an estimate of the heritability is obtained by taking the difference of the two.

We make two observations about the procedure outlined above. First, it can lead to negative estimates of heritability since the coefficients of the additional terms included in the 'full model' with interactions are fixed to the values obtained in the training set – there are no additional degrees of freedom compared to a model with only the PGS. Second, because of this fixing of coefficients and of the fact that the PGS was itself computed based on the training set (implying that the coefficients of terms involving the PGS computed in the training set may not translate well to the test set due to overfitting in the former), this approach can be thought of as yielding a *lower bound* on the true value of the heritability.

Two obtain an *upper bound*, we repeat the procedure above but using only the test set and letting the coefficients be determined from the regression in this set. Given the increase in degrees of freedom in this model and the fact that estimating these coefficients requires 'seeing' the phenotypes of the test set (something impossible in a real-life prediction scenario), we think of this as an upper bound.

Finally, we take the mid-point of the lower and upper bounds computed to get a final estimate of heritability, which we divide by the performance of the PGS on the held-out test set to obtain a relative measure of improvement in predictive performance.

3.6.7 *Identifying the targets of interaction signals*

For each independent interaction hit identified in the previous step we wish to find, *if possible,* the SNP(s) *within* the PGS driving the interaction signal, or potentially other SNPs in the genome which also interact with this hit. As discussed when motivating our approach, we believe it is likely that some SNPs may interact with a large group of other SNPs. For such SNPs, it may not be possible – and nor would it be sensible – to reduce the interaction to a pairwise one, or one involving a small set of SNPs. However, pairwise or other interactions involving a subset of SNPs that have a large (interaction) effect size may be captured by our interaction test, and for these it may be possible to identify the SNPs in the PGS that drive them.

To attempt to find which locus or loci in the PGS lie behind each independent interaction hit, for each of these hits we run a simple GWAS of pairwise interaction with every SNP in the genome (not only those in the PGS) one-by-one. We then use stepwise regression again to filter these down to a final set of pairwise interactions, including terms for the squares of the genotypes of all SNPs considered, as before. For many cases, we do not find GWS 'partners' for the independent SNP–PGS hits through this analysis; however, several independent hits can be linked to specific SNPs, and these pairwise interactions can provide interesting biological insights, as we will discuss in Chapters 5 and 6.

4 Simulations

In this chapter, we describe the simulation study we have undertaken to assess our method's performance in settings where the underlying genetic architecture is known.

In the description of our approach in a previous chapter, we presented some simple mathematical derivations that gave us a first glimpse into its potential pitfalls (specifically, sources of potential false positives) and power to detect associations. However, given the complexity of the data analysis pipeline, deriving precise mathematical properties would be very challenging and indeed we were only able to derive basic properties for a toy model. We have therefore built an extensive set of simulation scenarios to assess performance under realistic settings relevant to our analysis of real data.

When assessing a method for statistical association testing, it is natural to examine two different properties: the potential of the method to yield false-positive findings (type I error) and its power to detect true effects (the avoidance of type II error). As discussed in our exposition of the method, we are chiefly concerned with false positives: given how challenging detecting interactions has proven and the well-documented possibility of arriving at spurious results [148–150], we wish to be confident that our findings are likely to reflect true statistical interactions.

We have noted previously that the astronomical search space of potential interactions has limited the statistical power of methods for their detection. While our

approach has been designed to increase power – by testing for interactions with groups of SNPs – we are not able to fully circumvent this challenge and therefore expect our method's power to be only modest.

The main goal of this simulation study when it comes to assessing statistical power is therefore to *understand which types of interactions we have the greatest power to detect:* as explained in more detail below, we expect interactions with stronger effects or involving larger numbers of variants to be easier to detect, and we will indeed find this to be the case. This can then guide our interpretation of results in the aggregate: the lack of evidence for certain kinds of interactions – namely those with small effect sizes or involving few SNPs – need not imply that they do not exist, since our method is unlikely to detect them in any case.

To evaluate the occurrence of errors of types I and II, we build two classes of simulated traits based on the same dataset of imputed genotypes used in the subsequent analysis of real traits:

1. Purely-additive traits with a wide range of genetic architectures (varying in number and positioning of causal loci, their effect sizes, possible non-linear monotonic transformations of the combined additive genetic effects, and heritability) – we term these *'null simulations';* and
2. Traits that are partly determined by interactions, which we denote by *'positive simulations'.* For these, we keep the parameters relating to causal loci and their heritability fixed to mid-range values and instead explore variation in the number, strength and range of interactions as explained in detail below.

We now describe in turn the two classes of simulated traits and the empirical results from applying our method to them.

4.1 NULL SIMULATIONS

4.1.1 *Setup*

We begin by building a series of traits without any interactions. This requires selecting a set of SNPs, indexed by $k = 1, \ldots, K$, attributing them effect sizes β_k and obtaining phenotypes y_i for each observation $i = 1, \ldots, N$, given genotypes x_{ik}, as follows:

$$y_i = \sum_{k=1}^{K} \beta_k x_{ik} + \varepsilon_i,$$

where ε_i is a stochastic error term drawn independently for each observation from a zero-centred Normal distribution whose variance is set so as to obtain a desired level of (additive) heritability.

Since we do not want to mistake a global non-linear functional form of the genetics–trait mapping (as described in Section 3.4.4) for a localised interaction, we also build two additional versions of each trait in which we apply a non-linear function $f : \mathbb{R} \to \mathbb{R}$ as follows:

$$y_i = f\left(\sum_{k=1}^{K} \beta_k x_{ik}\right) + \varepsilon_i, \quad \text{or} \tag{4.1}$$

$$y_i = f\left(\sum_{k=1}^{K} \beta_k x_{ik} + \varepsilon_i\right), \tag{4.2}$$

i.e., we either transform the additive genetic effects and then add Normal noise, or we first add noise and then apply the transformation (we expect this latter transformation to not be a challenge to our approach since this non-linearity is not affected by noise and should thus be easily reversed by quantile normalisation of the trait). We choose f to be a scaled sigmoidal function with the scaling factor given by the standard deviation of the additive genetic effect: denoting the observed standard deviation of $\sum_{k=1}^{K} \beta_k x_{ik}$ across observations i by $\hat{\sigma}_S$, we set $f(x) = 1/[1 + \exp(-x/\hat{\sigma}_S)]$.

We make a series of simulated phenotypes by setting different parameters along five axes:

1. Number of causal SNPs;
2. Positioning of causal SNPs relative to one another;
3. Effect sizes of these SNPs;
4. Possible global, non-linear transformation of the additive genetic effects; and
5. Level of heritability,

as we now describe.

Number of causal SNPs: We set 100, 1000 or 10,000 causal SNPs to cover a realistic range for the number of independent associations observed for real traits.

Positioning of causal SNPs: For each number of causal SNPs, we either place these uniformly at random along the genome (i.e., any of the 12.7 million SNPs that survived variant filtering is equally likely to be chosen) – we term this a *'regular architecture'* – or we place half of them into clusters while the remaining half are placed uniformly at random again – we term this a *'clustered architecture'.*

More precisely, a clustered architecture is obtained as follows:

1. We first place (# causal SNPs)/(5 × 2) regions of physical length 1,000 kb uniformly at random along the genome into which we will place (# causal SNPs)/2 SNPs, resulting in an average of five variants per region;
2. We then draw the number of causal variants to place into each region from a Multinomial distribution with equal probabilities and thus place half the number of causal variants into these regions;
3. Finally, we place the remaining half of the causal variants at random anywhere on the genome.

Effect sizes: All effect sizes are drawn independently from a standard Normal distribution.

Since different causal SNPs have different frequencies, without further adjustment this setup would lead to rarer SNPs having a lower expected contribution to the heritability of the trait.[1] This is generally seen as undesirable because, under a theoretical model that assumes that the average value of a phenotype in a population is at its optimum, many newly arising mutations will have a negative effect on fitness and therefore be selected against and so kept at low frequency.

To incorporate this expected negative relationship between effect size and the allele frequency, it is common practice in the literature to multiply the effect sizes obtained from a standard Normal distribution by $[2p_k(1-p_k)]^{-0.5}$, a transformation that completely nullifies the positive relationship between allele frequency and heritability contribution. This approach has been adopted in the simulation studies undertaken for prominent recent methods in the field of statistical genetics [168–170].

However, a recent empirical assessment of variant-level contributions to heritability where these are modelled as being proportional to $[2p_k(1-p_k)]^{1+\alpha}$ suggests that a value of $\alpha = -0.25$ provides the best fit to real phenotypes [171].[2] We therefore choose to set two possible values for α of 0 (in which case we do not transform the effect sizes) and -0.5 (where we multiply the effect sizes by $[2p_k(1-p_k)]^{-0.25}$) with the aim of capturing the full realistic range for this parameter.

Possible non-linear transformation of the additive genetic effects: For each combination of number of causal SNPs, their positioning and the scaling of their effect sizes,

[1] The heritability contribution of a particular SNP x_k can be written as $\mathrm{Var}(\beta_k x_k) = \beta_k^2 \mathrm{Var}(x_k) = 2\beta_k^2 p_k(1-p_k)$, where β_k is fixed and p_k is the MAF of x_k, since x_k is binomial.

[2] In their paper, Speed et al. [171] model heritability as depending also on the level of LD between a SNP and its neighbours and the confidence in the accuracy of its imputation, factors which we ignore in this exposition.

we build three traits: one without any global transformation ($f(z) = z$), one where the transformation is applied before adding noise as in equation (4.1) and one where it is applied after adding noise as in equation (4.2). The function f is a scaled sigmoid as described above.

Heritability: Heritability is set to either 30% or 60%. To achieve this, we simply add to each individual's 'PGS' (i.e., their total additive genetic effect obtained above) an error term drawn from a zero-mean Normal distribution whose variance is set so that the ratio of the variance of this PGS to that of the resulting trait is 30% or 60%.[3]

To summarise, we have *three* possible numbers of causal SNPs, positioned in *two* possible ways, with *two* possible sets of effect sizes, with either no transformation or one of two alternative transformations (*three* possibilities) and *two* possible levels of heritability: this yields 72 null traits in total. After regressing out the covariates, the resulting residuals will be either quantile normalised (QN) or Winsorised, and therefore each of the 72 traits will be tested for interactions under these two alternative forms.

4.1.2 *Results*

Having conducted interaction testing on all 72 traits with both Winsorised and QN residuals, we now list the number of GWS hits in each GWAS, the resulting false-positive rate (FPR; number of GWS hits divided by the total number of SNPs tested, approximately 12.7 million) and the number of independent hits from among the GWS hits obtained through stepwise regression. Note that, for each trait, we run two GWASs, one with the LOCO PGS and the other with the full score.

[3]The Normal distributions are i.i.d. across individuals.

Table 4.1: False positives in null simulations with $\alpha = 0$, a heritability of 0.3 and QN residuals.

Each row shows data for one of 18 traits, with the first three columns containing trait information (number of causal SNPs, architecture – regular or clustered – and transformation – none, before adding noise or after adding noise). The GWS columns show the number of genome-significant hits in the GWAS, the FPR columns the FPR (number of GWS hits divided by the total number of SNPs tested, 12.7 million) and the Ind. columns the number of independent hits. The combined FPR across all 18 traits is shown in the last row for the LOCO and full PGS GWASs separately.

SNPs	Arc.	Tr.	LOCO PGS			Full PGS		
			GWS	FPR	Ind.	GWS	FPR	Ind.
100	Reg.	–	2	1.58×10^{-7}	1	2	1.58×10^{-7}	0
100	Reg.	Bef.	4	3.15×10^{-7}	2	1	7.88×10^{-8}	0
100	Reg.	Aft.	0	0	0	0	0	0
100	Cl.	–	0	0	0	0	0	0
100	Cl.	Bef.	0	0	0	0	0	0
100	Cl.	Aft.	1	7.88×10^{-8}	1	1	7.88×10^{-8}	0
1K	Reg.	–	0	0	0	0	0	0
1K	Reg.	Bef.	0	0	0	0	0	0
1K	Reg.	Aft.	0	0	0	0	0	0
1K	Cl.	–	0	0	0	0	0	0
1K	Cl.	Bef.	0	0	0	0	0	0
1K	Cl.	Aft.	0	0	0	0	0	0
10K	Reg.	–	0	0	0	0	0	0
10K	Reg.	Bef.	0	0	0	0	0	0
10K	Reg.	Aft.	0	0	0	0	0	0
10K	Cl.	–	1	7.88×10^{-8}	1	2	1.58×10^{-7}	0
10K	Cl.	Bef.	1	7.88×10^{-8}	1	6	4.73×10^{-7}	0
10K	Cl.	Aft.	1	7.88×10^{-8}	1	1	7.88×10^{-8}	0
Combined FPR				4.38×10^{-8}			5.69×10^{-8}	

Table 4.1 shows results for all combinations of number of causal loci, architecture and possible transformation of the additive genetic effects for traits with an α parameter of 0, a heritability of 0.3 and QN residuals. We see that the majority of traits have no false positives, while the ones that do have only a few.

We report also the combined FPR across all traits in the table, which is close the GWS p-value threshold as expected for a well-calibrated test in simulation scenarios

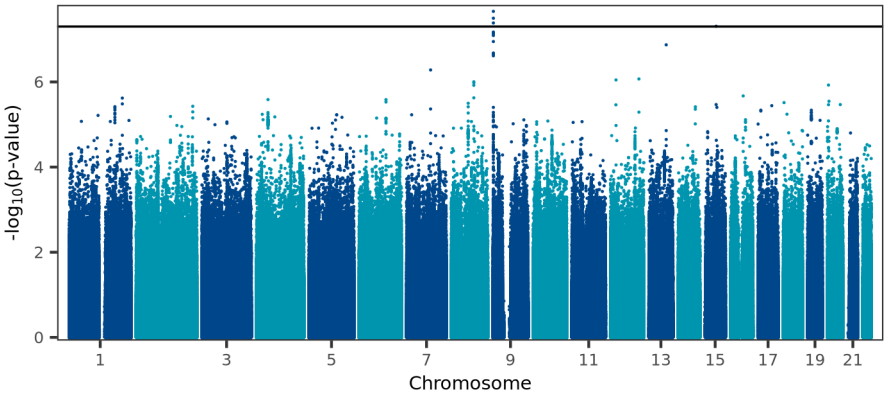
with no true interaction signal. Note that the combined FPR is slightly lower for the LOCO tests than for those with the full PGS as may be anticipated due to the potential for local LD to lead to false positives as discussed previously.

We now examine the results of the GWAS for the trait on the second row of Table 4.1, which has 100 causal loci, a regular architecture and a sigmoid transformation applied before noise (Figure 4.1). We see the peak with four genome-wide significant (GWS) SNPs on chromosome nine in the LOCO PGS GWAS and the corresponding deviation in the QQ plot below. In the full PGS GWAS, this peak does not reach genome-wide significance.

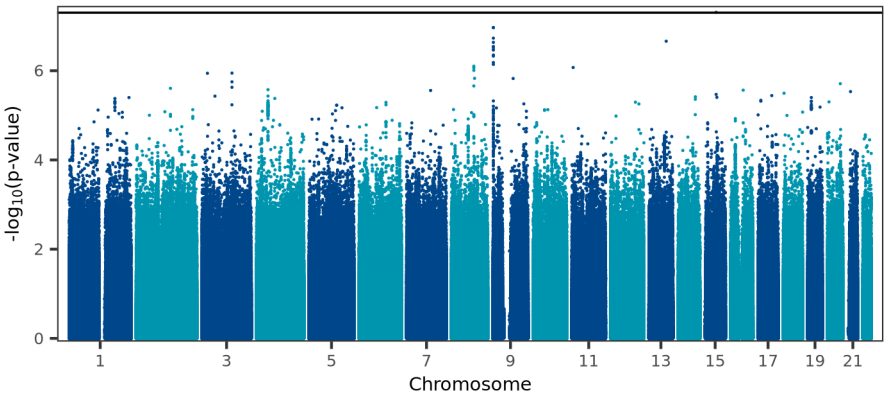
Looking at the Winsorised version of this trait in Figure C.1 in the appendix, we see that the situation is largely the same, suggesting general concordance between results from either method of transforming residuals for subsequent analysis (although for other traits there will be important differences between the two, with the QN version generally being less prone to false positives, as we will see below).

We continue by inspecting results from the opposite combination of parameters than in the previous table: Table 4.2 shows false positive hits for traits with an α parameter of -0.5, a larger heritability of 0.6 and Winsorised residuals. (Tables for the other six combinations of these three parameters are included in Appendix C.)

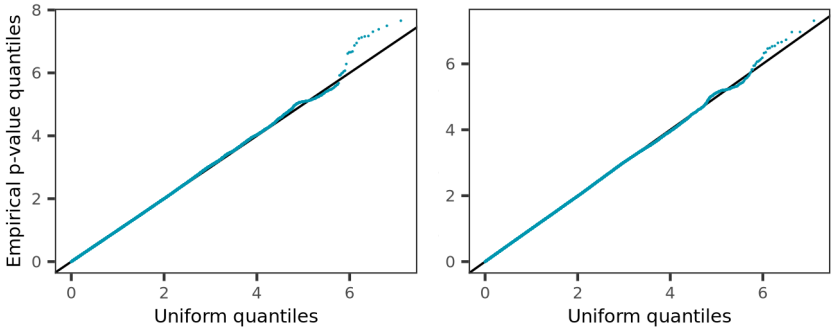
While the overall results are similar to the previous set of traits, one now stands out as having a high number of false positives: the trait with 100 causal SNPs, a clustered architecture and a sigmoid transformation applied before adding noise has 81 GWS hits in the interaction GWAS with the LOCO PGS and 62 in the full PGS GWAS, leading to FPRs of 6.38×10^{-6} and 4.89×10^{-6}, respectively. This is one of three traits in the full set of 144 (72 traits each analysed with two types of residuals) that show a high rate of false positives. We will now examine in more detail the one trait in these three that has the highest FPR and will return to this first trait below.



(a) LOCO PGS



(b) Full PGS



(c) LOCO PGS (d) Full PGS

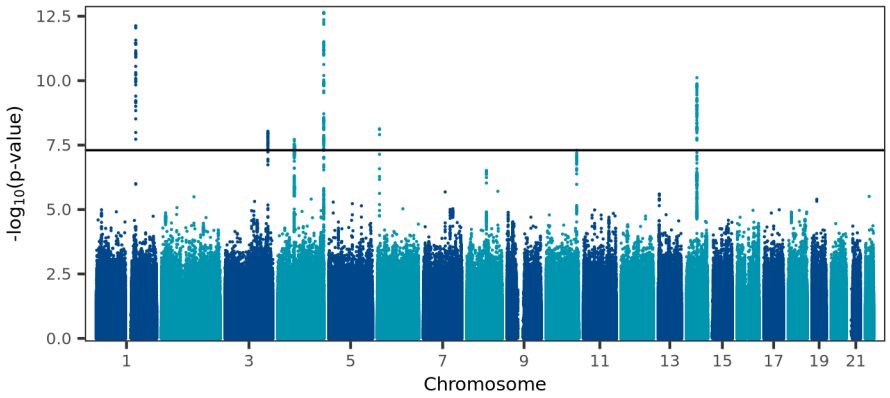
Figure 4.1: Manhattan and QQ plots of results from the SNP × PGS interaction GWASs with the LOCO or full PGSs: null simulation with 100 causal loci, a regular architecture, a sigmoid transformation applied before adding noise, an α parameter of 0, a heritability of 0.3 and QN residuals. The horizontal line represents the usual genome-wide significance threshold of 5×10^{-8}.

Table 4.2: False positives in null simulations with $\alpha = -0.5$, a heritability of 0.6 and Winsorised residuals.

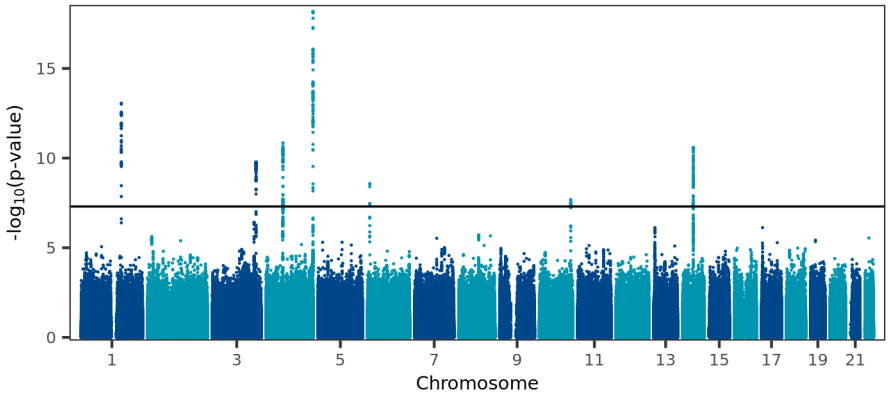
SNPs	Arc.	Tr.	LOCO PGS			Full PGS		
			GWS	FPR	Ind.	GWS	FPR	Ind.
100	Reg.	–	0	0	0	0	0	0
100	Reg.	Bef.	0	0	0	0	0	0
100	Reg.	Aft.	0	0	0	0	0	0
100	Cl.	–	4	3.15×10^{-7}	1	0	0	0
100	Cl.	Bef.	81	6.38×10^{-6}	6	62	4.89×10^{-6}	2
100	Cl.	Aft.	0	0	0	0	0	0
1K	Reg.	–	0	0	0	0	0	0
1K	Reg.	Bef.	0	0	0	0	0	0
1K	Reg.	Aft.	0	0	0	0	0	0
1K	Cl.	–	0	0	0	0	0	0
1K	Cl.	Bef.	1	7.88×10^{-8}	1	1	7.88×10^{-8}	0
1K	Cl.	Aft.	0	0	0	0	0	0
10K	Reg.	–	0	0	0	0	0	0
10K	Reg.	Bef.	0	0	0	0	0	0
10K	Reg.	Aft.	0	0	0	0	0	0
10K	Cl.	–	0	0	0	0	0	0
10K	Cl.	Bef.	0	0	0	0	0	0
10K	Cl.	Aft.	0	0	0	0	0	0
Combined FPR				3.76×10^{-7}			2.76×10^{-7}	

The trait with most concerning results across all the null simulations has 100 causal SNPs, a clustered architecture, a sigmoid transformation applied before noise is added, an α parameter of 0, a heritability of 0.6 and Winsorised residuals. As shown in Table C.5, for this trait we have 311 GWS hits in the LOCO GWAS and 494 in the full GWAS. (Note that no other trait in this table has any false positive hits at all.)

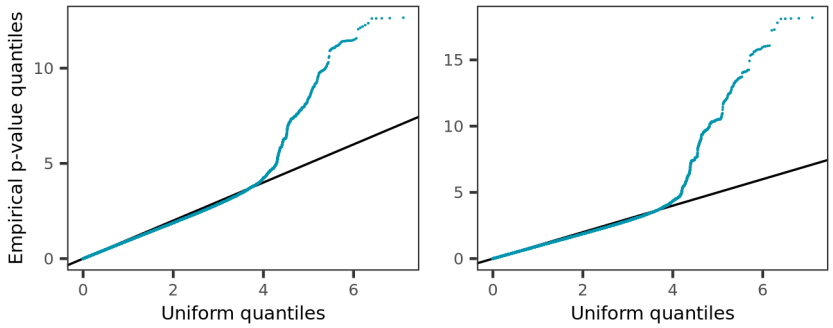
Looking at the Manhattan and QQ plots for this trait shown in Figure 4.2, we see that there are several clear peaks in different chromosomes clearly reaching significance. These are reflected in the QQ plots where we see large deviations from the expected distribution of p-values under a null uniform distribution as would be expected for a well-calibrated test.



(a) LOCO PGS



(b) Full PGS



(c) LOCO PGS (d) Full PGS

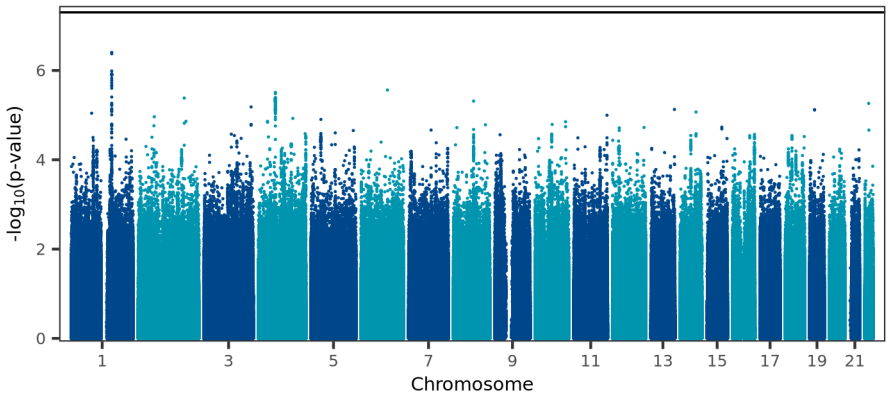
Figure 4.2: Manhattan and QQ plots of results from the SNP × PGS interaction GWASs with the LOCO or full PGSs: null simulation with 100 causal loci, a clustered architecture, a sigmoid transformation applied before adding noise, an α parameter of 0, a heritability of 0.6 and Winsorised residuals.

We note, however, that these results are for Winsorised residuals and that, in general, we expect results for QN traits to be more conservative and therefore less prone to false positives. (This is because the rank-based inverse Normal transformation is more wide-reaching in its impact in changing all the values of the variable to which it is applied, unlike Winsorisation which affects only outliers at the tails of the distribution.) And indeed, when we look at the results for the trait with the same architecture but with QN residuals shown in Figure 4.3, we see that there are no false positives at all.

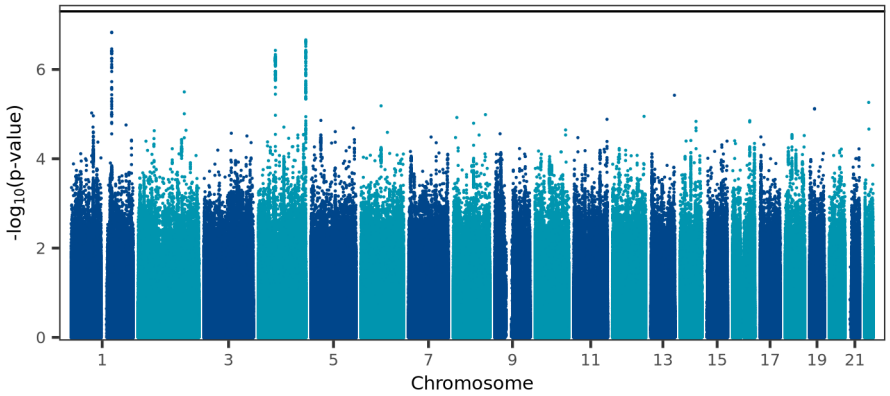
We mentioned previously that there are two other traits with a large number of false positives, one of which appears in Table 4.2 above. These other two traits again have 100 causal loci, a clustered architecture and a sigmoid transformation before noise is added; their α parameter is -0.5 and their heritability 0.6.

Unlike in the worst-case example above, here the Winsorised and QN residuals still individually yield a sizeable number of false positives. In the Winsorised case, there are 81 false positives in the LOCO GWAS and 62 in the full; in the QN case, there are only 3 false positives in the LOCO GWAS but 38 in the full. The Manhattan and QQ plots for these GWASs are shown in Figures C.2 and C.3.

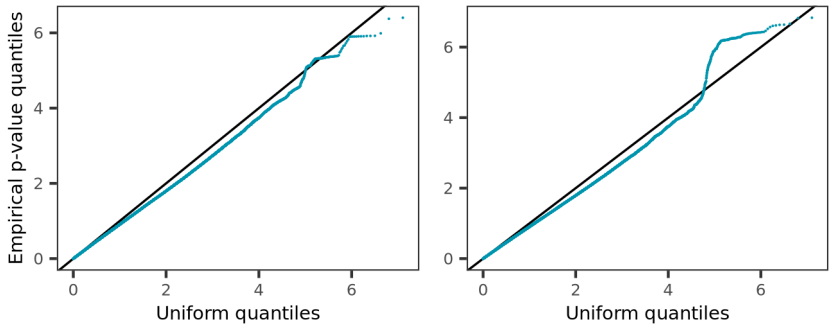
While, in this scenario, the quantile normalisation of residuals does not fully remove false positives, interaction testing with the LOCO PGS yields a very low number of spurious findings. Therefore, while this example shows that quantile normalisation is not guaranteed to yield an acceptably low FPR, fixing potential issues in the Winsorised version, it may reassure us in that the most conservative of our approaches (i.e., QN residuals tested with a LOCO PGS) is not generally prone to type I errors.



(a) LOCO PGS



(b) Full PGS



(c) LOCO PGS (d) Full PGS

Figure 4.3: Manhattan and QQ plots of results from the SNP × PGS interaction GWASs with the LOCO or full PGSs: null simulation with 100 causal loci, a clustered architecture, a sigmoid transformation applied before adding noise, an α parameter of 0, a heritability of 0.6 and QN residuals.

Finally, we note that the three traits across 144 for which we observe a large number of false positives have a very similar architecture which we believe makes them prone to spurious results. In particular, all three traits have:

- A small number of causal loci (100);
- A clustered architecture (in which 50 of the 100 loci are placed into small regions of 1 kb with an average of five SNPs in each);
- A non-linear transformation applied before noise is added (which makes it challenging to 'undo' it); and
- High heritability (leading to large effect sizes for the causal loci).

We included such traits in our simulations to assess the performance of our method under unfavourable circumstances and are reassured to see that it is only in this worst-case scenario that it leads to more than a few occasional false positives. We expect that most of the real traits that we analyse will not be characterised by these four problematic properties simultaneously and are therefore confident that the results from our analysis of real data are generally not spurious.

We conclude this exposition of the results from the null simulations by aggregating in Table 4.3 the FPRs across all traits, making a distinction only between QN and Winsorised residuals and LOCO and full PGSs.

Table 4.3: FPR for all null simulations by type of residuals and PGS.

	LOCO PGS	Full PGS
QN residuals	3.61×10^{-8}	7.44×10^{-8}
Winsorised residuals	4.48×10^{-7}	6.19×10^{-7}

As suggested above when looking at individual traits and aggregating FPRs across smaller groups, we see here in greater generality that interaction testing of QN residuals against the LOCO PGS yields the lowest FPR (which is lower than the usual

genome-wide significance threshold). We will therefore be most confident in the results from this test in our analysis of real data. We note also that the LOCO PGS yields fewer false positives than its full counterpart and that QN residuals similarly perform better than Winsorised residuals as expected.

Finally, we break down the rates in the table above by the type of transformation that is applied to the additive genetic component: either no transformation or a scaled sigmoid applied either after or before noise is added. These results are shown in Table 4.4.

Table 4.4: FPR for null simulations by type of residuals and PGS, split by type of transformation applied.

(a) Traits without non-linear transformations.

	LOCO PGS	Full PGS
QN residuals	3.28×10^{-8}	2.95×10^{-8}
Winsorised residuals	1.31×10^{-8}	3.28×10^{-9}

(b) Sigmoid transformation after adding noise.

	LOCO PGS	Full PGS
QN residuals	1.97×10^{-8}	6.57×10^{-9}
Winsorised residuals	0	0

(c) Sigmoid transformation before adding noise.

	LOCO PGS	Full PGS
QN residuals	5.58×10^{-8}	1.87×10^{-7}
Winsorised residuals	1.33×10^{-6}	1.86×10^{-6}

From this table we see that it is only when applying a sigmoid transformation *before* noise is added that we obtain larger FPRs that are above the threshold for genome-wide significance. This is as we might expect: if there is no transformation

and the trait is fully additive, our method infers very few false positives; similarly, a transformation that is applied after noise is added can be easily reversed; therefore it is only if noise is added on top of the transformation that our method does less well.

In summary, our method appears quite robust to false positives, breaking only in the worst possible scenario with few causal loci that are clustered, an impactful non-linear transformation to which noise is added – making its reversal difficult – and high heritability. Across these results, we see that the test with QN residuals and the LOCO PGS is the most reliable; however, other tests do not generally yield a high FPR and may capture useful signal, therefore needing only more careful interpretation.

4.2 SIMULATIONS WITH INTERACTIONS

We proceed by examining a series of simulated traits whose architecture includes interactions between two or more SNPs.

4.2.1 *Setup*

As explained in the introduction to this chapter, our primary goal in building simulated traits with interactions is to understand which types of interactions we are more likely to be able to detect. To this end, we start from one of the simplest frameworks in our null simulations (1,000 causal loci, a regular architecture, no transformation and a heritability of 0.6) and extend it by adding a variety of interactions to its architecture.

In total, we will have *eight traits:* we will allow the α parameter to vary (it can be 0 or -0.5 as before) and, as explained in more detail below, we will have four possible magnitudes for the interaction effect sizes. As before, we transform residuals either through quantile normalisation or by Winsorising them prior to subsequent analysis.

While we have a smaller number of simulated traits than in the null simulations, there is a significant degree of variety in the interactions included, as we now explain.

To create our simulated traits, we begin by randomly placing 22 'left-hand side' (LHS) SNPs at random along the genome; these are required to have an MAF greater than or equal to 5%. We then introduce variation in the types of interactions we create along four axes:

1. Whether or not the LHS has a direct effect on the trait;
2. How many SNPs each LHS interacts with;
3. The sign of the interaction; and
4. Its effect size,

as we now describe.

Presence of a direct effect on the trait: An important question when we analyse the interactions we will detect in real data is whether or not the interacting SNPs have a main effect: do they affect the trait in an additive manner in addition to their non-linear interaction with another SNP (or SNPs), or does their effect operate purely by modulating the effect of other variants? From the perspective of having enough power to detect interactions, we may also wonder whether variants with a direct effect are easier to detect, or on the contrary whether, by having their estimated main effect partially account for their interaction, they become more evasive.

To shed some light on these questions, we build interactions with different properties. Of the 22 LHS interacting SNPs, we choose eight from among those 1,000 SNPs that have a direct effect on the trait; eight from among those other SNPs in the genome which do not have such an effect; and, for a middle ground scenario, we choose six SNPs which, while not having a direct effect on the trait (i.e., their main coefficient is zero) will have a 'biased interaction' by interacting primarily – at a

four-to-one proportion as outlined below – with SNPs with a positive effect.[4] SNPs with a 'biased interaction' of this kind may be estimated to have a positive main effect even though this is zero in reality, since the OLS estimator of a main effect will partly capture interactions if these exist, as we have previously noted.

Number of SNPs that each LHS SNP interacts with: We have remarked before than an important motivation for our approach is the possibility that some SNPs may interact with several or even many others, and how our method will be especially powerful to capture such instances.

To test this, we vary the number of SNPs that each LHS SNP interacts with. More specifically, each LHS SNP may interact with either one other SNP with an additive effect, with 2% of the SNPs with an additive effect (i.e., those in the set of 1,000 causal SNPs, which we denote by PGS in this context), with 10% of these SNPs or with the full PGS (i.e., 100% of its SNPs).

The presence of a direct effect and the number of SNPs that each LHS interacts with are jointly distributed as follows:

- The eight LHS SNPs with a direct effect are divided into four groups of two, each affecting one SNP or 2%, 10% or 100% of SNPs in the PGS;
- The same is true for the eight LHS SNPs that do not have a direct effect;
- The six LHS SNPs that do not have a direct effect but will have a 'biased interaction' effect are divided into three groups of two SNPs, each interacting with: one other SNP which must have a positive effect; in the case where 2% of 1,000 PGS SNPs (i.e., 20 SNPs) are affected, 16 are chosen from among those with a positive main effect and the remaining four from among those with a negative effect; in the case where 10% of PGS SNPs (i.e., 100 SNPs) are

[4] This is without loss of generality: an interaction primarily with SNPs with a negative effect would have a symmetric effect.

affected, 80 are chosen to have a positive effect and the remaining 20 a negative effect.[5]

Note that the 22 LHS SNPs and their RHS counterparts are always the same across all eight simulated traits.

Sign of the interaction: Each LHS SNP will affect all the SNPs with which in interacts (if there is more than one) in the same way: it will be randomly chosen to have either a positive or a negative effect, where in the former case it increases the main effect of the SNPs with which it interacts (they become more positive if their main effect is positive and more negative if it is negative) while in the latter case it reduces the magnitude of the RHS SNPs.

Effect size: We will have *four traits* for each level of α (0 or −0.5) which arise from four different magnitudes of the interaction effect size: the interaction pairs selected as determined above will have an effect that is 1%, 10%, 50% or 100% of the effect size *of the RHS SNP,* respectively for each trait. We may therefore conceive of the interactions we simulate as rotating the increasing line connecting the additive PGS to the trait: if their interaction sign is negative, they decrease its slope (with this change being more pronounced the greater the proportion of PGS SNPs affected), while if it is positive they increase it.

Finally, all traits are constructed to have an *additive* heritability of 60%.

[5] Note that it is not possible to have a biased interaction effect if a LHS SNP affects 100% of the PGS SNPs as these are equally likely to have positive or negative main effects and therefore we expect there to be approximately 500 of each.

4.2.2 *Results*

In Table 4.5, we show the log-transformed p-values – with a sign given by that of the interaction term's estimated coefficient – from testing each of the 22 LHS SNPs for a SNP × PGS interaction for the four traits with $\alpha = 0$, QN residuals and the full range of effect sizes. (The results for the three remaining combinations of α and residual transformations are included in the appendix in Tables C.7, C.8 and C.9.)

From inspecting these four tables, we make the following observations:

1. We are not able to detect any LHS SNPs that interact with only one other SNP (pairwise interactions) regardless of the strength of this interaction;
2. As can be seen from the first two columns of signed p-values, we are not able to detect any LHS interacting SNPs *regardless of the number/proportion of SNPs in the PGS that they affect* when their effect size is only 1% of that of the RHS SNP with which they interact;
3. For a given effect size percentage, we are more likely to be able to detect LHS SNPs that interact with a larger fraction of the PGS;
4. For a given percentage of SNPs affected, our method performs better at detecting LHS SNPs whose interactions have a larger effect size;
5. We are able to detect LHS SNPs equally well regardless of whether or not they have a direct effect.

These results are generally as we expected: interactions that are stronger in effect size, or in number or proportion of PGS SNPs affected, are naturally easier to detect.

We note that neither the two different values of the α parameter, on the one hand, nor quantile normalisation instead of Winsorisation, seem to result in a clearly discernible effect on our ability to detecting interactions. The latter observation is in

contrast with our hypothesis that Winsorisation would yield greater power to detect interactions.

Our inability to capture pairwise interactions is not surprising since we are testing for an interaction with an entire PGS and a single SNP will have only a very modest effect on the whole score. This is in concordance with the motivation for our method, which was to find SNPs that interact with multiple others.

4.3 SUMMARY

In this chapter, we have described simulation scenarios we generated in order to test the propensity of our method to yield false-positive results and assess its power to detect true interactions.

The null simulations showed that in most simulation settings – when there is no strong non-linear transformation applied before adding Normal noise – our method leads to null or very moderate FPRs. Even considering also the more challenging settings where such transformations are applied, the aggregate FPR is still lower than 5×10^{-8} – as expected for a well-calibrated method – for the QN residuals tested for interaction with the LOCO PGS.

Therefore, in our analysis of real data, we will have more confidence in results from this test. However, we will still consider results from Winsorised residuals and the full PGS, as we expect most traits to not have the rather extreme characteristics – very low number of causal loci which are clustered and a strong non-linear transformation of the additive genetic effects – present in the few traits that yielded large numbers of false positives.

Our simulations of traits with interactions yielded the results we expected, showing greater power for stronger interactions and highlighting that we are more likely to detect interacting SNPs the greater the number of SNPS with which they interact.

Table 4.5: Log-transformed p-values with sign of corresponding coefficient obtained from SNP × PGS interaction testing for simulated traits with $\alpha = 0$ and QN residuals.

Each row shows one of 22 left-hand side SNPs with their MAF, coefficient, sign of interaction terms, type of effect and number or proportion of SNPs affected. Columns correspond to the effect size of the interaction terms as a fraction of the effect size of the right-hand side SNP. Genome-wide significant p-values are highlighted in bold.

MAF	Coeff.	Sign	Type	# SNPs affected	*Trait 1* Effect: 1%		*Trait 2* Effect: 10%		*Trait 3* Effect: 50%		*Trait 4* Effect: 100%	
					LOCO	Full	LOCO	Full	LOCO	Full	LOCO	Full
0.38	0.7	-	Direct effect	1	-0.1	0.2	-0.7	-0.9	-1.3	-1.7	-2.3	-2.6
0.42	-0.1	+	Direct effect	1	-0.3	-0.3	0.1	0.1	0.5	0.6	1	0.9
0.28	0.84	-	Direct effect	2%	0.1	0	0.5	0.5	-0.8	-0.8	**-7.9**	**-8.1**
0.46	1.09	-	Direct effect	2%	-0.2	-0.2	-1	-1.4	-3.9	-3.6	**-8.5**	**-8.2**
0.13	-1.04	+	Direct effect	10%	3.6	3.7	3.9	4.2	**59.2**	**61.2**	**214.3**	**216.9**
0.34	-1.86	+	Direct effect	10%	0.4	0.5	1.1	0.9	**32.1**	**34.3**	**128.6**	**127.5**
0.12	-0.18	-	Direct effect	100%	-1.2	-1	**-153**	**-142.8**	**-3426.4**	**-3242.4**	**-2210.1**	**-2089.4**
0.42	-0.01	+	Direct effect	100%	3.7	4.4	**350**	**341.4**	**7660.7**	**7425.4**	**5213.2**	**5055.9**
0.11	0	-	No direct effect	1	-0.1	-0.1	0.2	0.1	-1.2	-1.3	-0.2	-0.2
0.45	0	+	No direct effect	1	1.1	1	-0.3	-0.4	1	1	1.6	1.5
0.08	0	-	No direct effect	2%	-0.6	-0.6	-1	-1.3	-4	-3.6	-6.3	-6.1
0.29	0	-	No direct effect	2%	-1.7	-1.5	-0.7	-0.7	-3.4	-3.7	**-22**	**-22.6**
0.25	0	+	No direct effect	10%	0.1	0.1	1.7	1.8	**35.5**	**36.5**	**147**	**150.3**
0.39	0	-	No direct effect	10%	0.4	0.5	-3.1	-2.7	**-35.8**	**-34.9**	**-151.6**	**-147.2**
0.08	0	+	No direct effect	100%	4.3	3.5	**113.4**	**110**	**2061.2**	**2000.6**	**1428.7**	**1380.4**
0.29	0	+	No direct effect	100%	4.3	3.8	**320.6**	**296.2**	**6039.7**	**5660.4**	**4144.7**	**3869.2**
0.05	0	-	Biased inter. sign	1	0.5	0.6	-1	-1.3	-2.7	-2.7	-1	-1.1
0.1	0	-	Biased inter. sign	1	0.1	0.2	0.1	0	0.2	0.1	0.1	0.1
0.08	0	+	Biased inter. sign	2%	0.2	0.3	0.3	0.4	2.8	3.4	**8.2**	**8.8**
0.08	0	+	Biased inter. sign	2%	0.3	0.3	0.6	0.5	0.6	0.6	1.9	2
0.14	0	+	Biased inter. sign	10%	-0.1	0	3.5	4	**33.9**	**33.5**	**145.3**	**135.5**
0.3	0	+	Biased inter. sign	10%	0.9	0.7	1	0.9	**32.9**	**22**	**190.8**	**128.9**

5 Results: alkaline phosphatase

We begin our analysis of real data by examining the results of interaction testing for a specific blood biomarker, the levels of serum alkaline phosphatase. We chose to focus on this phenotype for two main reasons: first, it is one of the traits for which we find the greatest number of interactions; and second, at least three pairs of interacting loci have already been identified in the literature, so it serves as a positive control. As we will see, we are able to recover these three known interactions and to significantly extend the set of interacting loci for ALP.

5.1 BACKGROUND

Alkaline phosphatases are a group of isoenzymes widely present in nature: they form a super family of proteins produced by many organisms, from archaea to humans [172, 173]. In vertebrates, ALP is an ectoenzyme that binds to the outer side of the cell membrane and whose basic function is to catalyse the hydrolysis of phosphate groups in an alkaline environment, thus freeing inorganic phosphate [173, 174].

5.1.1 *Location in humans*

In humans, ALP is found, in decreasing order of concentration, in the placenta, ileal mucosa, kidney, bone and liver, with the majority of ALP *in serum* – which is the object of our analysis in this chapter – being released from the liver and bone (with a

small amount released from the intestine also present in some individuals as will be explained in greater detail below) [172].

ALP can be classified according to its origin and location as either tissue specific (from the intestine, placenta or germinal tissues) or tissue non-specific, which comprises most of the amount circulating in serum [172].

5.1.2 *Variation in ALP levels in healthy individuals*

ALP levels in healthy individuals change with age: they are high during childhood and puberty (due to bone growth and development, one of the key known functions of ALP), then decrease from the ages of 15 to 50 (when they are slightly higher in men than women) and gradually increase again in old age [172]. Women in the third trimester of pregnancy tend to have elevated ALP levels due to an influx of placental ALP into blood [175]. ALP levels are known to be affected by many drugs, food intake, smoking, alcohol intake and medical conditions [172].

5.1.3 *Physiological functions*

In spite of their ubiquitousness in nature, suggestive of an important role in fundamental biochemical processes, the precise function of alkaline phosphatases is still not well understood [173, 174, 176]. It is believed that ALP plays a mediating role in processes such as molecular transportation (including of lipids), metabolism and gene expression [176, 177].

One of the functions of ALP that was first suggested (by Robert Robinson in 1923 [176]) – and which partly justifies its widespread use as a clinical biomarker – is that of skeletal mineralisation, specifically for tissue-non-specific ALP. This process is essential in vertebrates as it contributes to the maintenance of body mass and the

preservation of soft tissues and enables movement and mastication [174]. ALP is also thought to play a role in vitamin B6 metabolism and neurogenesis [173].

5.1.4 *Disease associations and clinical relevance*

The level of ALP in serum is one of the most commonly measured biomarkers in clinical medicine [176]. While it is especially used to assist in the diagnosis of liver and bone diseases [176], ALP levels have also been linked to disorders of the prostate and endocrine system [177].

High levels of ALP have been linked to cholestatic liver disease (patients with this condition have some of the highest recorded levels of this biomarker), biliary obstruction (due to cancer, choledocholithiasis, biliary stricture and sclerosing cholangitis), intrahepatic cholestasis, infiltrative liver disease, severe alcoholic hepatitis and AIDS [172], among others. Conversely, abnormally low levels of ALP have been linked to Wilson's disease, chronic myelogenous leukemia and aplastic anemia [177].

5.1.5 *Relationship with ABO blood group and secretor status*

Levels of ALP have long been known to depend on ABO blood group and ABH secretor status, which determines whether soluble ABO blood antigens are present in bodily fluids. The accumulated evidence from multiple studies in different populations [178–182], beginning with a first article by Bamford et al. in 1965 [178], has yielded the following conclusions:

1. Total serum ALP levels are higher for individuals with B/O blood type than those with A/AB;

2. Total serum ALP levels are higher for ABH secretors than non-secretors. This seems to be driven by IALP, which is more frequently detectable in the serum of secretors;
3. Within secretors, B/O blood types have higher ALP levels than those with blood types A/AB, a pattern again caused at least partly by differences in IALP;
4. Finally, *the difference in IALP levels across blood types – decreasing with A antigens – is more pronounced in secretors than non-secretors.*

As explained by Nakano et al. in a recent article [182], two alternative but non-mutually-exclusive hypothesis have been proposed to explain the lower levels of serum IALP seen in secretors with blood type A compared to those with blood type B/O. In 1968, Langman et al. [179] suggested that this could be due to a genetic down-regulation of IALP in A types leading to a lower rate of production of ALP by the mucosal cell which is then absorbed into the blood. Alternatively, in 1980, Bayer et al. [180] found that red cells of blood group A bind IALP to a much higher degree than erythrocytes of groups B/O and proposed instead that it is the regulation of the clearance of IALP in plasma through adsortion by blood group antigens that underlies the difference in levels of IALP. To our knowledge, this debate has not been resolved but the literature appears to lean towards the latter hypothesis [183–185].

This physiological interdependence between serum ALP levels (in particular, IALP), ABO blood group and secretor status, combined with *the differential effect on IALP levels of blood antigens depending on secretor status,* and the fact that both blood group and secretor status are genetically determined, suggests that a statistical interaction between variants contributing to the determination of blood group or secretor status may impact ALP. This is indeed the case and such an interaction has been found and successfully replicated, as we will see below.

5.1.6 *Genetics*

The production of the different types of ALP has been linked to four different genes [174]:

- Tissue-non-specific ALP (present in the liver, bone and kidney) is encoded by *ALPL* in chromosome 1;

Among the tissue-specific types:

- Placental ALP is encoded by *PLAP* in chromosome 2;
- Germ cell ALP is encoded by *GCAP* in chromosome 2; and
- IALP is encoded by *ALPI*, also in chromosome 2.

Early GWASs naturally identified *ALPL* loci as being associated with ALP levels in blood plasma or serum. For example, in [186] four associations were found at *ALPL*, *GPLD1*, *JMJD1C-REEP3* and *ABO* in a discovery cohort of 7700 individuals of European ancestry; *ABO* was the most strongly associated locus. In a Japanese cohort of 14,700 individuals, *ALPL*, *ABO* and *GPLD1* were again found to be associated with ALP and a newly-associated locus in region 15q26 was identified [187]. And in another 'smaller scale' GWAS of 56,400 primarily-European samples, Chambers et al. [188] identified 14 hits at the 1×10^{-8} significance level and ten new genes involved in determining ALP levels, including *FUT2* and *ASGR1*.

Of direct relevance to the previous section on the relationship between ALP and blood types, Li et al. [189] replicated the association of *ABO* and found in their sample of 8800 Chinese Han individuals that blood type correlated with serum ALP, with A, AB, B and O types showing ALP levels in increasing order. They also found a gene-by-environment interaction between rs651007, a variant upstream of *ABO*, and obesity.

More recently, modern biobanks have enabled the discovery of many more genetic associations with ALP. For example, Sinnott-Armstrong et al. [190] used a set of 363,000 samples of primarily White British ancestry in the UKB (which largely overlaps our own) to identify four protein-truncating variants at three different genes (*ADH1C*, *GPLD1* and *RAP1GAP*) and an additional 86 protein-altering variants impacting 65 different genes, where 74 of these associations were novel. In another article from the same year, Backman et al. [191] used the exome sequencing data in the UKB to identify 123 rare coding variants (i.e., with minor allele frequency not greater than 1%) in 19 genes associated with ALP in 431,000 European samples.

5.1.7 *Interactions*

The genetic architecture of ALP has been shown to include a few interactions. One of these was first reported by Masuda et al. [184], who analysed 3000 Japanese subjects and identified a positive interaction between rs550057 (C to T), located in an intron of *ABO*, and rs2071699 (C to T), a non-synonymous exonic variant of *FUT1*: in a linear regression model including rs550057, rs2071699 and their interaction, both variants had negative main effects but their interaction was positive; these terms explained 12.4, 0.9 and 0.3% of the total variability in serum ALP levels, respectively. The authors also examined the mode of inheritance of these mutations and estimated rs2071699 to be additive while the T allele was thought to be dominant for rs550057.

Masuda et al. also imputed rs1047781 (A to T), a missense mutation in *FUT2* not sequenced in their data, and found evidence for an interaction with rs550057 affecting ALP. However, rs1047781 is in high LD with rs2071699 (the variant interacting with *ABO* discussed in the previous paragraph) in the Japanese population (squared correlation of 0.59) and therefore these two interactions may not represent truly distinct signals.

To understand the functional significance of this interaction, we begin by noting that rs550057 is an important determinant of ABO blood group (in both Japanese and European individuals), with the T allele being (positively) associated with the number of A-type blood alleles [184, 192]: in fact, Masuda et al. show that the number of A alleles matched the number of alternative T alleles at rs550057 for 98% of their samples. Therefore, individuals with one or two mutated alleles at rs550057 will likely have blood types AO/AB or A, respectively.

Moreover, the interacting SNP rs2071699 is thought to reduce the amount of ABH antigens in erythrocytes: the *FUT1* gene that it affects is responsible for H antigen formation, which undergoes further modification to result in A or B antigens, and the SNP rs2071699 has been shown to reduce this activity [184]. Given the high correlation between *FUT1* SNP rs2071699 and *FUT2* SNP rs1047781, and the fact that the latter SNP also shows a significant interaction signal, this interaction may also be thought to more directly involve the ABH secretor status – the secretor phenotype is primarily determined by the missense mutation rs1047781 in the Japanese population [184] (and in Chinese [193] and East and Southeast Asian populations more generally [194]). We will interpret the second interaction found between *ABO* polymorphism rs550057 and *FUT2* SNP rs1047781 as it shows the same interaction patterns (negative main effects of T mutations on either SNP and positive interaction coefficient) and is more directly linked to the ABH secretor status, whose relationship with ABO blood type in determining ALP levels has been extensively studied, as discussed previously.

The positive interaction between rs550057 and rs1047781, with both SNPs having negative main effects, matches the variation in ALP levels by ABO blood type and ABH secretor status reported above. More specifically:

- Given the nearly-perfect correlation between the genotype at rs550057 and the number of A alleles, the fact that individuals with genotypes CT/TT at rs550057 had lower ALP levels than those with genotype CC implies that individuals with blood types A/AB/AO had lower ALP levels than types B/O, as expected from the literature;
- Similarly, given that ABH secretor status is determined by *FUT2* SNP rs1047781 (with homozygosity for the missense mutation T leading to the absence of a functional FUT2 enzyme and to the consequent lack of ABH antigens in bodily fluids, i.e., non-secretor status [195, 196]), the negative main effect for rs1047781 reflects the fact that non-secretors have lower ALP levels than secretors, again as expected;
- Finally, the positive interaction term means that the decrease in ALP levels with A blood antigens is more pronounced for secretors than non-secretors, as had been previously reported.

This interaction between *ABO* rs550057 and *FUT2* rs1047781 in a Japanese cohort has recently been successfully replicated in the UKB by Lyon et al. [185], who report a positive interaction between *ABO* rs635634 and *FUT2* rs281379. While the specific variants identified at each gene are different, their effect on blood group and secretor status are very similar.

Considering first the ABO blood group, rs635634 (C to T) was found to discriminate between A and non-A blood types with an accuracy of 92% in a sample of 1000 Han Chinese individuals [197] and is in high LD (squared correlation of 65%) with rs550057 in our dataset of White British individuals in the UKB. Regarding the ABH secretor status, while the missense mutation rs1047781 is the most prevalent non-functional *FUT2* allele in East and Southeast Asian populations [194], the equivalent

SNP in Europeans is rs601338 (G to A) [194, 195, 198, 199], which is in high LD (squared correlation of 84.1%) with rs281379.

The interaction between *ABO* rs550057 and *FUT1* rs2071699 does not replicate in Lyon et al.'s analysis of the UKB. However, and as noted by Masuda et al., this may be due to the much lower frequency of the *FUT1* SNP in European compared to Japanese populations (0.022 vs. 0.36 in the HapMap database [184]).

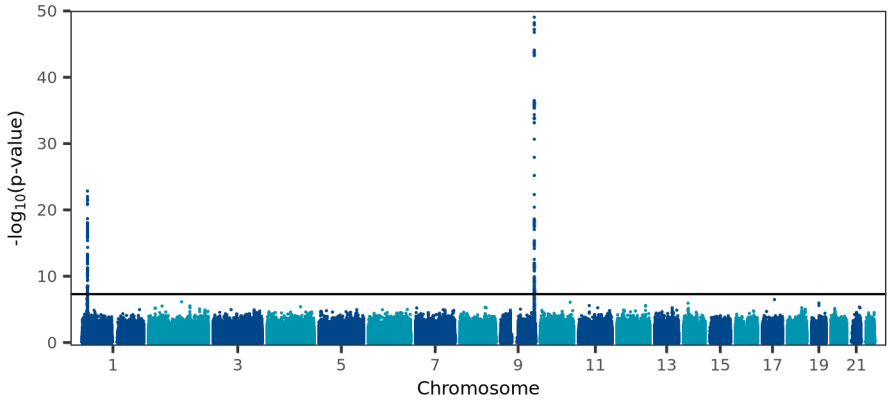
Finally, Lyon et al. report two new interactions affecting ALP levels, between *TREH* rs12225548 and *FUT2* rs281379 and between *ABO* rs635634 and *TREH* rs12225548, both with positive effects.

5.2 SNP × PGS INTERACTION GWAS

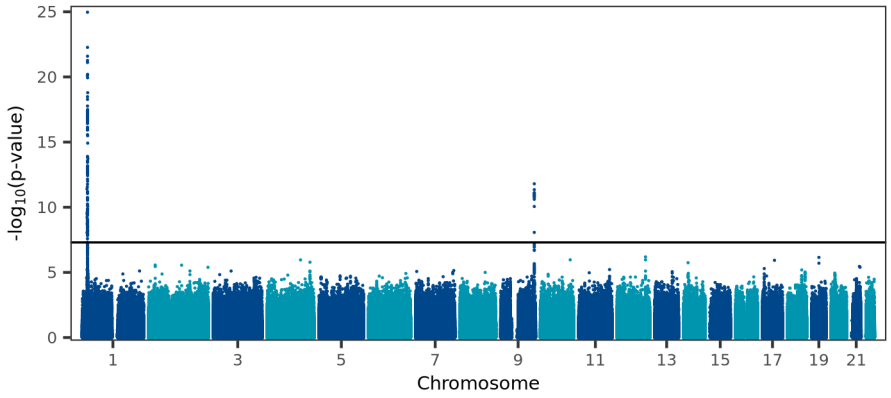
Figures 5.1 and 5.2 show Manhattan and QQ plots of the summary statistics from the SNP × PGS interaction GWASs with LOCO and full PGSs for ALP with QN and Winsorised residuals, respectively.

In the QN results in Figure 5.1, we see two clear GWS peaks on chromosomes 1 and 9 in both the LOCO and the full PGS results. In the results for the Winsorised version of the trait, we can see an additional peak in the long arm of chromosome 1 as well as in chromosome 19 and, in the full PGS results only, a peak in the short arm of chromosome 1 and in chromosomes 3, 4 and 22.

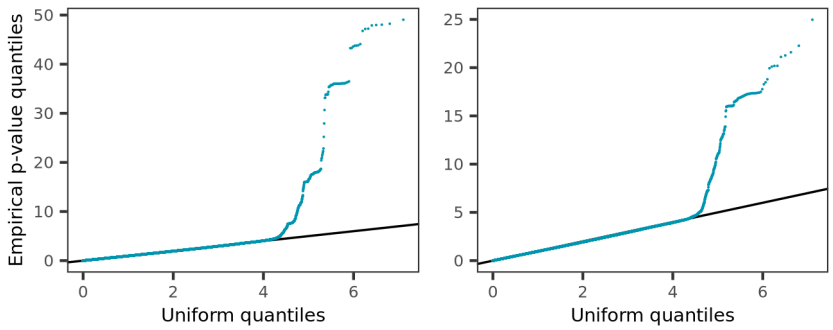
We note that these GWASs show no obvious signs of inflation. To investigate a possible correlation between results in the original GWASs (Manhattan and QQ plots for these are shown in Figure D.1 in the Appendix) and these GWASs of interactions with the PGS, we plot log-transformed p-values (with sign given by that of the coefficients) in the original against interaction GWAS for all SNPs that are GWS in at



(a) LOCO PGS

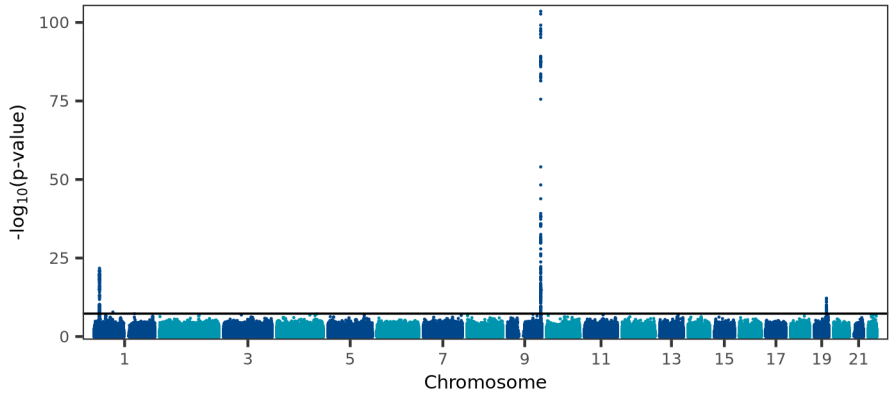


(b) Full PGS

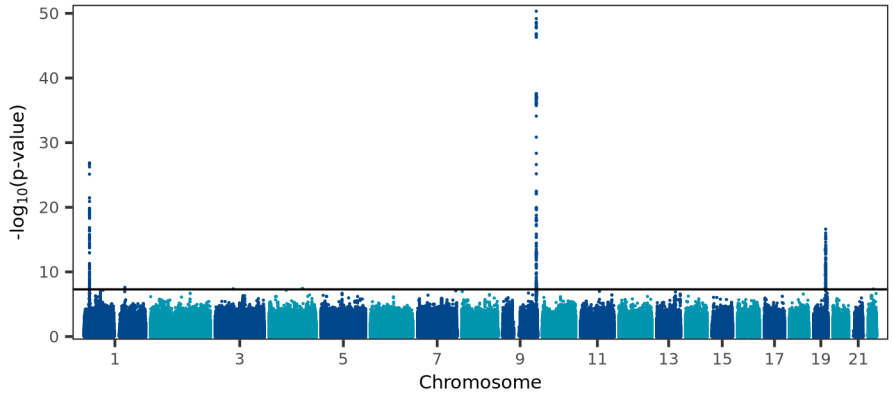


(c) LOCO PGS (d) Full PGS

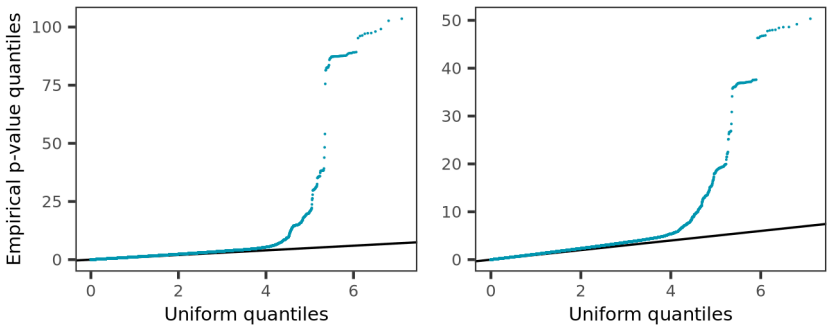
Figure 5.1: Manhattan and QQ plots of results from SNP × PGS interaction GWASs with LOCO or full PGS: ALP with QN residuals.



(a) LOCO PGS



(b) Full PGS



(c) LOCO PGS (d) Full PGS

Figure 5.2: Manhattan and QQ plots of results from SNP × PGS interaction GWASs with LOCO or full PGS: ALP with Winsorised residuals.

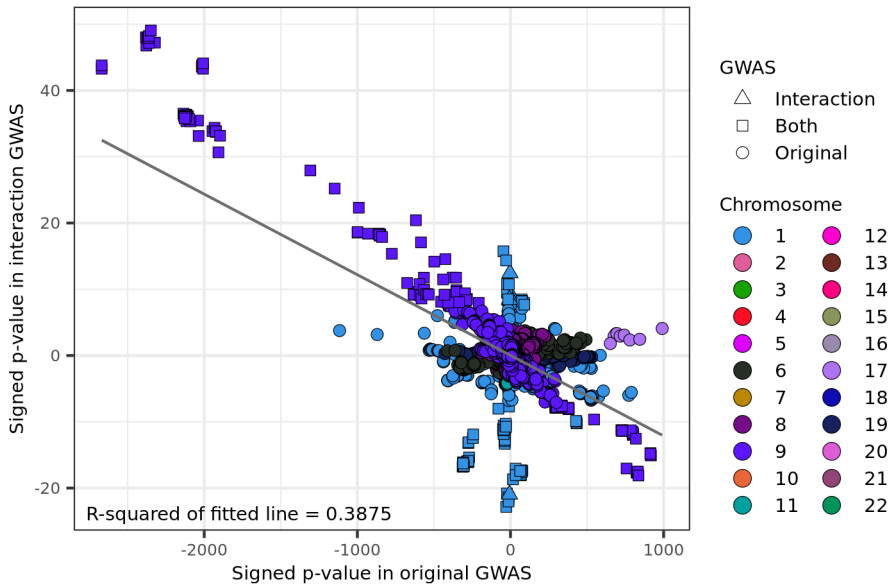
least of one the two GWASs. These scatter plots are shown in Figure 5.3 for the QN trait and in Figure 5.4 for the Winsorised version.[1]

From inspecting the R-squared coefficient of the linear regression fitted for each of the four plots, we note that the correlation between original and interaction GWAS varies significantly by type of PGS and residual, ranging from 2.3% for the full PGS and QN residuals to 70.2% for the LOCO PGS with Winsorised residuals. By examining also the Manhattan plots in Figures 5.1 and 5.2, it appears that the higher correlations observed in the Winsorised results compared to their QN counterparts are due in part to the strong interaction signal in chromosome 19, which seen only in the former and is also present in the original GWAS (Figure D.1).

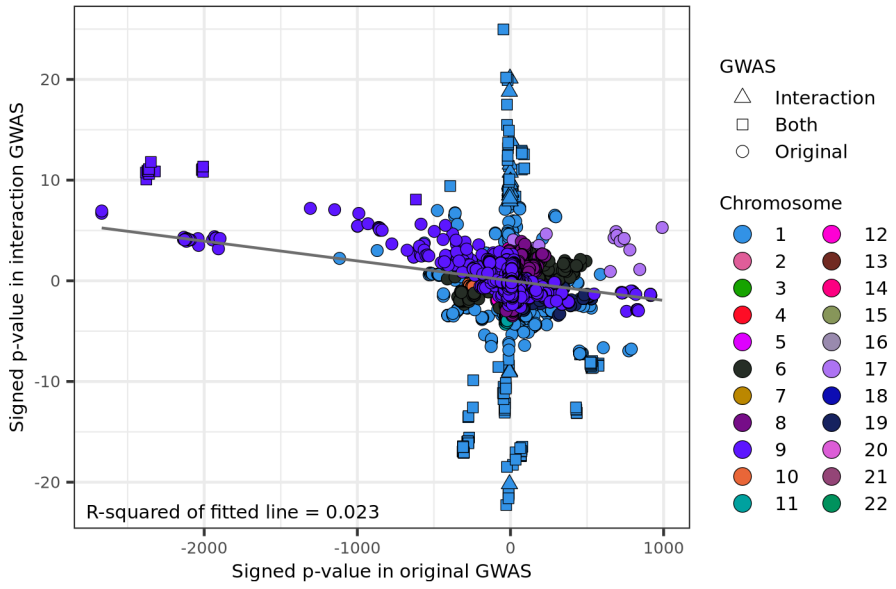
By examining the patterns of correlation for the different SNPs and chromosomes, we see clearly distinct trends for different groups of SNPs, even within the same chromosome. In particular, some SNPs show no signal in the original GWAS but are significant in the interaction GWAS, with either positive or negative sign. In contrast, other SNPs have a strong direct association with the trait but no evidence of interaction. Interestingly, there are several SNPs that have a significant positive effect in the original GWAS but a negative interaction effect, or vice-versa.

This brief analysis allows us to conclude that there is no consistent pattern of correlation, either in strength or sign of effect, between the original and interaction GWAS results, as one might fear if the latter were simply a spurious truncation of the former. However, there are strong general tendencies reflecting, as we expected, that interactions are more likely to involve SNPs that also have a direct effect on the trait detected in the original GWAS.

[1] Note that, for making these plots, we included all GWS SNPs at a locus, i.e., we did not apply LD Clumping or any other technique to identify independent hits.

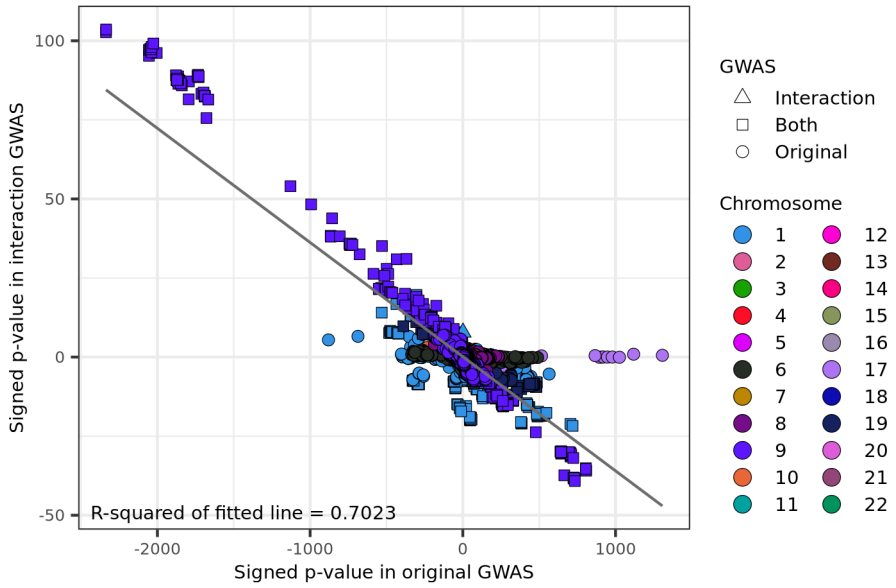


(a) LOCO PGS

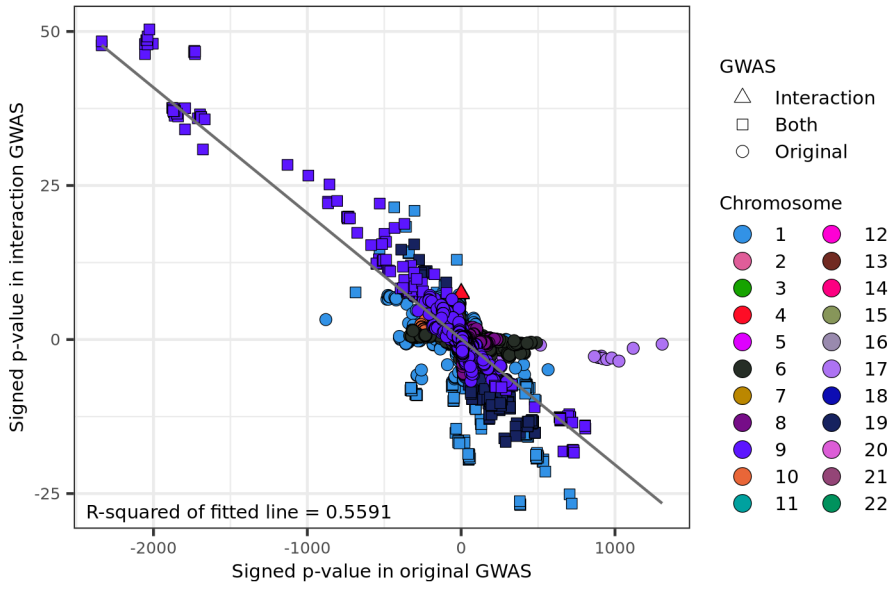


(b) Full PGS

Figure 5.3: Scatter plot of signed log-transformed p-values from original and SNP × PGS interaction GWASs with LOCO or full PGS: ALP with QN residuals.



(a) LOCO PGS



(b) Full PGS

Figure 5.4: Scatter plot of signed log-transformed p-values from original and SNP × PGS interaction GWASs with LOCO or full PGS: ALP with Winsorised residuals.

5.2.1 *Independent LHS hits*

The independent LHS hits obtained from the SNP × PGS interaction GWASs are listed in Tables 5.1 and 5.2 for the QN and Winsorised versions of the phenotype, respectively.

Beginning with the QN phenotype, we see that our stepwise regression procedure yielded eight independent hits: six of these, from two different regions, arose from the interaction with the LOCO PGS and the remaining two from the full. The LOCO results show five different SNPs located in chromosome 1: two of these are intronic variants of *ALPL* while the remaining three are linked to nearby *EIF4G3*, *LDLRAD2* and *LINC00339*. We note that four of these had significant direct signal in the original GWAS, while *EIF4G3* variant rs77569513 did not.

The LHS hit in chromosome 9 is rs635634, a variant upstream of *ABO* that, as we saw previously, accurately discriminates between ABO blood types A and non-A. To visualise this interaction, we plot the linear relationship between the PGS and the phenotype (after regressing out covariates) by the allele count at this SNP in Figure 5.5. We see that the slopes of the lines increase with the allele count, corresponding to an interaction with positive sign – of the conceptual examples given in Figure 1.1 in Chapter 1, this real data plot resembles the third panel the most.

Finally, the two SNPs in chromosome 1 that interact with the full PGS fall in intergenic regions: the first SNP is located between *NBPF3* and *ALPL*, and the second between *MIR4418* and *ZBTB40*, according to the ANNOVAR database [200].

Turning now to the Winsorised phenotype, we have eight LHS hits from the interaction GWAS with the LOCO PGS and an additional three hits from the GWAS with the full PGS. We first note that we identify three distinct *ALPL* SNPs as interacting with the PGS. The first of these, rs75746521, is in high LD (squared correlation of

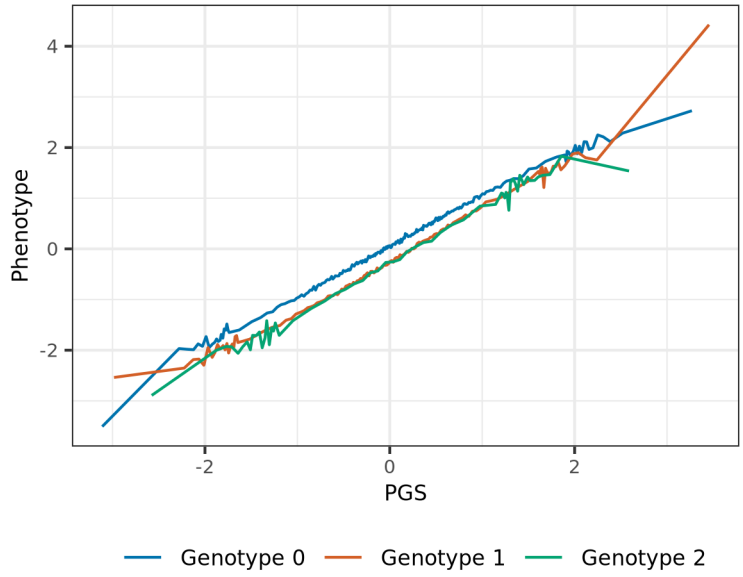


Figure 5.5: Plot of relationship between the LOCO PGS and the phenotype (after regressing out covariates) for ALP with QN residuals by genotype at rs635634.

Lines shown are obtained by linear interpolation through median PGS values and average phenotype value across a set of bins of size 1000 for genotypes 0 and 1 (these contain 217,700 and 100,000 individuals, respectively) and of size 500 for genotype 2 (11,100 individuals), with 20 smaller bins of sizes 50/50/10 for genotypes 0/1/2, respectively, at either end of the PGS distribution.

63%) with rs10465542, a neighbouring SNP identified in the QN results, and may therefore constitute the same signal. The second *ALPL* SNP, rs149356924, had also been identified in the QN results, but the third SNP, rs1697405 (a 3'-UTR variant), is new and very common (MAF of 39.6%). Regarding *ABO*, we identify two separate SNPs (rs550057 and rs635634) as independent signals even though they are in high LD (squared correlation of 65%), as mentioned in Section 5.1.7 – both SNPs have strong signal for interaction in the QN results but rs550057 does not survive stepwise filtering once rs635634 (which is more significant) is included. Between LOCO and full results, five new genes are identified as interacting with the PGS: *NBPF3*, *ST6GALNAC3*, *MAMSTR*, *LMX1A* and *RP3-434P1.6*. We note that, of the LOCO hits, only one (*ST6GALNAC3* variant rs143104384) does not show a significant direct

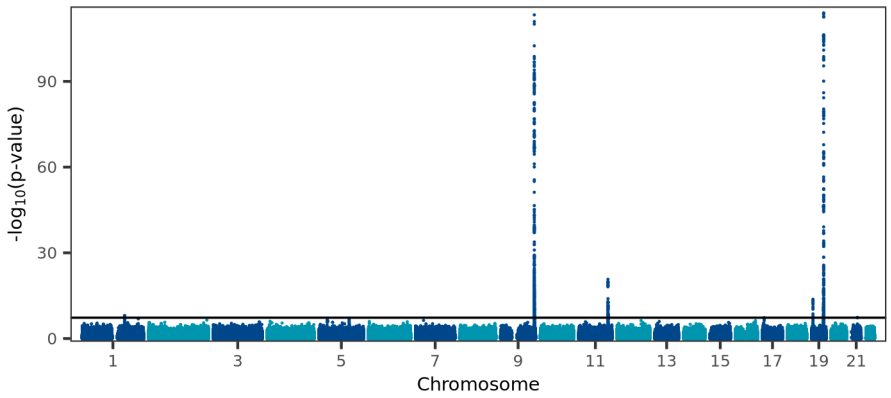
association in the original GWAS. By contrast, none of the full PGS hits show such a direct association – given that they also have very low MAFs, there is a strong possibility that they might be spurious associations.

5.3 PAIRWISE INTERACTION GWAS

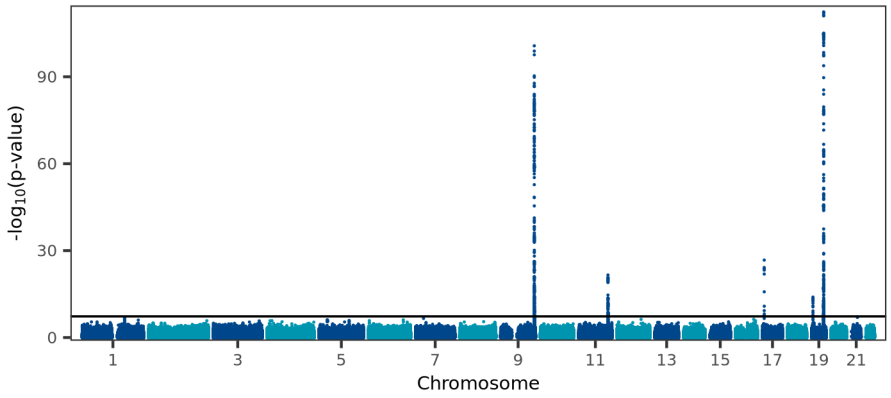
Having identified independent LHS hits, and as detailed in Chapter 3, we proceed by running a simple pairwise interaction GWAS for each LHS hit.

The results from such a GWAS for *ABO* variant rs635634 are shown in Figure 5.6 for both QN and Winsorised residuals. In both cases, we see strong interaction signals in chromosomes 9, 11 and 19; however, the peak in chromosome 9 is likely to be driven by LD as this is where *ABO* is located. We also see, in the QN phenotype, smaller peaks in chromosomes 1 and 21 and, in the Winsorised version, a prominent peak in chromosome 17.

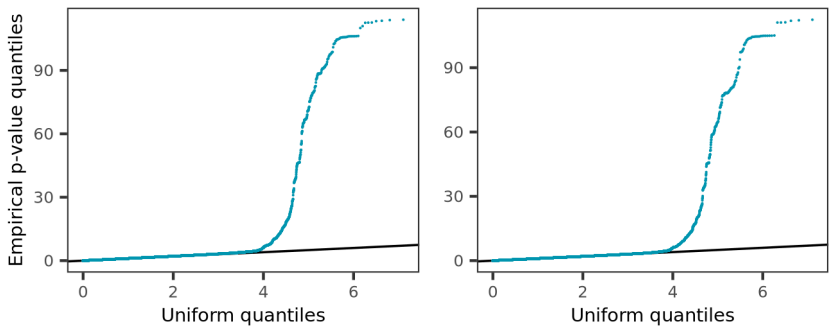
Running a stepwise regression algorithm on the results from these pairwise GWASs yields a set of independent pairwise interactions which we list in Tables 5.3 and 5.4 for the QN and Winsorised residuals, respectively. The interactions listed survived two different filtering steps. First, we only list interactions between SNPs on different chromosomes to ensure that these are not driven by LD, as we suspect is the case for the interaction peak in chromosome 9 above, for example. Second, we exclude rare interactions: as noted in Chapters 2 and 3, rare variants can lead to false-positive associations; while we removed rare *individual* SNPs (MAF < 0.001) from our data at the start of this analysis as explained in Chapter 2, the *interaction* between two SNPs with MAF ≥ 0.001 can still be very rare as its frequency will depend on the *product* of the frequencies of individual SNPs. Therefore, and for consistency with



(a) QN residuals



(b) Winsorised residuals



(c) QN residuals (d) Winsorised residuals

Figure 5.6: Manhattan and QQ plots of results from pairwise interaction GWASs with LHS hit rs635634: ALP with QN or Winsorised residuals.

our filtering of individual variants, we remove from our results interactions whose frequency[2] is less than 0.1%.

A first conclusion that we may draw from examining these tables is that we have uncovered evidence for *networks* of interacting loci where a single LHS SNP interacts with multiple RHS SNPs. (We note, however, that our approach is more powerful for detecting interactions involving multiple other variants as discussed in Chapters 3 and 4; therefore, the fact that most of the interactions we identify here are of this kind does not imply that this is the most common kind of interaction affecting complex traits – we only wish to emphasise that such networks do exist.)

We note also that the same LHS SNP can have interactions with different *signs* for different RHS SNPs: for example, in the QN results, *ALPL* intron variant rs149356924 has two interactions with positive sign (with SNPs in chromosomes 3 and 12, whose effects in the original GWAS are negative and positive, respectively) and five other interactions with negative sign (the effects of these five SNPs in the original GWAS are also negative except for rs7684999, the second chromosome 4 SNP).

Moreover, while, as we saw previously, we have a greater number of LHS hits for the Winsorised residuals than for their QN counterparts (as expected), we nevertheless have fewer independent pairwise hits for the former than the latter – this again suggests that the QN LHS hits are more robust than those obtained with the Winsorised version of the trait.

[2] Defined as the ratio of the number of samples with non-zero allele count at both loci to the total number of samples. We note that this may overestimate the frequency of rare interactions since our genetic data has imputed dosages (which can take any value between 0 and 2) rather than integer allele counts for most SNPs – we will address this in future work.

5.3.1 *Results for QN residuals*

We now highlight four findings from our list of independent pairwise interaction hits, focusing first on the QN version of the trait. We begin by discussing three instances in which a LHS SNP interacts with several RHS SNPs, forming networks of interacting loci. However, we caution that all three LHS SNPs are rare (MAF below 1%) and therefore that further evidence is needed for us to be fully confident that these interactions are non-spurious. In contrast – and as implied by our filtering procedure – most of the RHS SNPs that interact with these rare LHS SNPs are reassuringly somewhat-common. In general, we will be most confident in interactions between two non-rare variants.

We begin by considering the LHS hit rs77569513, an intronic variant of *EIF4G3*[3] which is found to interact with five different loci linked to as many genes in four different chromosomes. We noted before that this SNP does not show a direct effect on ALP; however, other *EIF4G3* loci have been associated with this trait [190, 202].

Second, we find seven pairwise interactions involving rs149356924, an intronic variant of *ALPL* that shows a strong signal of direct association with the trait. That a variant at the primary gene responsible for producing the protein whose levels in blood are being measured as a phenotype shows evidence of interactions with several others is in a sense expected. However, interactions involving *ALPL* and affecting ALP levels had not been found previously, to the best of our knowledge.

It is interesting that, of the two *ALPL* SNPs with independent evidence for an interaction with the PGS (rs10465542 and rs149356924, as can be seen in Table 5.1), we are only able to find RHS partners for the latter, which has a *weaker* signal for interaction (which would imply lower power for detecting interaction pairs). This

[3] *EIF4G3* is involved in mRNA cap recognition and in the transport of mRNAs to the ribosome [201].

may mean that the interaction identified between rs10465542 and the PGS may be capturing an interaction of this SNP with a substantial proportion of the score that cannot be adequately captured by a few pairwise interactions.

Third, we identify a large network of interactions involving rs573303250, a variant upstream of *LDLRAD2*, and at least nine other genes. This variant has a strong direct association with the trait; moreover, three other SNPs linked to this gene had previously been associated with ALP levels [190]. This interaction may shed light on a biological pathway involving *LDLRAD2* and other genes which merits further investigation.

Having briefly examined these first three groups of interacting loci, which as we noted above require caution in their interpretation as their LHS SNPs are rare, we proceed to consider in greater detail a key network of interacting loci driven by *ABO* variant rs635634, an *ABO* upstream variant that correlates strongly with blood type, as mentioned previously. This variant was found to interact with five other loci/genes: *C1orf105*, *TREH*, *FUT2*, *FUT6* and *MIS18A*. Since the LHS SNP driving these interactions, rs635634, is common (MAF of 18.6%), and only one of the interacting SNPs (*MIS18A* variant rs533852698) has a low MAF of 0.2% – and given also the previous evidence for interactions involving this locus, as we discuss below – we are very confident that this network of interactions is real and biologically meaningful.

The interaction between *ABO* rs635634 and *FUT2* rs681343 is the same interaction between variants determining ABO blood group and secretor status that was originally identified by Masuda et al. [184] in a Japanese sample and recently replicated in the UKB [185]. As described previously, Lyon et al. [185] found a positive interaction between *ABO* rs635634 and *FUT2* rs281379. The *ABO* SNP we identify is

the same[4] while the *FUT2* variant is different. However, rs681343 is in nearly-perfect LD (squared correlation of 99.8%) with rs601338, a stop-gain mutation in the same gene and which, as we noted previously, is the relevant non-functional *FUT2* allele in European populations, determining ABH secretor status [194, 195, 198, 199]. In fact, rs601338 has nearly as significant an interaction with the *ABO* SNP rs635634 as rs681343, but only the latter survives our stepwise regression procedure. Finally, the sign of the interaction we report is negative while previous results show it to be positive; however, this is an artefact caused by a different allele being counted: the non-functional *FUT2* mutation at rs601338 is G to A, but the count of allele A is nearly-perfectly *negatively* correlated with the count of the *minor* allele C of rs681343, which is used in our regression. Therefore, the interaction in fact has a positive sign if the non-functional *FUT2* allele is counted, as expected.

We have seen that we recover, with identical sign, the interaction between *ABO* variant rs635634 and the non-functional *FUT2* mutation in Europeans, rs601338. We also identify a positive interaction between the same *ABO* variant and the *TREH* missense variant rs2276065. rs2276065 is in very high LD (squared correlation of 97.2%) with rs12225548, the *TREH* variant identified by Lyon et al. as interacting positively with *ABO*. Since rs12225548 is an intronic variant while rs2276065 is a missense mutation, it seems likely that the latter is the 'causal' variant. In lieu of a functional interpretation of this interaction, we simply note that trehalase, the protein encoded by *TREH*, hydrolyses trehalose to release glucose [203], and that this *TREH* mutation has been linked to trehalase deficiency, according to the VEP database.

Beyond replicating the interactions of the key *ABO* blood group-tagging variant in Europeans with the *FUT2* secretor status mutation and a *TREH* missense variant,

[4] In the results for the Winsorised residuals, we identify both rs635634 and rs550057 – the *ABO* intronic SNP originally identified in the Japanese sample and which is in high LD (squared correlation of 65%) with rs635634 in our data – as can be seen in Table 5.4 in the Appendix.

we extend this network of interactions to include three other newly-reported loci: *FUT6*, *C1orf105* and *MIS18A*.

FUT6 is a gene involved in the processing of sugar molecules that plays a role in the biosynthesis of blood group systems, according to the pathways database PathCards [204]. We note that rs12019136, the *FUT6* variant in question, has a direct association with the trait (log-p-value of 7.89) is not in LD with *FUT2* variant rs681343 discussed above (squared correlation of approximately zero).

Turning now to *C1orf105* SNP rs2301455, we first note that it does not have a direct association with the trait. Moreover, while both VEP and ANNOVAR link this SNP to *C1orf105*, it is located in an intron of *PIGC*, as reported by Open Targets Genetics [205, 206]. *PIGC* is part of a gene complex of which multiple genes (including *PIGH*, *PIGV* and *PIGW*) have been associated with severe neurodevelopmental and skeletal anomalies, sometimes accompanied by significantly elevated ALP levels (hyperphosphatasia) [207–209]. Like other genes in this family, *PIGC* is involved in glycosylphosphatidylinositol (GPI) lipid anchor biosynthesis, whose function is to anchor proteins to the cell surface [201].

ALP is a GPI-anchored protein, and the mechanisms by which genetic variation at another *PIG–* gene, *PIGW*, relates to circulating ALP levels have recently been explored in connection with a condition known as hyperphosphatasia with mental retardation syndrome (HPMRS) or Mabry syndrome [208, 210]. As reviewed in ref. [208], Chiyonobu et al. [210] first found that the levels of cell-surface GPI-anchored ALP in blood granulocytes were lower in a patient with this syndrome, something they attributed to two loss-of-function mutations in *PIGW*. In cells defective in PIGW, the GPI anchor is never attached to ALP proteins, which become soluble and are secreted, thus entering circulation. A similar mechanism involving *PIGV* had also been previously found for HPMRS in a separate study [207].

Lastly, an interaction between *ABO* and *MIS18A* for ALP has not been previously reported. However, given that the *MIS18A* SNP is rare (MAF of 0.2%) and that it does not have a direct effect on the trait (log-p-value of 0.18), this association may be a false positive.

5.3.2 *Results for Winsorised residuals*

We conclude our analysis of pairwise interactions for ALP by examining the results for the Winsorised version of this trait (shown in Table 5.4) and comparing them with the QN results discussed in the preceding paragraphs.

First, we note that we also detect interactions between rs635634 (as the LHS SNP) and the same variants in *FUT2*, *FUT6* and *TREH* (as RHS SNPs) as for the QN case, and that all interactions have the same sign in both versions of the trait – this adds strength to our evidence for an epistatic network involving these loci. We do not detect *ABO–C1orf105* or *ABO–MIS18A* interactions as before[5] but find a novel interaction between rs635634 and rs186021206, a variant 13.5 kb away from *ASGR1*.[6]

The *ASGR1* gene plays a crucial role in the homeostasis of circulating glycoproteins [201], of which ABO antigens are an example, and rs186021206 closely tags (squared correlation of 0.86 in Icelanders [211]) a 12-base-pair deletion in intron 4 of this gene that leads to a truncated protein prone to degradation [211]. Interestingly in our context, heterozygous carriers of this deletion were found to have lower levels of non-HDL cholesterol and a lower risk of coronary artery disease than non-carriers, which is evidence for a role of this variant in lipid-related processes; carriers were also

[5] We noted above that the latter may be a false positive.

[6] This interaction has a log-p-value of 7.27 (also with a negative sign and with an effect of approximately similar magnitude) in the QN results. This adds credibility to this signal, which is only genome-wide significant in the Winsorised trait. Notice the difference in the evidence for interaction at this locus in the QN vs. Winsorised Manhattan plots in Figure 5.6.

found to have higher ALP and vitamin B12 levels, but lower levels of triglycerides. We return to this SNP and its tagged deletion below.

Second, rs550057, a second independent *ABO* LHS hit, is associated with the same variants of *TREH*, *ASGR1*, *FUT2* and *FUT6*, and with the same sign, as before – this would be expected if these two *ABO* loci represent the same functional signal. However, for rs550057, we identify an additional interaction with an intronic variant of *PDE1C* (but given its very low frequency (0.2%), this finding may be spurious).

The evidence for the two signals at *ABO* (rs635634 and rs550057) being distinct is mixed. On the one hand, they survive as independent LHS hits in the Winsorised results (though not in the QN version). On the other hand, they are in reasonably-high LD in our White British sample (squared correlation of 65%, as noted above) and are seen to interact with the exact same variants (and with the same sign) of *TREH*, *ASGR1*, *FUT2* and *FUT6*. Further research is needed to understand whether these two SNPs have distinct biological functions in the determination of serum ALP levels.

Third, we note that the *ASGR1* variant rs186021206 interacts not only with the two *ABO* mutations but also with rs281380, a very common (MAF of 35.8%) variant downstream of *MASMTR*. In turn, rs281380 interacts also with *ABO* variant rs550057 and *TREH* variant rs10892251. While rs281380 is physically closest to *MASMTR* (distance of 8.5 kb), the gene reported as associated with this SNP in the VEP database, the evidence from eQTL (expression quantitative trait loci) studies aggregated by the Open Target Genetics Variant-to-Gene (V2G) pipeline [205, 206] assigns this variant firstly to *FUT2*. Moreover, its squared correlation with *FUT2* synonymous variant rs681343 is 51.5%. We may therefore consider the interactions involving rs281380 as the LHS SNP with *ABO*, *TREH* and *ASGR1* to correspond in fact to interactions between *FUT2* and these three genes. This would then replicate the interaction

between *TREH* rs12225548 and *FUT2* rs281379 previously reported by Lyon et al. [185], given also that the LD between these two variants and our corresponding variants rs10892251 and rs281380 is 96.8% and 57%, respectively.[7]

Again considering the rs186021206 SNP, which we have seen interacts with both *ABO* and *MAMSTR*/*FUT2* and is a strong proxy for the 12-base-pair deletion discussed above, we highlight that the strong functional relevance of this *ASGR1* deletion, combined with the large effect sizes of rs186021206's interactions with *ABO* and *MAMSTR*/*FUT2* (-0.22, -0.26 and 0.16 standard deviations for *ABO* SNPs rs550057 and rs635634, and *MAMSTR*/*FUT2* SNP rs281380, respectively), emphasise the likely functional importance of this novel interaction in the mechanistic system underlying serum ALP levels.

Lastly, we note that interactions involving loci at *ALPL*, the gene encoding tissue-non-specific ALP, are only seen for one somewhat-rare SNP (rs149356924 with MAF of 0.4%, covered above). The interactions we identify instead appear to reflect functional interdependencies between genes that do not directly encode ALP proteins but instead function together with the *ALP* genes in determining serum ALP levels.

5.4 SUMMARY

In this first results chapter, we have analysed the output of our method to detect interactions on the widely-used blood biomarker ALP.

We saw first that our approach produces sensible results for this non-simulated phenotype, with the interaction GWASs showing no inflation and a different signal signature to the original GWASs.

[7] These correlations were computed in the 'GBR' population of the 1000 Genomes Project using the Ensemble LD Calculator.

On closer examination, we were able to identify several loci that interact with a PGS for this trait, and for some of these could detect specific loci driving the interactions. We recovered the three previously-know interaction pairs (importantly, we identified an ABO blood group-tagging mutation as interacting with the expected positive sign with the main non-functional *FUT2* mutation in European populations) and discovered several new interaction pairs.

This first analysis also confirmed that LOCO results are likely more robust than those for the full PGS, while QN residuals similarly seem less prone to false positives than their Winsorised counterparts. This was expected from the simulation results described in the previous chapter and will inform our analysis of the remaining traits, to which we now turn.

Table 5.1: Independent 'LHS' hits with signed log-transformed p-values from SNP × PGS GWAS of ALP with QN residuals.

The ID column shows a 'positional' identifier for each SNP (CHR:POS:REF:ALT), the A1 column the minor allele whose effect is measured in the GWAS and the Consequence and Gene columns functional information taken from the VEP dataset. The p-values shown in the final two columns are log-transformed and have the same sign as the corresponding coefficient.

rsID	ID	A1	MAF	PGS	Consequence	Gene	SNP × PGS	Orig.
rs77569513	1:21449335:G:A	A	0.009	LOCO	Intron var.	*EIF4G3*	12.4	-3.3
rs10465542	1:21859507:A:G	G	0.222	LOCO	Intron var.	*ALPL*	-22.8	-27.7
rs149356924	1:21875506:T:C	C	0.004	LOCO	Intron var.	*ALPL*	15.7	-46.4
rs573303250	1:22138578:G:A	A	0.001	LOCO	Upstream gene var.	*LDLRAD2*	10.8	-19.1
rs145223095	1:22361255:C:T	T	0.008	LOCO	Downstream gene var.	*LINC00339*	14.4	-28.5
rs635634	9:136155000:C:T	T	0.186	LOCO	Upstream gene var.	*ABO*	49.1	-2348.5
rs578180252	1:21826729:G:A	A	0.019	Full	Intergenic var.	–	13.9	4.3
rs190334281	1:22726028:T:A	A	0.005	Full	Intergenic var.	–	18.8	-7.3

Table 5.2: Independent 'LHS' hits with signed log-transformed p-values from SNP × PGS GWAS of ALP with Winsorised residuals.

rsID	ID	A1	MAF	PGS	Consequence	Gene	SNP × PGS	Orig.
rs1976403	1:21766453:A:C	C	0.408	LOCO	Upstream gene var.	*NBPF3*	-21.7	719
rs75746521	1:21854901:T:C	C	0.152	LOCO	Intron var.	*ALPL*	-20	48.7
rs149356924	1:21875506:T:C	C	0.004	LOCO	Intron var.	*ALPL*	7.9	-28.1
rs1697405	1:21904206:C:T	T	0.396	LOCO	3'-UTR var.	*ALPL*	19.7	-303.3
rs143104384	1:76767763:C:T	T	0.035	LOCO	Intron var.	*ST6GALNAC3*	7.8	1.1
rs550057	9:136146597:C:T	T	0.257	LOCO	Intron var.	*ABO*	103.6	-2336.3
rs635634	9:136155000:C:T	T	0.186	LOCO	Upstream gene var.	*ABO*	99.1	-2026.5
rs281380	19:49214470:T:C	C	0.358	LOCO	Downstream gene var.	*MAMSTR*	-12.2	289.4
rs141920409	1:165285620:C:T	T	0.005	Full	Intron var.	*LMX1A*	7.6	0.5
rs183514251	4:135666299:T:C	C	0.001	Full	Intergenic var.	–	7.5	0
rs117640538	22:38819061:C:T	T	0.005	Full	Upstream gene var.	*RP3-434P1.6*	7.4	0.5

Table 5.3: Independent pairwise interaction hits for ALP with QN residuals. *(continued on the next page)*

The Beta and P-value columns show the coefficient and associated log-transformed p-value, respectively, of the interaction term of the two SNPs in a regression in which the main effects (and the squares of their allele counts) are also present.

	LHS SNP					RHS SNP					Beta	P-value
PGS	rsID	Chr.	MAF	Consequence	Gene	rsID	Chr.	MAF	Consequence	Gene		
LOCO	rs77569513	1	0.009	Intron var.	*EIF4G3*	rs144406757	3	0.012	Intron var.	*NAALADL2*	-0.56	7.86
LOCO	rs77569513	1	0.009	Intron var.	*EIF4G3*	rs182976181	8	0.007	Intron var.	*UNC5D*	-0.64	7.64
LOCO	rs77569513	1	0.009	Intron var.	*EIF4G3*	rs7329258	13	0.403	Intron var.	*ENOX1*	0.12	7.87
LOCO	rs77569513	1	0.009	Intron var.	*EIF4G3*	rs1801248	13	0.042	Synonymous var.	*ATP7B*	-0.28	7.49
LOCO	rs77569513	1	0.009	Intron var.	*EIF4G3*	rs3918095	14	0.022	Intron var.	*CCNK*	-0.41	8
LOCO	rs149356924	1	0.004	Intron var.	*ALPL*	rs12636326	3	0.26	Regulatory region var.	–	0.19	7.6
LOCO	rs149356924	1	0.004	Intron var.	*ALPL*	rs62297580	4	0.065	Intron var.	*FIP1L1*	-0.34	7.68
LOCO	rs149356924	1	0.004	Intron var.	*ALPL*	rs7684999	4	0.057	Regulatory region var.	–	-0.35	7.55
LOCO	rs149356924	1	0.004	Intron var.	*ALPL*	rs17123355	12	0.062	Intergenic var.	–	0.36	8.6
LOCO	rs149356924	1	0.004	Intron var.	*ALPL*	rs149786005	13	0.201	Intron var.	*SGCG*	-0.2	7.72
LOCO	rs149356924	1	0.004	Intron var.	*ALPL*	rs147817132	13	0.034	Intergenic var.	–	-0.45	7.35
LOCO	rs149356924	1	0.004	Intron var.	*ALPL*	rs8065491	17	0.45	Intergenic var.	–	-0.17	7.95
LOCO	rs573303250	1	0.001	Upstream gene var.	*LDLRAD2*	rs76571364	5	0.057	Intergenic var.	–	-0.7	9.08
LOCO	rs573303250	1	0.001	Upstream gene var.	*LDLRAD2*	rs9297078	6	0.441	Intergenic var.	–	-0.29	7.43
LOCO	rs573303250	1	0.001	Upstream gene var.	*LDLRAD2*	rs13229792	7	0.135	Intron var.	*DOCK4*	-0.46	8.86
LOCO	rs573303250	1	0.001	Upstream gene var.	*LDLRAD2*	rs34819186	8	0.473	Intron var.	*TS–RE1*	0.3	7.83
LOCO	rs573303250	1	0.001	Upstream gene var.	*LDLRAD2*	rs4745057	9	0.251	Intron var.	*TRPM3*	-0.34	8.38
LOCO	rs573303250	1	0.001	Upstream gene var.	*LDLRAD2*	rs7043546	9	0.475	Intron var.	*GLDC*	-0.31	8.14

Table 5.3: *(continued from the previous page)*

	LHS SNP					RHS SNP					Beta	P-value
PGS	rsID	Chr.	MAF	Consequence	Gene	rsID	Chr.	MAF	Consequence	Gene		
LOCO	rs573303250	1	0.001	Upstream gene var.	*LDLRAD2*	rs2796866	13	0.331	Intergenic var.	–	-0.35	9.96
LOCO	rs573303250	1	0.001	Upstream gene var.	*LDLRAD2*	rs9540399	13	0.132	Intergenic var.	–	-0.48	9.43
LOCO	rs573303250	1	0.001	Upstream gene var.	*LDLRAD2*	rs12884672	14	0.282	Intron var.	*RP11-1152H15.1*	0.33	7.58
LOCO	rs573303250	1	0.001	Upstream gene var.	*LDLRAD2*	15:76145831:TTA:T	15	0.036	Intron var.	*UBE2Q2*	-0.85	8.04
LOCO	rs573303250	1	0.001	Upstream gene var.	*LDLRAD2*	rs1634419	17	0.366	Upstream gene var.	*SLC5A10*	0.33	9.04
LOCO	rs573303250	1	0.001	Upstream gene var.	*LDLRAD2*	rs72891768	18	0.044	Intron var.	*FHOD3*	-0.73	8.08
LOCO	rs573303250	1	0.001	Upstream gene var.	*LDLRAD2*	rs528777301	21	0.469	Intron var.	*SH3BGR*	0.34	9.99
LOCO	rs573303250	1	0.001	Upstream gene var.	*LDLRAD2*	rs1893383	21	0.241	Intergenic var.	–	-0.36	8.42
LOCO	rs145223095	1	0.008	Downstream gene var.	*LINC00339*	rs534037265	6	0.024	Intergenic var.	–	-0.38	7.33
LOCO	rs145223095	1	0.008	Downstream gene var.	*LINC00339*	rs145058027	8	0.014	Intergenic var.	–	-0.46	8.01
LOCO	rs635634	9	0.186	Upstream gene var.	*ABO*	rs2301455	1	0.441	Intron var.	*C1orf105*	0.03	8.05
LOCO	rs635634	9	0.186	Upstream gene var.	*ABO*	rs2276065	11	0.228	Missense var.	*TREH*	0.05	20.75
LOCO	rs635634	9	0.186	Upstream gene var.	*ABO*	rs681343	19	0.492	Synonymous var.	*FUT2*	-0.1	113.07
LOCO	rs635634	9	0.186	Upstream gene var.	*ABO*	rs12019136	19	0.042	Intron var.	*FUT6*	-0.08	13.56
LOCO	rs635634	9	0.186	Upstream gene var.	*ABO*	rs533852698	21	0.002	Intron var.	*MIS18A*	0.31	7.41
Full	rs578180252	1	0.019	Intergenic var.	–	rs149363855	6	0.002	Regulatory region var.	–	-1.06	7.51
Full	rs190334281	1	0.005	Intergenic var.	–	rs192184224	22	0.006	Intergenic var.	–	-1.06	7.71

Table 5.4: Independent pairwise interaction hits for ALP with Winsorised residuals.

	LHS SNP					RHS SNP					Beta	P-value
PGS	rsID	Chr.	MAF	Consequence	Gene	rsID	Chr.	MAF	Consequence	Gene		
LOCO	rs1976403	1	0.408	Upstream gene var.	*NBPF3*	rs117735761	9	0.033	Intron var.	*FREM1*	-0.04	7.33
LOCO	rs149356924	1	0.004	Intron var.	*ALPL*	rs17123355	12	0.062	Intergenic var.	–	0.28	7.97
LOCO	rs550057	9	0.257	Intron var.	*ABO*	rs554626792	7	0.002	Intron var.	*PDE1C*	-0.25	8.09
LOCO	rs550057	9	0.257	Intron var.	*ABO*	rs2276065	11	0.228	Missense var.	*TREH*	0.04	27.37
LOCO	rs550057	9	0.257	Intron var.	*ABO*	rs186021206	17	0.007	Intergenic var.	–	-0.22	22.14
LOCO	rs550057	9	0.257	Intron var.	*ABO*	rs681343	19	0.492	Synonymous var.	*FUT2*	-0.09	192.71
LOCO	rs550057	9	0.257	Intron var.	*ABO*	rs12019136	19	0.042	Intron var.	*FUT6*	-0.07	15.71
LOCO	rs635634	9	0.186	Upstream gene var.	*ABO*	rs2276065	11	0.228	Missense var.	*TREH*	0.04	21.6
LOCO	rs635634	9	0.186	Upstream gene var.	*ABO*	rs186021206	17	0.007	Intergenic var.	–	-0.26	26.27
LOCO	rs635634	9	0.186	Upstream gene var.	*ABO*	rs681343	19	0.492	Synonymous var.	*FUT2*	-0.08	111.56
LOCO	rs635634	9	0.186	Upstream gene var.	*ABO*	rs12019136	19	0.042	Intron var.	*FUT6*	-0.07	13.77
LOCO	rs281380	19	0.358	Downstream gene var.	*MAMSTR*	rs550057	9	0.257	Intron var.	*ABO*	-0.07	106.44
LOCO	rs281380	19	0.358	Downstream gene var.	*MAMSTR*	rs10892251	11	0.228	Intron var.	*TREH*	-0.02	8.45
LOCO	rs281380	19	0.358	Downstream gene var.	*MAMSTR*	rs186021206	17	0.007	Intergenic var.	–	0.16	13.71
Full	rs141920409	1	0.005	Intron var.	*LMX1A*	rs140913877	15	0.056	Intron var.	*AGBL1*	0.27	7.66
Full	rs183514251	4	0.001	Intergenic var.	–	2:177097855:GA:G	2	0.486	Intergenic var.	–	-0.25	7.59

6 Results: remaining traits

In this second results chapter, we give an overview of the interactions we found (both SNP × PGS and pairwise interactions) for the remaining traits examined in this project.

As we saw in the simulation results and also in the previous chapter on ALP, interactions involving the LOCO PGS for QN traits appear to be the most reliable, yielding the lowest number of false positives. Moreover, and as discussed in the second chapter on methods, LD between potentially-interacting variants is a primary concern as it can lead to such false positives. Therefore, in this chapter we focus exclusively on interactions detected with the LOCO PGS, for the QN version of traits and between variants on *different chromosomes*. Further work may allow us to identify additional interactions that do not obey these three requirements but are still not the result of statistical artefacts.

We will begin by analysing the set of independent LHS interactions (i.e., interactions with the LOCO PGS) and will then move on to consider pairwise interactions involving some of these LHS hits.

6.1 DISTRIBUTION OF LHS HITS ACROSS TRAITS

Our statistical approach identified 144 independent LHS hits across 52 distinct traits, which we list in full in Table E.1 in the appendix. We now consider the distribution of these 144 interacting hits across the different traits plotted in Figure 6.1.

We first note that LHS hits are concentrated in molecular biomarkers, and in particular in traits in the blood biochemistry and blood cell count categories. Glycated haemoglobin, total bilirubin and direct bilirubin are the traits with the highest number of LHS hits, in that order, followed by ALP, which we analysed in detail in the previous chapter.

Despite the pre-eminence of interactions affecting biomarker traits, other categories are also represented. Among these, impedance measures of body fat are the most numerous, and fat-related body size measures such as waist circumference, BMI or waist-to-hip ratio also show interactions. We highlight the conspicuous absence of the different measures of height in our list of traits (standing and sitting height, and leg length) for which – and despite their well-known polygenicity – we were not able to find interactions with the PGS.

Finally, we see a few interactions for blood pressure traits (three for pulse rate and one for diastolic blood pressure) and two interactions for ankle spacing width, which may be considered a proxy for the amount or distribution of body fat.

6.2 COMPARISON OF ASSOCIATIONS IN THE ORIGINAL AND INTERACTION GWAS

We proceed by comparing the signal for association of each LHS hit in the original vs. the SNP × PGS interaction GWAS. As discussed before in this thesis, a potential concern when developing a GWAS-based method for interaction testing is that it will

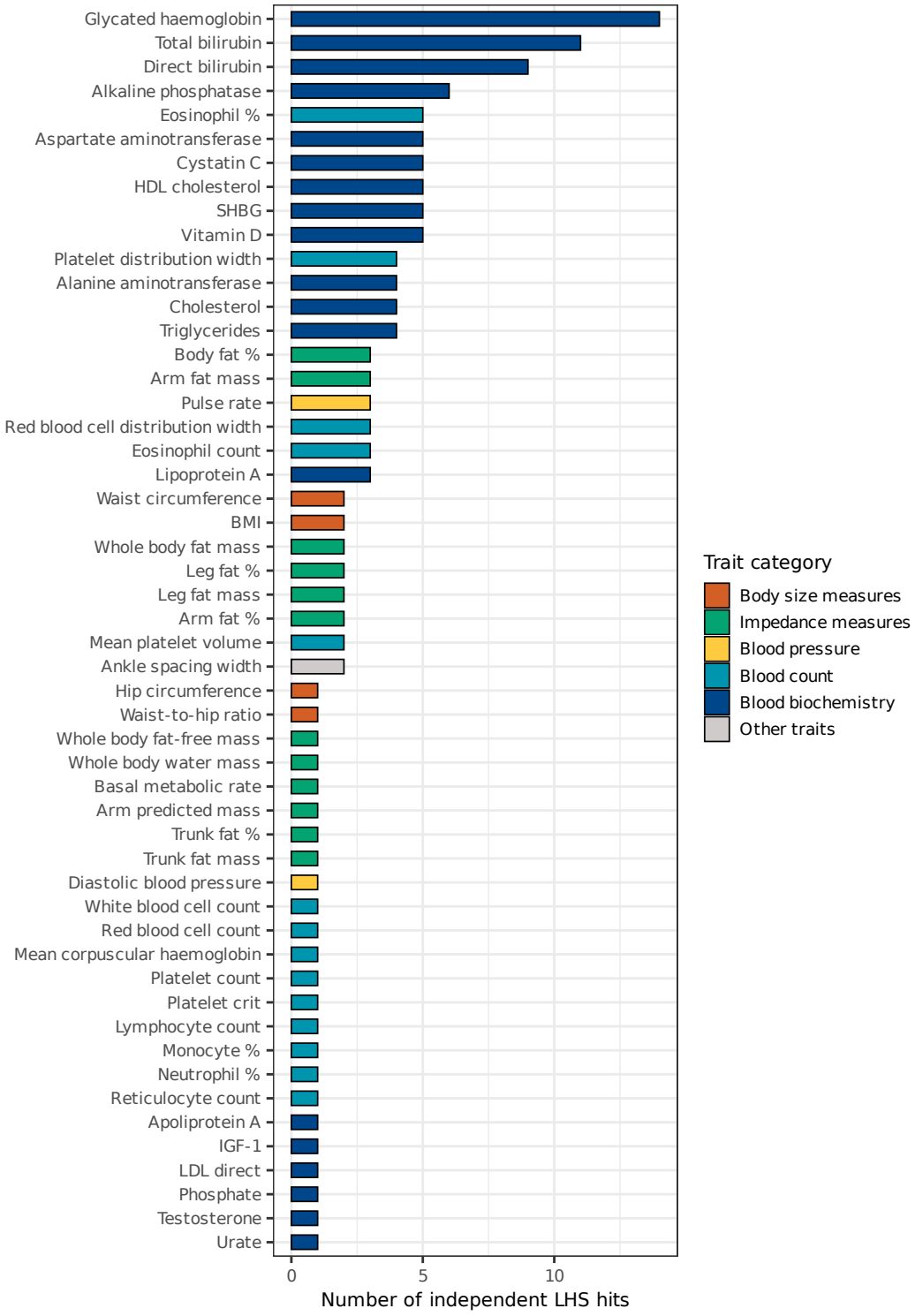


Figure 6.1: Number of independent LHS hits from SNP × PGS interaction GWAS with LOCO PGS and QN residuals by trait. Traits not shown do not have hits.

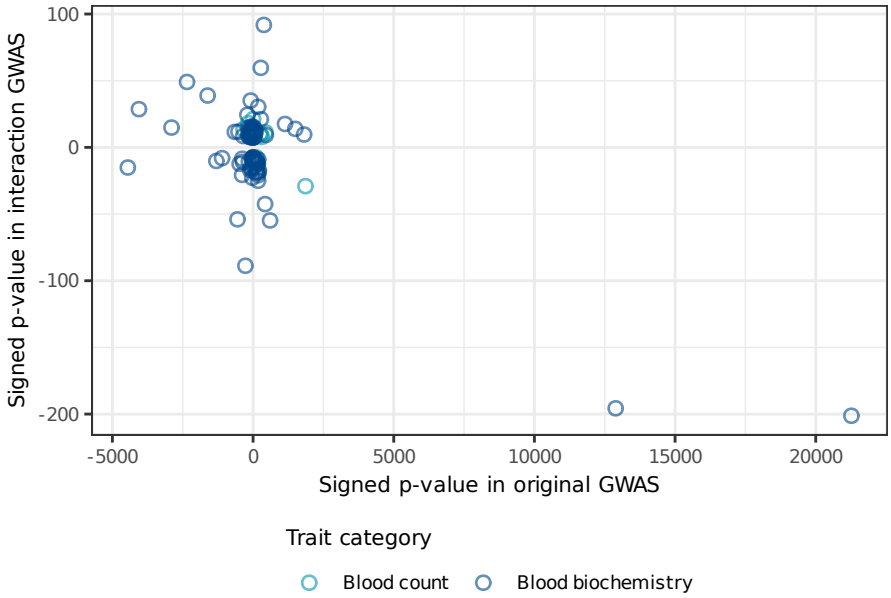
produce results that are merely a truncation of the evidence for direct association of the different SNPs tested. Therefore, in the previous chapter on ALP, we examined, for all SNPs with GWS signal in *either* original or interaction GWASs, the correlation between these signals. We found that there was moderate correlation but clearly differentiated patterns, both across and within chromosomes. In keeping with this approach, we now plot the original against the interaction *signed* log-transformed p-values for all the LHS hits, coloured by trait category, in Figure 6.2.

We see from this figure that there is again no obvious correlation between the original and interaction p-values, and this is also true within individual traits (data not shown). This is as expected if our test is capturing real signal as opposed to a transformed version of the original GWAS associations. We also note that, similarly to what we saw for ALP, some hits show strong evidence of interaction but have weak, or no, signal for direct association with the trait.

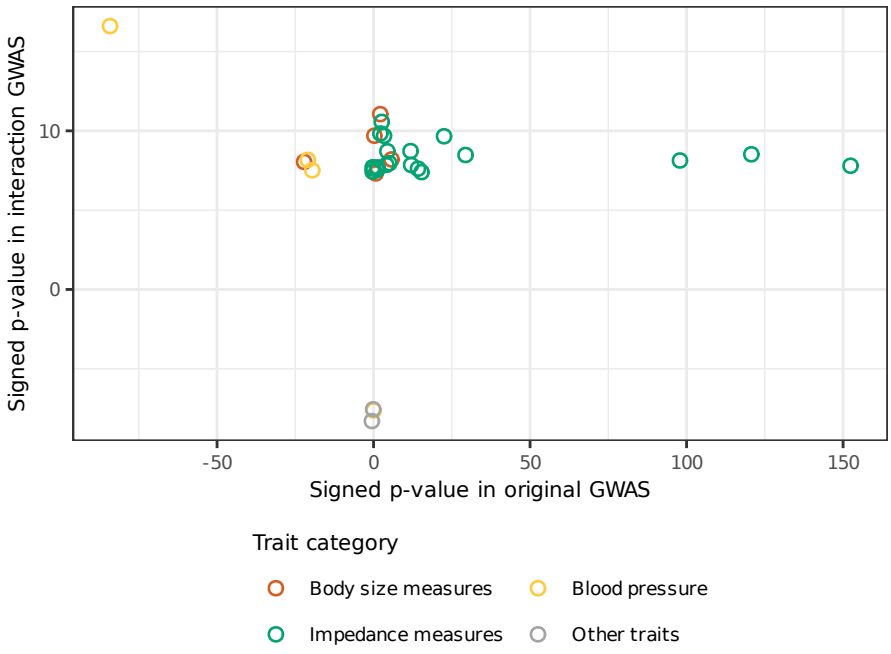
6.3 LHS HITS ARE ENRICHED FOR FUNCTIONAL RELEVANCE

A natural question when examining a moderately-sized set of genetic associations is how functionally relevant they are, and one way to answer this question is to compare the set in question to the genome at large. For this, we need a measure of functional significance for each variant considered. Fortunately, the VEP database of annotations [128] not only includes variant consequences – there are 29 different categories in our full set of 12.9 M variants – but also provides a ranking of these consequences by severity.

In Figure 6.3, we show the distribution of variants by consequence categories for our set of 144 independent LHS hits and for a random sample of 100,000 of the 12.9 M variants that passed our initial filters. (We chose to take a small random sample



(a) Blood count and blood biochemistry traits.



(b) Remaining traits.

Figure 6.2: Scatter plot of signed log-transformed p-values from original and SNP × PGS interaction GWASs for LOCO LHS hits with QN residuals.

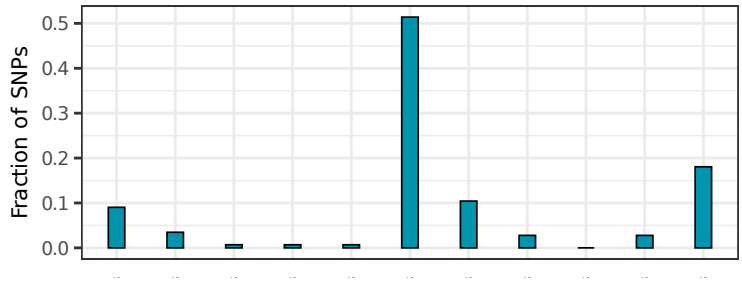
rather than consider the full set because we wish to test for a difference in the two distributions and standard statistical tests for this comparison assume independent distributions in the two groups being compared. By considering only 1% of the full set of variants, we hope to obtain a set of approximately independent variants for comparison with the interaction hits, which are independent within traits and are sufficiently low in number that significant overlap across traits is unlikely.)

From this figure, we can see that the most frequent category in both the set of LHS hits and the sample of variants along the genome is 'Intron variant', followed by 'Intergenic variant'. The plot of the ratio of the fraction of hits in each category of the LHS hits to the full set of SNPs, shown in the bottom panel, shows a marked enrichment of the interaction hits in categories of greater functional severity.

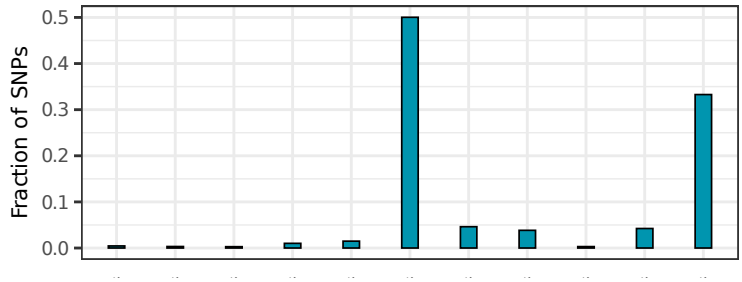
The hypothesis that the set of LHS interaction hits tends to have more severe consequences than the set of variants randomly chosen from the genome can be formally tested by specifying as the *null hypothesis* the assumption of *stochastic equality:* for any two values taken at random from each of the two sets, the probability that the first ranks higher than the second is equal to the probability that the second ranks higher than the first.

A standard nonparametric test of the assumption of stochastic equality is the Mann-Whitney U Test, also known as the Wilcoxon Rank-Sum Test [213–215]. A *one-sided* test of this null hypothesis against the alternative hypothesis that the probability of a LHS interaction hit having a more severe consequence than a randomly chosen variant is greater than its converse yields a highly-significant p-value of 4.65×10^{-7}.

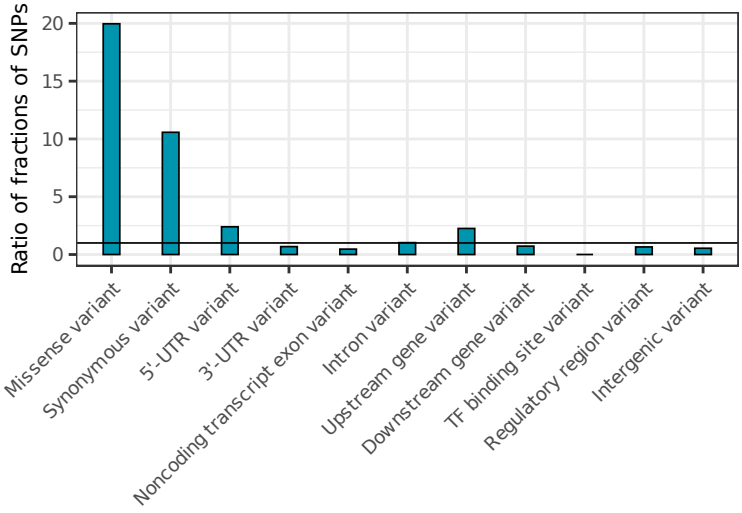
Despite the popularity of the Mann-Whitney U Test, it has recently been remarked in the applied statistics literature that this test makes the very strong assumption of *exchangeability,* which is likely to be violated in practice, in which case the test can yield highly-inflated type I error rates [215]. In the context of this test, exchangeability



(a) 144 LHS hits with LOCO PGS and QN residuals.



(b) Random sample of 100,000 of the full set of 12.9 M SNPs.



(c) Ratio of fractions in LHS hits to random sample of all SNPs.

Figure 6.3: Histograms of distribution of SNPs by VEP Consequence functional categories for the 144 LHS hits with LOCO PGS and QN residuals, for a random sample of all SNPs considered in the analysis, and for their ratio. Consequence categories are sorted from left to right in decreasing order of severity according to the ranking provided by Ensembl [212]. Only categories whose share of the random sample of 100,000 variants is greater than or equal to 0.1% are shown in either plot. The horizontal line in the plot of ratios is set at 1.

means that, under the null hypothesis, the two distributions are not only stochastically equal but are in fact fully identical [215]. We would like to be able to test stochastic equality without making such an assumption as this would allow, for example, for one distribution to have greater dispersion than the other without being 'shifted', something which could reasonably happen in our comparison of interaction hits against the genome.

A more recently proposed test that adequately controls type I errors even if the exchangeability assumption is violated is the Brunner-Munzel Test, also known as the Generalised Wilcoxon Test [215, 216], which is implemented in R in the package `brunnermunzel` [217]. A one-sided test of the same hypothesis as above now yields a higher but still highly-significant p-value of 1.11×10^{-6}.

We therefore conclude that there is strong evidence that the SNP × PGS interactions identified by our approach tend to have more severe consequences than a random set of variants.

6.4 SNP × PGS INTERACTIONS ACCOUNT FOR LITTLE HERITABILITY

Two motivations for searching for epistatic interactions are that these may shed light on the biological mechanisms underlying different traits and diseases, and that they may account for some of the estimated heritability of such phenotypes. It is therefore relevant, once new interactions have been found, to attempt to estimate the proportion of variance of the trait in the population for which they account.

As discussed in the chapter on methods, we do not currently have a well-tested and calibrated procedure for estimating the heritability of our interactions. However, we have computed approximate estimates of heritability for the combined LHS hits

found for each trait based on out-of-sample predictive performance and, while not precise, this can give us a first sense of how much heritability they account for.

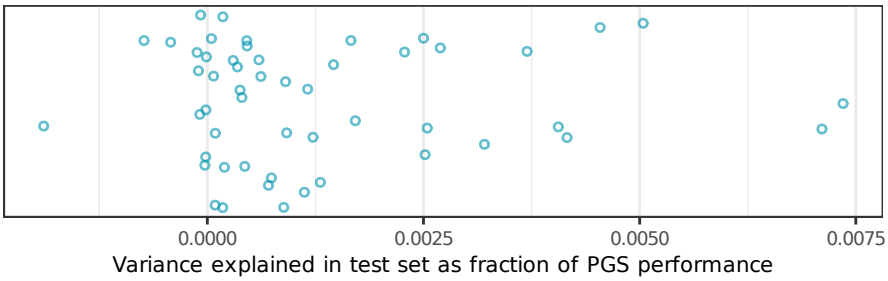
We show three plots displaying the distribution of these estimates for the 52 traits for which we have identified interactions in Figure 6.4. (We recall that it is possible for the heritability estimates produced by this naive approach to be negative since, for the computation of the lower bound, we compare the performance of adding interaction terms to a linear model including the PGS in a held-out test set where the coefficients of the interactions and of the PGS are fixed based on estimates obtained in the training set, and therefore do not increase the degrees of freedom of the model.)

From these plots, it is clear that the LHS hits explain very little heritability even when combined (for the traits that have multiple such interactions) and taken as a fraction of the performance of the PGS on the held-out test set – the two largest estimates are for vitamin D (0.0074) and pulse rate (0.0071), followed by glycated haemoglobin (0.005). However, our (approximate) heritability estimates are nevertheless positive for most traits, which is evidence for a (limited) role for interactions in their genetic architecture.

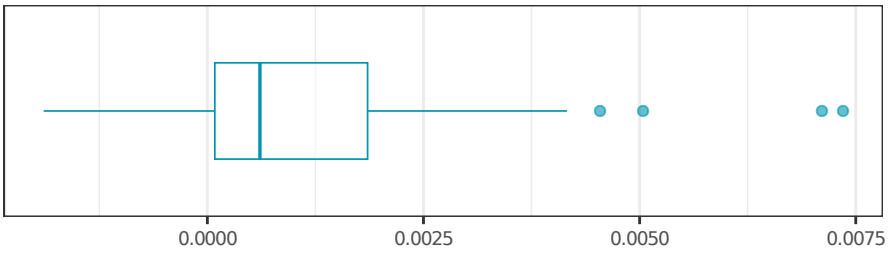
6.5 DISTRIBUTION OF PAIRWISE HITS ACROSS TRAITS

So far in this chapter, we have focused exclusively on the LHS hits, i.e., the independent interacting loci identified through the SNP × PGS interaction GWASs. We now proceed by considering the independent GWS *pairwise* interactions detected through subsequent standard pairwise GWASs.

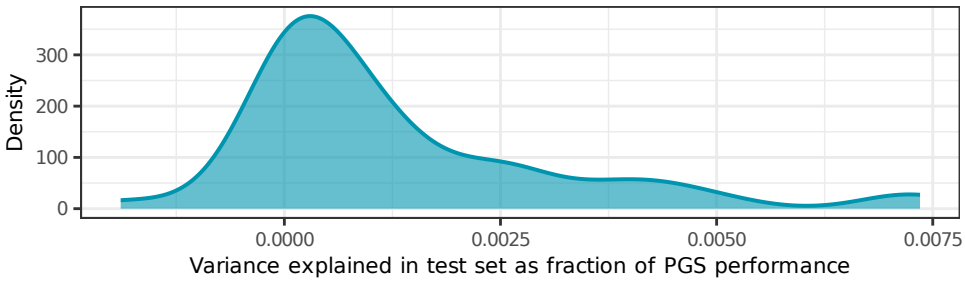
We identified 145 independent pairwise interactions across 26 different traits. The full list of interactions is given in Table F.1 and their distribution by trait is plotted in Figure 6.5.



(a) Scatter plot with jittered points.



(b) Box-and-whiskers plot.



(c) Gaussian kernel density estimate.

Figure 6.4: Distribution of estimates of combined heritability explained by all LOCO LHS hits as a fraction of the performance of the PGS on the held-out test set for traits with QN residuals for which we identified SNP × PGS interactions.

Box plot details: the central line is the median; the lower and upper hinges of the box correspond to the 25th and 75th quartiles, respectively; the upper whisker extends to the largest data point within 1.5 × Inter-quartile Range from the upper box hinge, and similarly for the lower whisker; data points beyond the limits of the whiskers are plotted individually.

From this plot we can see that, generally in line with the LHS results that form the basis of the procedure to identify pairwise hits, the great majority of pairwise interactions were found for molecular biomarkers and in particular for blood bio-

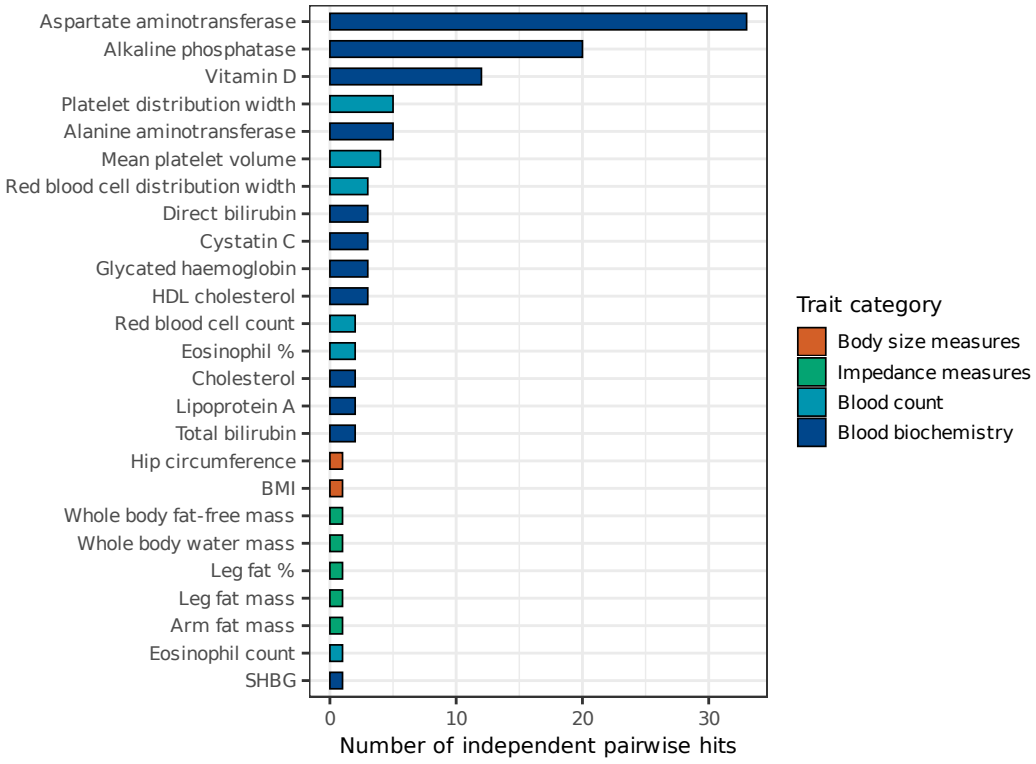


Figure 6.5: Number of independent GWS pairwise interaction hits with LOCO PGS and QN residuals by trait. Traits not shown do not have hits.

chemistry traits. We also found a few pairwise interactions for blood cell count traits (especially platelets, as we will see below), as well as for impedance measures and two fat-related body size traits (hip circumference and BMI).

6.6 HIGHLIGHTS OF PAIRWISE INTERACTIONS

In this final section, we highlight some of the pairwise interactions detected by our statistical approach. In contrast to our analysis of the *ABO–FUT2* interaction for ALP, for which we were able to provide a reasonably-detailed biological interpretation, here we do not attempt to analyse potential mechanisms driving the observed interactions but instead only provide additional context. The brief comments on six different traits

and their interactions written below are therefore only suggestive of the potential functional importance of these hits, which could be explored in further research.

Platelet-related traits: For two traits related to platelets, mean platelet volume and platelet distribution width, we detected multiple interactions that have the *ARHGEF3* intron variant rs1354034 as the LHS SNP. This variant is very common (MAF of 39.9% in our sample) and had previously been associated with multiple traits relating to both platelets and other types of blood cells [202, 218, 219]. Interestingly, in our original GWAS, it is extremely-highly associated (log-transformed p-value of 1862.8) with mean platelet volume but shows no direct association with distribution width (log-transformed p-value of 0.1). This SNP is an eQTL for *ARHGEF3* [205, 206], which is likely to be one of the two driving genes in these interactions as it has been associated with blood cell maturation (specifically, the formation of erythroids); its relevance for bone cell biology is also well known [220].

The RHS SNPs for these interactions are linked to a few different genes. For mean platelet volume, these are *RNPEPL1*, *IQGAP2* (a missense mutation), *WDR66* and *RUNX1*. The specific SNPs involved in the first two interactions have both been previously associated with platelet counts [219, 221]. *WDR66* encodes a protein belonging to the WD family and which is thought to be involved in determining mean platelet volume [201]. Finally, *RUNX1* is a transcription factor that is believed to be important in normal haematopoiesis [201] – a monoallelic mutation of this gene causes *RUNX1* familial platelet disorder, a condition whose characteristics include platelet functional defects [222]. This *RUNX1* SNP shows a GWS direct signal for platelet count and crit, a non-GWS but moderate signal for mean platelet volume (log-p-value of 6.08) and no direct association with platelet distribution width.

Turning now to platelet distribution width, the RHS genes appear to be *RP5-1086L22.1*, *CTB-111H14.1*, *ODF3*, *WDR66* again and *BET1L*. The mutation involved in the interaction for *RP5-1086L22.1* has been linked with red blood cell-related traits [221, 223], the *CTB-111H14.1* SNP with platelet distribution width [223] and mean platelet volume and count [219, 221, 223], and the *ODF3* and *BET1L* variants also with platelet crit [223] in addition to mean platelet volume and count [219, 221, 223].

The fact that all the RHS SNPs have been linked to platelets or other blood cell types gives us confidence that these interactions represent true functional interdependencies between different genes rather than being statistically spurious.

Eosinophil count: We detected a negative interaction between a variant upstream of *IL33* and a missense variant of *ALOX15*. Both SNPs have highly-significant direct effects on the trait, the former with a positive and the latter a negative sign – this negative interaction can therefore be seen as reducing the positive effect of the *IL33* variant, or as further increasing in magnitude the negative effect of the *ALOX15* mutation. *IL33* is involved in the maturation of eosinophils while *ALOX15* encodes an enzyme that has been shown to regulate inflammation and immunity processes [201]. We tentatively suggest that this interaction may shed light on the aetiology of asthma as a recent study *in mice* showed that *ALOX15* plays a role in reducing *IL33*-induced airway eosinophilic inflammation [224].

Alanine aminotransferase and lipoprotein(a): Alanine aminotransferase (ALT) is an enzyme found in many tissues but primarily in the liver that plays an important role in amino acid metabolism and gluconeogenesis [225]. Serum levels of ALT are the most widely used biomarker of hepatic health in the clinic [226]. Lipoprotein(a)

(LPA) is a lipoprotein produced primarily by the liver [227] whose levels are highly heritable [228]. Lipoproteins play a crucial role in the transport of lipid proteins in circulation [229]. The function of LPA specifically is not well understood [227, 229] but is thought to include wound healing, tissue repair and vascular remodelling [230]. LPA is used as a clinical biomarker as high levels of this lipoprotein are recognised as a risk factor for atherosclerotic cardiovascular disease (ASCVD) [228, 231].

For ALT, we identified a positive interaction between *APOE* missense variant rs429358 and rs2176521, an intergenic variant in chromosome 3. For LPA, we detected a negative interaction between a different *APOE* missense variant, rs7412, and rs571848809, an intron variant of the *LPA* gene. The *APOE* variants rs429358 and rs7412 are well know for their strong associations with Alzheimer's Disease (AD) risk, with rs429358 increasing risk and rs7412 being protective [232].

Again very tentatively, we propose that this interaction may partly explain a recent suggestion in the literature of a potential link between AD and liver function [233].

Aspartate aminotransferase: This is the trait for which we find the greatest number of independent pairwise interactions (43). Forty-one of these have the same LHS SNP, the *CNNM1* intronic deletion variant at position 101,101,889 on chromosome 10, with reference allele GT and alternative G.[1] This variant is rare (MAF of 0.2% in our sample) but has strong signals in both the original and SNP × PGS interaction GWASs (log-transformed p-values of 84.2 and 35.1, respectively). The second most significant hit at this locus in the SNP × PGS GWAS, intron variant rs527506472 which is in high LD (squared correlation of 91.7%) with the lead SNP, has also been previously associated with aspartate aminotransferase levels [234].

[1] We were not able to find an rsID for this SNP.

We chose to highlight the group of interactions discussed in the previous paragraph because they share the same LHS SNP, which suggests a large network of interacting loci. However, given the very low frequency of this variant, we cannot be fully confident in these results without further testing, potentially in an external cohort.

A second interaction for this trait that involves only two SNPs but for which the functional evidence is perhaps more convincing is a negative interaction between rs3747207, a *PNPLA3* intron variant, and rs13150834, a variant upstream of *MIR5705*. The first SNP has been found to strongly associate with levels of aspartate aminotransferase [235], including in the context of high alcohol intake [236]. The RHS SNP, rs13150834, shows a highly-significant signal of association with this trait in our own original GWAS (log-transformed p-value of 41.6).

Cholesterol: Our final highlight is (total) cholesterol and HDL-cholesterol, for which we identify two and three pairs of interacting loci, respectively. Beginning with total cholesterol, we have first a negative interaction between a *BUD13* deletion (TA to T) at position 116623213 of chromosome 11[2] and rs56043443, an intron variant of *CTC-559E9.6*, both of which are common (MAFs of 13.9% and 15.1%, respectively). The *BUD13* deletion has previously been associated with triglycerides [237] and shows a very strong (positive) direct signal of association with cholesterol in our GWAS (log-p-value of 79.95), while the *CTC-559E9.6* SNP rs56043443 (which ANNOVAR links instead to *ZNF56*) does not have a direct association.

The second interaction, also with a negative coefficient, is between *APOC1* SNP rs814573 and *ZNF259* SNP rs964184, another two common variants (MAFs of 18.8% and 13.2%, respectively). rs814573 has previously been found to associate very

[2]The UKB does not provide an rsID for this variant but other sources suggest rs66505542 and rs398017699 have been used to refer to this deletion.

strongly with LDL-cholesterol [238] (increasing its levels) and has an extremely low p-value (log-p-value of 453.84) for direct association in our own GWAS for total cholesterol, and *APOC1* more generally is known to be relevant for this phenotype [239]. rs964184, a 3'-UTR variant of *ZNF259*[3] (a gene which is known to play a key role in the metabolism of high-density and very low-density lipoproteins [201]), has previously been found to have a negative association with LDL-cholesterol levels [238], but a positive association with triglycerides and HDL-cholesterol [240], and has a strong positive association with cholesterol in our data. Given the common frequency of both SNPs in each interaction pair and their previous association with lipid-related traits, these two interactions seem likely to constitute real signals.

Continuing with HDL-cholesterol, we highlight two interactions involving missense variants as LHS SNPs: one between deleterious *CD300LG* missense variant rs72836561 and *STX6* intron variant rs147648373 (with a large negative effect of -0.26 standard deviations (s.d.)), and the other between *APOE* variant rs7412 and NAALADL2 intron variant rs62284708 (with a large positive effect of 0.36 s.d.). Beginning with the first interaction, we note that, while rs72836561 is somewhat rare (MAF of 3.2%), it has previously been found to be the lead variant in its region affecting both triglycerides and HDL-cholesterol [241]; this missense mutation affects the immunoglobulin domain of CD300LG, which binds to lymphocytes and is expressed in endothelial cells – however, the function of *CD300LG* in lipid metabolism is not yet well understood [241]. Moreover, the protein encoded by the *STX6* gene with which it interacts binds cholesterol, regulating the transport of cholesterol-rich domains to and from the cell surface [242]. The *CD300LG* variant in this interaction has a strong negative effect on HDL-cholesterol levels (−0.18 standard deviations with log-p-value of 139.07) while the *STX6* mutation does not have a direct association with the trait.

[3]This gene is also known by the alias *ZPR1*.

Lastly, we note regarding the second interaction that its LHS SNP, *APOE* variant rs7412, is the same well-know mutation that is protective against Alzheimer's disease and that we identified above as interacting with an *LPA* variant to affect LPA levels. rs7412 is known to have a large positive effect on total cholesterol levels and has been linked to familial type III hyperlipoproteinemia [241], and the *APOE* gene that it affects plays an important role in maintenance of cholesterol levels through the reverse transport pathway that directs excess HDL cholesterol to the liver for elimination [243]. This variant has a significant positive direct effect on HDL-cholesterol in our GWAS, while the *NAALADL2* variant with which it interacts does not.

6.7 SUMMARY

In this chapter, we have given an overview of the results of our interaction testing framework for all traits other than ALP, which we examined in detail in the preceding chapter.

Focusing exclusively on the hits obtained from an interaction with the LOCO PGS and for the QN versions of the traits, we saw first that we were able to identify 144 independent LHS hits across 52 different traits. The great majority of these interactions affects molecular biomarker traits: 86 of 144 affect blood biochemistry measurements and another 26 impact counts of different blood cell types. However, there are also several interacting loci with an effect on body fat traits. We noted that our method did not detect any interactions for height-related traits in spite of their great polygenicity.

We examined the correlation between the evidence of association in the original and SNP × PGS GWASs for all the LHS hits and found no clear pattern of strong

dependence between the two, an observation consistent with our statistical approach capturing real interactions.

By comparing the distribution of functional consequences ranked by severity for the LHS hits against the genome as a whole, we found that these interacting loci are enriched for functional relevance. Finally, we saw that, even when aggregated within each phenotype, the LHS loci account for little heritability.

We proceeded by considering independent pairwise interaction hits, of which we found 145 affecting 26 different traits. Naturally, given the results of the SNP × PGS interaction testing, the large majority fell under the blood biochemistry trait category (120 out of 145), followed by blood cell count traits (18).

We concluded our analysis by highlighting some of the pairwise interactions we detected. Key highlights were several pairwise hits for two different platelet-related traits where the genes associated with both LHS and RHS SNPs had previously been linked to phenotypes in this vein, suggesting these statistical findings are credible and worthy of further investigation. A second interesting finding was an interaction between *IL33* and *ALOX15* for eosinophil count which may have a potential connection to the aetiology of asthma. Finally, we saw interactions involving two key *APOE* AD risk modifiers for liver proteins ALT and LPA, interesting in the light of recent suggestions of a connection between liver function and this disease.

7 Interactions involving *HLA* genes

Contributions to this chapter: *The imputation of* HLA *alleles for the UKB used for the interaction analysis described in this chapter was undertaken by Esther Ng. This work was jointly supervised by Yang Luo and Simon R. Myers.*

In this chapter, we describe in-progress work on a smaller project in which we apply our interaction testing framework to genetic variation in human leukocyte antigen (*HLA*) genes. We will begin by introducing the major histocompatibility complex (MHC) – in which *HLA* genes are located – describing its structure and function, its unique characteristics related to genetic variation, and its known associations with many complex diseases. We will then report on our imputation of this region for the UKB and on the results of interaction testing involving *HLA* genetic variants.

7.1 BACKGROUND

7.1.1 *Structure and function*

The MHC is a gene-dense, highly-polymorphic genomic region located in the short arm of chromosome 6 (6p21). Genes located in this region, including *HLA* genes, encode proteins that are key to the functioning of the immune system, controlling

antigen processing and presentation and thus regulating both innate and adaptive immune responses [244, 245].

More than 200 genes are located in the MHC region [246]. These are divided into three subclasses (denoted I, II and III) but most research into this region has focused on the first two as these are responsible for the key process of antigen presentation to T cells and are highly polymorphic. Class I includes classical *HLA* genes *HLA-A*, *HLA-B* and *HLA-C*; class II includes *HLA-DRB1*, *HLA-DQA1*, *HLA-DQB1*, *HLA-DPA1* and *HLA-DPB1* [247]. These eight classical genes will be the focus of our analysis below.

Both class I and class II genes encode glycoprotein molecules that bind peptides and present them to receptors on T cells [244, 247]. Class I molecules are expressed by nearly all nucleated cells; they bind antigenic epitopes derived from *intracellular* pathogens and present antigens to $CD8^+$ T cells. By contrast, class II molecules are primarily expressed by cells with a dedicated antigen-presenting function such as dentritic cells, macrophages, B cells or thymic epithelial cells; these molecules primarily bind peptides of *extracellular* origin and present them to $CD4^+$ T cells [248, 249]. The biochemical tripartite interaction between HLA cells, peptides and T-cell receptors is crucial in enabling the adaptive immune system to mount a response to infection while maintaining self-tolerance and so preventing autoimmune disease [247].

7.1.2 *Genetic variation*

In addition to its high gene-density, two other defining genomic characteristics of the MHC locus are the strong LD observed within it and its extreme polymorphism: more than 15,000 alleles have been identified for class I and II genes [247]. This diversity is thought to maximise the probability that at least a small number of individuals in the human population would be able to survive a new infection by mounting an effective immune response [247]. Polymorphisms in *HLA* genes often result in

nonsynonymous amino acid changes in the peptide-binding groove of the encoded molecules that specify which foreign peptides are presented to T-cell receptors [247, 249]. The extreme genetic diversity in this region suggests the action of a strong balancing selective pressure, but the driving mechanisms of this diversity are still poorly understood [244].

HLA polymorphisms are reported according to a specific, standardised nomenclature whose development is overseen by the WHO Nomenclature Committee for Factors of the *HLA* System, established in 1968 [250]. The need for a common naming system was recognised from the early days of the field, and this nomenclature evolved in parallel with methods to detect *HLA* variation [250]. The designation of each *HLA* locus begins with the gene name followed by an asterisk (e.g., *HLA-C**). This is followed by allele designations that are unique numbers comprised of up to four fields (sets of digits) separated by colons. The first two fields, each with two digits, are the most commonly used. The first of these two-digit fields refers to the *type* (often corresponding to the serological antigen) and the latter to the *subtype,* a unique identifier of the amino acid sequence [244, 250]. We will analyse genetic variation data in this format, i.e., *HLA* alleles determined at the two-field resolution.

As noted above, *HLA* genes are in tight LD, and it has been observed that short blocks of conserved DNA sequence containing specific *HLA* allele combinations (haplotypes) can be shared by many unrelated individuals in a population [251]. Longer blocks, termed 'conserved extended haplotypes', also exist, sometimes spanning several megabases [244, 251]. The mechanisms driving the formation and maintenance of these haplotypes is not fully understood and multiple explanations have been proposed. These include instances of geographic isolation in a population's history, lower recombination at this locus and the maintenance of functionally-coordinated sets of

alleles [244]. It is thought that these haplotypes originally arose through imperfect sequence duplication combined with insertions and deletions [245, 252].

For our analysis, we will define two sets of haplotypes for classes I and II separately: one set for the *HLA-A*, *HLA-B* and *HLA-C* genes, and another for the *HLA-DRB1*, *HLA-DQA1*, *HLA-DQB1*, *HLA-DPA1* and *HLA-DPB1* genes. We will analyse these jointly with alleles at the four-digit resolution as explained below.

7.1.3 *Disease associations*

Genetic variation in *HLA* genes has been associated with a large number of diseases. While this region makes up only 0.3% of the human genome, it was reported to account for 6.4% of the significant SNP associations in the EMBL-EBI GWAS Catalog [2, 253]. *HLA* variants have been linked to nearly all autoimmune diseases as well as to multiple infectious and inflammatory diseases [254] – in line with the crucial role played by this region in the functioning of the human immune system – but also to cancers and neurological diseases.

The first disease to be linked to *HLA* variation was Hodgkin's lymphoma in 1967 [254–256]. Other early findings were associations with lupus [257, 258] and coeliac disease [258, 259]. More recent research on the latter has identified variation within *HLA* genes as driving this association: the DQ2 molecule encoded by alleles *DQB1*02:01* and *DQA1*05:01*, which was found to be present in 95% of coeliac patients compared to 20–30% of controls in a sample of Northern European ancestry [260].

Two of the strongest disease associations with *HLA* variation are for two autoimmune diseases, ankylosing spondylitis (AS) and Behçet's disease. For AS, the *B*27* allele has an odds ratio larger than 100 [253] while, for Behçet's disease, *B*51* accounts for 32–52% of cases [253, 261].

Other diseases whose connection to *HLA* variation has been extensively studied include type 1 diabetes (T1D), rheumatoid arthritis (RA) and multiple sclerosis (MS). For T1D, the class II loci *HLA-DRB1* and *HLA-DQB1* have been established as having the strongest associations [262]: risk variant *DQB1*03:02*, which encodes an alanine at DQB1 position 57 and whose link to T1D was first identified in 1987 [263], was recently found to explain 15.2% of the phenotypic variance, with an odds ratio of 6.29, in a European sample [264]; haplotypes of DR3 and DR4 are also known to increase risk, while a DR2 haplotype is protective [265]. For RA, the association of *HLA* variation to disease risk has been shown to be almost completely explained by positions 11, 71 and 74 at DRB1, by a single SNP in *HLA-B* and by DPB1 [266]. For MS, an early association was found in the 1970s between class II serotype DR2 [267] which was later fine-mapped to the extended haplotype *DRB5*01:01~DRB1*15:01~DQA1*01:02~DQB1*06:02* [268, 269]; the *DRB1*15:01* allele is currently thought to be the strongest genetic risk factor for this disease [270, 271], with an odds ratio of 3.92 in individuals of European ancestry [270].

With respect to infectious diseases, multiple associations have been identified between *HLA* variation and the development of AIDS for carriers of HIV. Class I alleles *B*35*, *C*04* and *C*07* are linked to rapid development of AIDS in Europeans [272, 273]; conversely, alleles *B*57:01*, *B*27:05*, *B*14/C*08:02*, *B*52* and *A*25* delay disease progression [249, 273]. Finally, very recently a strong association has been found between *B*15:01* and asymptomatic SARS-CoV-2 infection [274].

We conclude this brief section on disease associations by noting that, while much evidence has been gathered on the statistical connection between variation in *HLA* genes and a vast array of diseases, the biological mechanisms linking *HLA* variants to disease aetiology remain poorly understood. Perhaps the most common model to explain this association is that the biochemical interactions between HLA

molecules, the peptides which they bind and the T-cell receptors to which they are presented – which underlie discrimination between self and nonself – are disrupted in immunopathology [247].

7.1.4 *Epistasis*

In addition to the disproportionately large number of disease associations located in this small segment of the genome, another interesting finding from studies of *HLA* and disease has been a series of epistatic interactions involving *HLA* alleles that affect disease risk.

The most well-established example is perhaps the interaction of allele *B*27* with *ERAP1* missense variant rs30187 for AS: this *ERAP1* SNP increases AS risk only in *B27*-positive individuals [275]. This interaction is thought to be informative about disease aetiology since the primary function of the ERAP1 enzyme is to trim peptides to the ideal length for presentation by the HLA-B molecule [68]. A second interaction between this variant and another *HLA* allele, *B*40*, has also been identified [276].

Polymorphisms at *ERAP1* have been found to interact with other *HLA* loci for at least two other autoimmune diseases: Behçet's disease, for which *ERAP1* SNP rs17482078 increases risk only in *B*51*-positive individuals [277], and psoriasis, for which *HLA-C* variant rs10484554 and *ERAP1* mutation rs27524 interact to increase disease risk [87].

A second important non-MHC locus found to interact with *HLA* variation to affect disease risk is the cluster of killer-cell immunoglobulin-like receptors (*KIR*) genes. For AIDS, the activating *KIR* allele *KIR3DS1* is associated with delayed progression to the disease in individuals infected with HIV-1 who also carry the *B4-80Ile* allele [278]. For psoriatic arthritis, activating *KIR* alleles *KIR2DS1* and *KIR2DS2* were found to disease risk more when the *HLA-C* groups 2 and 1, respectively, were

absent [279]. For T1D, positivity for the inhibitory *KIR2DL1* allele interacts with *HLA-C* alleles to increase risk, while positivity for *KIR2DL1* interacts with *HLA-C* to decrease risk; positivity for *KIR3DL1* interacts with *HLA-Bw4* alleles to reduce risk [280].

Evidence for non-MHC interactions with *HLA* alleles or genes in the MHC region more generally exists for various other diseases. For example, for systemic lupus erythematosus (SLE), risk-increasing interactions were found between *CTLA4* risk allele rs231775 and both the rs3131379 SNP of *MSH5* and the rs1270942 SNP of *CFB* [281], where the last two genes are located in the MHC class III region [244]. As a second example, an epistatic interaction between *HLA-G* 3'-UTR variant rs9380142 (known to play a role in mRNA stability) and *LILRB1* SNP rs9380142 [282] was found to associate with vitiligo. Finally, the simultaneous presence of the *DRB1*04* allele and the homozygous T to C mutation at position 809 of *DSG1* increased susceptibility to pemphigus foliaceus in a European sample [283].

Beyond interactions between *HLA* and non-MHC loci, many interactions have been reported *among HLA* genotypes for multiple diseases including AS [258], MS [270, 284–286], T1D [258, 264, 287, 288], adult glioma [289], coeliac disease [288] and RA [288].

7.2 IMPUTING *HLA* ALLELES FOR THE UKB

The vital role of *HLA* alleles in immune-mediated diseases, together with the existing evidence for epistatic effects involving *HLA* genes, motivated us to apply our new method for interaction testing to this region of the genome, which is not well covered by the main UKB imputed dataset used for the analyses discussed so far in this thesis.

However, this task – as well as *HLA* association testing more generally – has been complicated by challenges to the genotyping of variants within *HLA* genes: its extreme genetic diversity, which is also population specific, renders traditional genotyping methods slow and costly [246, 290]. These challenges have led to the development of several methods for statistical imputation of *HLA* alleles based on array sequencing which are inexpensive, fast and reasonably accurate [290–294].

For our analysis of interactions involving *HLA* alleles, we imputed genotypes in this region using the array data in the UKB together with a high-resolution multi-ancestry reference panel (version 2) [294]. This imputation was performed using Minimac4 on the Michigan Imputation Server (MIS) [295]. Phasing was performed as a first step with Eagle2 [296] and the default quality control procedures of the MIS pipeline were applied.

This imputation procedure yielded genotypes for 423 *HLA* alleles at two-field resolution: 251 for *HLA-A*, *HLA-B* and *HLA-C*, and the remaining 172 for *HLA-DRB1*, *HLA-DQA1*, *HLA-DQB1*, *HLA-DPA1* and *HLA-DPB1*. We then applied filters based on frequency of the alleles, which we required to be greater than or equal to 0.001, and imputation quality score, which we required to be at least 0.7. This filtering step resulted in a final set of 93 two-field resolution alleles, 44 for class I genes and the remaining 49 for class II.

From the set of 93 filtered alleles, we also identified long-range haplotypes [246] separately for genes in classes I and II – when carrying out association testing, we then treated alleles and haplotypes equally and processed them jointly. In the process of identifying haplotypes, we first removed any whose sample count was lower than ten *as well as the individuals carrying those haplotypes;* we then removed haplotypes whose frequency was (strictly) lower than 0.1% and also the individuals who carried two of the rare haplotypes (separately for classes I and II). This resulted in 151

haplotypes for 323,836 individuals for class I genes, and in 122 haplotypes and 331,258 individuals for class II.

7.3 RESULTS OF INTERACTION TESTING

We processed this set of 93 two-field alleles and 273 haplotypes through our pipeline for interaction testing, which we described extensively in previous chapters. We began by running standard direct association testing for these variants and then proceeded to test them for interactions against the PGSs we had built for our 97 traits. Finally, we identified independent LHS hits from the *HLA* allele/haplotype × PGS interaction results. We report the 25 independent interaction hits identified in the *LOCO* test results for *either Winsorised or QN versions* of the residuals in Table 7.1.

This work is ongoing and we have not yet been able to run pairwise interaction GWASs to attempt to identify pairs of interacting loci driving the LHS hits we found, but will do so in the future.

We begin by noting that we used a less stringent p-value threshold of 10^{-6} to identify the independent LHS hits in Table 7.1 because we are performing a much smaller number of tests compared to the genome-wide framework we employed in previous chapters. While for a test for associations of only 366 variants – if taken in isolation – we could have used a much higher threshold, this would have been problematic for a few different reasons. First, an even less stringent threshold would be likely to increase the rate of false positives. Second, interaction hits with high p-values have a lower 'signal-to-noise' ratio and it would therefore be less likely that we would be able to identify corresponding RHS hits for these hits. And finally, while we undertook this work separately from the main project described in this thesis, we are nevertheless merely extending the interaction tests we had already carried

out for the whole genome to a small additional region – seen in the context of this larger work, the interaction testing of these few *HLA* variants is effectively part of a genome-wide test and therefore should approximately satisfy the usual stringent threshold for significance for us to be confident that its results are unlikely to be false positives.

Now examining the 25 LHS hits listed in Table 7.1, we first draw attention to the fact that the general relationship between interaction and original GWAS signal across variants is similar to that which we observed for non-HLA SNPs before. In particular, while most variants (16/25) also showed a direct association (at the 10^{-6} significance threshold) with the trait in question, a sizeable proportion appear to affect that trait only via modulating the effects of other variants on different chromosomes.

Second, we observe that there are more hits with the Winsorised traits (16/25) than with their QN counterparts (9/25). We note that only one allele (*DQA1*03:01*) shows an interaction for both QN and Winsorised versions of a trait, glycated haemoglobin.

Third, we found no evidence of interactions involving haplotypes that could not be explained by alleles showing a stronger signal for an interaction. As mentioned before, we treated alleles and haplotypes equally when carrying out interaction testing. Therefore, when distilling the set of all significantly associated alleles/haplotypes for a particular trait, we applied stepwise regression considering both types of variants jointly. While a few haplotypes (13 across six traits) did show significant interaction signal, these either did not survive our stepwise filtering procedure (indicating that a single allele whose p-value was lower rendered the haplotype no longer significant

once included in the regression) or looked very likely to be false positives and so we decided to exclude them.[1]

We proceed by examining some of our interaction hits more closely and highlight the interaction involving *DQA1*03:01* for glycated haemoglobin, which as noted above is seen in both QN and Winsorised results. This interaction has a positive effect, while the *DQA1*03:01* allele does not show a direct association with the trait. This finding is of potential relevance for diabetes since *DQA1*03:01* has previously been linked to T1D risk [297], and glycated haemoglobin is routinely used as a diagnostic and clinical management tool for this disease [298].

Many of the remaining *HLA* alleles for which we found evidence of interaction with the PGS for one or more traits have also previously been found to associate with disease risk. Beginning with the class I genes, we draw attention to the following interactions involving:

- *A*01:01*, for fat-related traits: waist circumference, BMI and arm fat mass. This allele has been linked to T1D [299];
- *A*02:01*, for eosinophil count and percentage. This allele has been linked to MS [300], AS [301] and spontaneous clearance of hepatitis C virus (HCV) [302]; and
- *B*07:02*, for total bilirubin, which has also been connected to progression of HIV infection [273].

Now turning to alleles of the class II genes, we found interactions involving, among other alleles:

[1] Five haplotypes for eosinophil percentage and gamma glutamyltransferase did survive stepwise regression but were very rare (frequency of at most 0.5%), were all associated only with the Winsorised version of the traits, had abnormally large effect size estimates (but still moderate p-values, indicating large standard errors) and did not show direct association with the trait.

- *DPB1*03:01*, for aspartate aminotransferase, which has been associated with MS susceptibility [303];
- *DQA1*01:01*, also for aspartate aminotransferase, linked to MS [270];
- *DQB1*06:02*, for eosinophil count, which is associated with multiple autoimmune diseases including narcolepsy, MS, SLE, sarcoidosis, primary sclerosing cholangitis and T1D [304];
- *DRB1*03:01*, for eosinophil count and percentage, and aspartate aminotransferase. This allele is known to influence SLE susceptibility [305, 306];
- *DRB1*04:01*, for FEV1/FVC ratio, linked to treatment-resistant Lyme arthritis [307];
- *DRB1*07:01*, for alanine aminotransferase, previously found to confer higher risk of asparaginase allergies [308];
- *DRB1*14:01*, for BMI and leg fat mass, known to be linked to SEL [306];
- and, finally, *DRB1*15:01*, for eosinophil percentage, which has been reported to be the strongest genetic risk factor for MS [270, 271] as mentioned before and is also linked to Epstein–Barr virus infection [309], a well-known environmental risk factor for this disease. Interestingly, eosinophilic esophagitis (EoE), a chronic allergic condition in which eosinophils build up in the esophagus leading to inflamation, was recently found to correlate with excess risk for a series of autoimmune diseases, including MS, both in EoE patients and in their first-degree relatives [310].

7.4 Summary and future work

In this chapter, we described an in-progress extension of our work on identifying interactions to *HLA* genetic variants. After reviewing the structure and functionality of the MHC region and the *HLA* genes within it, we proceeded to consider previous

HLA associations with different diseases, paying special attention to existing evidence for epistasis. We then described how we imputed *HLA* alleles for our UKB dataset and finally reported the results of our *HLA* allele/haplotype × PGS interaction testing.

Our approach identified *25 independent* HLA alleles *interacting with the PGS for 15 different traits.* We then concluded by examining these different hits – of which 17 are unique – and listed their previously-known associations with various diseases.

Future work on this project will consist first of running pairwise interaction GWASs for each of the 25 independent LHS hits identified above. We will then attempt to detect significant pairwise interactions based on the results from these pairwise GWASs as we did for the rest of the genome in the previous chapters.

Finally, we will consider extending the analysis above to include disease phenotypes, with a particular emphasis on autoimmune diseases. Given the well-known relevance of *HLA* genetic variation for susceptibility to different diseases (especially those related to the immune system) and the several epistatic associations previously identified for this locus, this research direction may allow us to expand the set of known *HLA* interactions.

Table 7.1: Independent 'LHS' hits with beta coefficients and log-transformed p-values from *HLA* allele × PGS and original GWASs with QN or Winsorised residuals for all traits.

Trait	Residuals	Allele	Freq.	Int. beta	Int. p-value	Orig. beta	Orig. p-value
Waist circumference	QN	*A*01:01*	0.193	0.017	6.1	0.004	0.7
BMI	QN	*A*01:01*	0.193	0.014	6.4	0.011	3.7
BMI	Winsorised	*DRB1*14:01*	0.021	-0.042	6.1	0.021	2.3
Leg fat mass	Winsorised	*DRB1*14:01*	0.021	-0.049	8.9	0.006	0.5
Arm fat mass	QN	*A*01:01*	0.193	0.016	6.3	0.014	5.0
Red blood cell dist. width	QN	*B*08:01*	0.144	0.018	6.0	0.039	28.8
Neutrophil count	Winsorised	*C*05:01*	0.113	-0.020	6.8	0.050	46.7
Eosinophil count	Winsorised	*A*02:01*	0.280	-0.023	17.9	0.037	62.5
Eosinophil count	Winsorised	*C*03:03*	0.059	-0.025	6.3	0.021	6.3
Eosinophil count	QN	*DRB1*03:01*	0.147	-0.020	7.6	-0.063	72.5
Eosinophil count	Winsorised	*DQB1*06:02*	0.144	0.020	8.5	-0.029	24.6
Eosinophil %	Winsorised	*A*02:01*	0.280	-0.026	21.6	0.032	45.1
Eosinophil %	QN	*DRB1*03:01*	0.147	-0.019	6.9	-0.038	26.5
Eosinophil %	Winsorised	*DRB1*15:01*	0.145	0.017	6.1	-0.019	10.6
Alanine aminotransferase	Winsorised	*DRB1*07:01*	0.136	-0.020	7.7	0.009	3.1
Apoliprotein A	QN	*DQB1*02:01*	0.247	0.019	10.3	-0.011	3.3
Aspartate aminotransferase	Winsorised	*B*08:01*	0.144	0.038	27.1	-0.033	50.7
Aspartate aminotransferase	QN	*DRB1*03:01*	0.147	0.021	7.9	-0.078	108.0
Aspartate aminotransferase	Winsorised	*DQA1*01:01*	0.141	-0.019	6.7	0.022	23.4
Aspartate aminotransferase	Winsorised	*DPB1*03:01*	0.100	-0.026	9.4	0.020	13.9
Glycated haemoglobin	QN	*DQA1*03:01*	0.209	0.020	11.7	-0.005	0.9
Glycated haemoglobin	Winsorised	*DQA1*03:01*	0.209	0.017	8.7	0.002	0.5
Total bilirubin	Winsorised	*B*07:02*	0.148	-0.025	11.0	0.017	8.7
Triglycerides	Winsorised	*DQA1*05:01*	0.236	0.018	10.0	-0.024	22.0
FEV1/FVC ratio	Winsorised	*DRB1*04:01*	0.111	0.025	8.8	0.041	26.3

8 Conclusion

8.1 SUMMARY

In this thesis, we have introduced a new statistical approach to the testing of epistatic interactions between genetic variants affecting quantitative traits. This approach consists of first building a PGS exhausting all additive signal for a trait, and then testing for interactions between each SNP and this PGS in a genome-wide analysis. Genetic loci found to be significant in this test can also be examined further to identify pairwise interactions.

As discussed in Chapter 1, our method is related to existing approaches to detect epistasis: in particular, it has similarities with methods that aim to to prioritise loci for interaction testing in a two-step manner. However, it is distinct from these methods in that the central test of interaction between a SNP and the PGS through which SNPs can be prioritised for pairwise testing is scientifically meaningful in itself. Specifically, we conjecture that, give our understanding of regulatory mechanisms, it is plausible that a mutation could interact statistically with multiple other mutations whose effect operates downstream along a biological pathway.

Having described our method in Chapter 3 – along with a series of statistical challenges to identifying true interaction effects in this manner, and how we designed our method to overcome them – we proceeded in Chapter 4 to examine its perform-

ance in simulated data. We constructed a sizeable set of phenotypes based on a wide and realistic range of genetic architectures and tested our method's statistical power and propensity to yield false-positive findings. We found that type I-error could be effectively controlled by carefully dealing with outliers in the phenotypic distribution (through quantile-normalisation of the trait) and testing only for *trans* interactions (between variants in different chromosomes). Regarding statistical power, we concluded, as expected, that testing for SNP × PGS interactions is more effective for detecting interactions with a large impact, either in the number of SNPs affected or in their effect size.

We applied our method to the UKB (described in Chapter 2) and divided our exposition of the results of this analysis into two chapters. In Chapter 5, we focused on ALP and gave extended context on the biology of this phenotype, including the functional interaction between ABO blood group and secretor status that has long been known to affect serum ALP levels. Our interaction results identified an interaction between *ABO* variant rs635634 and *FUT2* variant rs681343 – the former is highly correlated with blood group, while the latter corresponds to the functional *FUT2* allele that determines secretor status. This interaction, which was previously known, therefore models the functional interdependence between these two regions. Beyond merely detecting this known interaction, we are able to extend it to a set of interacting loci involving *TREH* (also identified recently) as well as variants nearby *C1orf105*, *FUT6* and *MIS18A*, which are novel. In addition to these and a few other pairwise interactions, we also identify multiple other independent SNPs whose evidence for interaction with the PGS we are unable to distill into pairwise interactions – this is suggestive of a non-linear effect involving large groups of variants.

In Chapter 6, we described the results of applying our method to the remaining 96 traits analysed in this project. Examining these results in the aggregate allowed us

to draw the following general conclusions. First, we find that interaction signals are quite common: we observe 144 independent hits for interaction with the PGS across 52 different traits. Second, there is no consistent pattern of correlation between the results of our test for SNP × PGS interactions and those of a standard GWAS – this is significant as it is evidence that the non-linear relationships identified through this approach are not mere truncations of additive signal and are therefore more likely to be capturing meaningful biological relationships. Third, we saw that the 'independent LHS hits' for interaction with the PGS are enriched for functional relevance, again highlighting their potential functional importance. Fourth, our (very) approximate estimates of the heritability explained by the independent SNP × PGS interactions that we identified suggests that they account for nearly zero phenotypic variation, which is in agreement with previous literature. Fifth, we found that both LHS and pairwise hits were concentrated in molecular traits in the blood biochemistry and blood cell count categories – we comment on this finding below.

In the remainder of Chapter 6, we discussed some of the pairwise interactions identified for these traits. Among these, we highlight a network of interacting loci for mean platelet volume and platelet distribution width: rs1354034, a very common intronic variant of *ARHGEF3*, was found to interact with variants near *RNPEPL1*, *IQGAP2*, *WDR66* and *RUNX1* for platelet volume, and with variants near *RP5-1086L22.1*, *CTB-111H14.1*, *ODF3*, *WDR66* again and *BET1L* for platelet distribution width. All these 'RHS genes' had previously been linked to platelets or other blood cell traits, suggesting that this network of interactions is likely to be meaningful. Other significant findings include: two interactions, in opposite directions, between two key Alzheimer's disease mutations in *APOE* and the liver-associated proteins alanine aminotransferase and LPA, suggestive in line of recent evidence for a possible

link between AD and liver function; and an interaction between *IL33* and *ALOX15* for eosinophil count which may be related to the aetiology of asthma.

Finally, in Chapter 7, we described an on-going effort to apply our interaction-testing approach to genetic variation in *HLA* genes, a promising project given the many disease-associations found in this region and the substantial evidence that exists for epistatic effects. The work we have undertaken so far allowed us to identify 25 independent LHS hits associated with 15 different traits. A highlight from this preliminary analysis is an interaction between *DQA1*03:01* and the PGS for glycated haemoglobin, which could be relevant to the mechanisms underlying the development of diabetes given the known link between this allele and that disease and the importance of glycated haemoglobin as a diagnostic and clinical management tool.

8.2 LIMITATIONS

The project described in this thesis suffers from multiple limitations. First regarding our statistical approach, we are restricted to analysing quantitative (as opposed to case–control) traits because our strategies for normalising the phenotype (either through quantile-normalisation or through Winsorisation), as well as our iterative approach to building a PGS that accounts for all additive signal, do not directly apply to binary phenotypes.

A second limitation related to our interaction-testing strategy is that, while our PGS construction algorithm is well-motivated and theoretically sound, we did not consider alternative algorithms or perform a comparison with them in simulated data. It is therefore possible that a current state-of-the-art method could lead to greater power for the detection of interactions – our method's performance is not as good

as that of leading PGS construction algorithms[1] – while maintaining an appropriate type-I error rate.

Yet another limitation related to our method is that, as we postulated using simple mathematical derivations in Chapter 3 and later confirmed in simulated data as described in Chapter 4, we are not able to identify interactions that involve a small number of SNPs with small effect sizes. This means that, although we identify a large number of interactions for a sizeable fraction of our traits, a more powerful method could potentially detect many more.

A fourth shortcoming is that we cannot reliably test for *cis* interactions – i.e., interactions between variants in close physical proximity – as our approach is liable to yield false-positives in such settings. However, given the importance of cis-regulation of gene expression, it is conceivable that the majority of interactions would involve neighbouring variants. The detection of *cis* interactions in a way that is robust to statistical artefacts caused by potential unmeasured variants in LD with those being tested remains an important open problem in the field.

Turning now to our empirical analysis of the UKB, an important limitation is that we did not attempt to replicate our findings in other biobank datasets. This would be valuable in the future by giving us greater confidence in the reliability of the interactions that could be seen in multiple cohorts; a challenge is that it requires access to raw genetic data for a large number of individuals of closely-related genetic backgrounds to the White British sample from the UKB that we used in this analysis.

A sixth limitation is that our analysis and interpretation of possible functional mechanisms underlying the identified interactions is limited – we hope to improve on this in future work.

[1] For height, for example, our predictive performance on a held-out test set is approximately 86% of that achieved by BOLT-LMM.

8.3 DISCUSSION

We now proceed to consider the broader implications of our work, and to compare our findings with those of a recent related method to detect marginal epistasis that we introducted in Chapter 1. First, our analysis concurs with previous literature in showing that epistatic effects generally explain very little heritability. This well-established finding suggests that the focus of attention for *improving the predictive performance of PGSs* should be on further refining and expanding our knowledge of direct associations through larger samples of more diverse ethnicities, and better annotations of function.

Second, while the interactions we identified do not contribute significantly to phenotypic variation and consequently to heritability, a first examination of their functional implications suggests that much can be learned about key biological processes, such as gene regulation – and, more generally, the pathways underlying different traits and diseases – by modelling non-linear genetic effects. A key open problem in statistical genetics remains the large gulf that exists between the results of statistical association testing and the functional basis of phenotypes – greater attention to epistatic interactions could contribute towards bridging this gap.

Third, we noted above that the great majority of the interactions detected through our method affect only molecular traits, whose genetic architecture is simpler and which are generally more heritable than more polygenic traits. However, complex traits like height or BMI are in effect an aggregate result of multiple simpler biological traits and functions. Therefore, it appears likely to us that highly-polygenic traits – for which we have found only a few interactions – nevertheless contain several epistatic interactions in their architecture, and in fact more so than simpler biomarkers. This

would suggest that a lack of statistical power stands in the way of the identification of many more interactions for human complex traits.

Fourth, we argued when motivating our approach that it is biologically meaningful to test for interactions between single variants and groups of other variants that have a known direct effect on the trait: such interactions could represent the functioning of homeostatic mechanisms or the final breakdown of a genetically-robust system whose functioning is hampered in disease. That we identified nearly 150 such interactions for a variety of traits, and that that we could not find pairwise epistatic signals driving many of these, suggests that this type of interaction is meaningful and deserving of further study.

Turning now to a comparison with the most closely related existing approach to interaction testing in the literature, the test for marginal epistasis of Fu et al. [98], and its application to the UKB, we first note that our analysis covers a larger number of phenotypes (97 vs. 53). However, even taking this difference into account, our method is able to identify many more interactions between independent LHS SNPs and a PGS (these correspond much more closely to marginal epistatic effects than pairwise interactions): we identify 144 hits across 52 different traits, while Fu et al. detect only 16 trait–loci pairs. On the surface, the approach proposed in this thesis therefore appears to be more powerful. Moreover, and as we noted in Chapter 1, interactions with a PGS are more specific in scope and therefore easier to examine in downstream analyses – however, a more parsimonious model may be able to identify different kinds of effects not covered by our framework, such as pure variance effects.

Comparing the 16 trait–loci pairs identified by Fu et al. to our independent LHS associations, we note that we are able to identify at least seven of these, as we find either the same variant or a neighbouring SNP in the same region and for the same trait. This partial overlap between the results of the two methods gives strength to

both approaches: first, by making the overlapping interactions more credible, thus validating these two alternative ways of identifying epistatic effects; and second, by emphasising that the two methods can capture different effects, as might be expected given their different designs.

8.4 FUTURE WORK

There are multiple avenues for extending the work we have done so far that may prove fruitful, and we hope to explore some of them. Beginning with the main project presented in this thesis, future work could focus first on enlarging the list of interactions found by carefully examining hits for Winsorised traits or where the interaction was with the full PGS – since such interactions are more likely to be false positives, this would require careful manual examination. Second, the functional interpretation of the interactions detected could be improved. Third, we could test the robustness of our hits by verifying whether they can be replicated in external datasets, such as FinnGen [311] or the American biobank All of Us [312].

A fourth promising direction for future research is to consider interactions with sub-components of the PGS determined from biological knowledge, and we believe this might help overcome some of the shortcomings of our current approach. For example, one could identify all the variants in a PGS that are located on or near the binding sites of a particular transcription factor through an analysis of binding motifs, and then test for an interaction of mutations affecting this transcription factor with a 'sub-PGS' comprised of only those variants. In this way, we may be able to identify promising interactions whose functional interpretation would be powerful and clear.

Now considering the smaller project on *HLA* variation that we described in Chapter 7, future work could involve testing for pairwise interactions for the inde-

pendent LHS hits that we have identified, investigating the functional meaning of the SNP × PGS and pairwise interactions found – perhaps with special attention paid to the interaction of *DQA1*03:01* for glycated haemoglobin – and, finally, extending the analysis to a set of diseases for which a reasonable number of cases can be found in the UKB – given the importance of *HLA* variation for disease biology, particularly of autoimmune diseases, this may lead to relevant and clinically-useful findings.

References

1. Visscher, P. M. *et al.* 10 Years of GWAS Discovery: Biology, Function, and Translation. *The American Journal of Human Genetics* **101,** 5–22 (2017).
2. Sollis, E. *et al.* The NHGRI-EBI GWAS Catalog: Knowledgebase and Deposition Resource. *Nucleic Acids Research* **51,** D977–D985 (2023).
3. Watanabe, K. *et al.* A Global Overview of Pleiotropy and Genetic Architecture in Complex Traits. *Nature Genetics* **51,** 1339–1348 (9 2019).
4. Klein, R. J. *et al.* Complement Factor H Polymorphism in Age-Related Macular Degeneration. *Science* **308,** 385–389 (2005).
5. Tam, V. *et al.* Benefits and Limitations of Genome-Wide Association Studies. *Nature Reviews Genetics* **20,** 467–484 (8 2019).
6. Weissbrod, O. *et al.* Functionally Informed Fine-Mapping and Polygenic Localization of Complex Trait Heritability. *Nature Genetics* **52,** 1355–1363 (12 2020).
7. Schipper, M. & Posthuma, D. Demystifying Non-Coding GWAS Variants: An Overview of Computational Tools and Methods. *Human Molecular Genetics* **31,** R73–R83 (2022).
8. Sivakumaran, S. *et al.* Abundant Pleiotropy in Human Complex Diseases and Traits. *The American Journal of Human Genetics* **89,** 607–618 (2011).
9. Cano-Gamez, E. & Trynka, G. From GWAS to Function: Using Functional Genomics to Identify the Mechanisms Underlying Complex Diseases. *Frontiers in Genetics* **11,** 424 (2020).
10. Hampe, J. *et al.* A Genome-Wide Association Scan of Nonsynonymous SNPs Identifies a Susceptibility Variant for Crohn Disease in ATG16L1. *Nature Genetics* **39,** 207–211 (2 2007).
11. Moschen, A. R., Tilg, H. & Raine, T. IL-12, IL-23 and IL-17 in IBD: Immunobiology and Therapeutic Targeting. *Nature Reviews Gastroenterology & Hepatology* **16,** 185–196 (3 2019).
12. Frayling, T. M. *et al.* A Common Variant in the FTO Gene Is Associated with Body Mass Index and Predisposes to Childhood and Adult Obesity. *Science* **316,** 889–894 (2007).

13. Scuteri, A. *et al.* Genome-Wide Association Scan Shows Genetic Variants in the FTO Gene Are Associated with Obesity-Related Traits. *PLOS Genetics* **3,** e115 (2007).

14. Loos, R. J. F. & Yeo, G. S. H. The Bigger Picture of FTO—the First GWAS-identified Obesity Gene. *Nature Reviews Endocrinology* **10,** 51–61 (2014).

15. Zhao, X., Yang, Y., Sun, B.-F., Zhao, Y.-L. & Yang, Y.-G. FTO and Obesity: Mechanisms of Association. *Current Diabetes Reports* **14,** 486 (2014).

16. Lan, N. *et al.* FTO – A Common Genetic Basis for Obesity and Cancer. *Frontiers in Genetics* **11** (2020).

17. Rusu, V. *et al.* Type 2 Diabetes Variants Disrupt Function of SLC16A11 through Two Distinct Mechanisms. *Cell* **170,** 199–212.e20 (2017).

18. Choi, S. W., Mak, T. S.-H. & O'Reilly, P. F. Tutorial: A Guide to Performing Polygenic Risk Score Analyses. *Nature Protocols* **15,** 2759–2772 (2020).

19. Torkamani, A., Wineinger, N. E. & Topol, E. J. The Personal and Clinical Utility of Polygenic Risk Scores. *Nature Reviews Genetics* **19,** 581–590 (2018).

20. Khera, A. V. *et al.* Genome-Wide Polygenic Scores for Common Diseases Identify Individuals with Risk Equivalent to Monogenic Mutations. *Nature Genetics* **50,** 1219–1224 (2018).

21. Uffelmann, E. *et al.* Genome-Wide Association Studies. *Nature Reviews Methods Primers* **1,** 1–21 (1 2021).

22. Ni, G. *et al.* A Comparison of Ten Polygenic Score Methods for Psychiatric Disorders Applied Across Multiple Cohorts. *Biological Psychiatry* **90,** 611–620 (2021).

23. Zhang, Q., Privé, F., Vilhjálmsson, B. & Speed, D. Improved Genetic Prediction of Complex Traits from Individual-Level Data or Summary Statistics. *Nature Communications* **12,** 4192 (2021).

24. Mavaddat, N. *et al.* Polygenic Risk Scores for Prediction of Breast Cancer and Breast Cancer Subtypes. *The American Journal of Human Genetics* **104,** 21–34 (2019).

25. Schumacher, F. R. *et al.* Association Analyses of More than 140,000 Men Identify 63 New Prostate Cancer Susceptibility Loci. *Nature Genetics* **50,** 928–936 (2018).

26. Khera, A. V. *et al.* Genetic Risk, Adherence to a Healthy Lifestyle, and Coronary Disease. *New England Journal of Medicine* **375,** 2349–2358 (2016).

27. Inouye, M. *et al.* Genomic Risk Prediction of Coronary Artery Disease in 480,000 Adults. *Journal of the American College of Cardiology* **72,** 1883–1893 (2018).

28. Bonifacio, E. *et al.* Genetic Scores to Stratify Risk of Developing Multiple Islet Autoantibodies and Type 1 Diabetes: A Prospective Study in Children. *PLOS Medicine* **15,** e1002548 (2018).

29. Sharp, S. A. *et al.* Development and Standardization of an Improved Type 1 Diabetes Genetic Risk Score for Use in Newborn Screening and Incident Diagnosis. *Diabetes Care* **42,** 200–207 (2019).

30. Desikan, R. S. *et al.* Genetic Assessment of Age-Associated Alzheimer Disease Risk: Development and Validation of a Polygenic Hazard Score. *PLOS Medicine* **14,** e1002258 (2017).

31. Oram, R. A. *et al.* A Type 1 Diabetes Genetic Risk Score Can Aid Discrimination Between Type 1 and Type 2 Diabetes in Young Adults. *Diabetes Care* **39,** 337–344 (2016).

32. Patel, K. A. *et al.* Type 1 Diabetes Genetic Risk Score: A Novel Tool to Discriminate Monogenic and Type 1 Diabetes. *Diabetes* **65,** 2094–2099 (2016).

33. Daunt, P. *et al.* Polygenic Risk Scoring Is an Effective Approach to Predict Those Individuals Most Likely to Decline Cognitively Due to Alzheimer's Disease. *The Journal of Prevention of Alzheimer's Disease,* 1–6 (2020).

34. Polygenic Risk Score Task Force of the International Common Disease Alliance. Responsible Use of Polygenic Risk Scores in the Clinic: Potential Benefits, Risks and Gaps. *Nature Medicine* **27,** 1876–1884 (2021).

35. Lee, A. *et al.* BOADICEA: A Comprehensive Breast Cancer Risk Prediction Model Incorporating Genetic and Nongenetic Risk Factors. *Genetics in Medicine* **21,** 1708–1718 (2019).

36. Carver, T. *et al.* CanRisk Tool—A Web Interface for the Prediction of Breast and Ovarian Cancer Risk and the Likelihood of Carrying Genetic Pathogenic Variants. *Cancer Epidemiology, Biomarkers & Prevention* **30,** 469–473 (2021).

37. Cecile, A., Janssens, J. W. & Joyner, M. J. Polygenic Risk Scores That Predict Common Diseases Using Millions of Single Nucleotide Polymorphisms: Is More, Better? *Clinical Chemistry* **65,** 609–611 (2019).

38. Curtis, D. Clinical Relevance of Genome-Wide Polygenic Score May Be Less than Claimed. *Annals of Human Genetics* **83,** 274–277 (2019).

39. Abu-El-Haija, A. *et al.* The Clinical Application of Polygenic Risk Scores: A Points to Consider Statement of the American College of Medical Genetics and Genomics (ACMG). *Genetics in Medicine* **25,** 100803 (2023).

40. Polderman, T. J. C. *et al.* Meta-Analysis of the Heritability of Human Traits Based on Fifty Years of Twin Studies. *Nature Genetics* **47,** 702–709 (2015).

41. Need, A. C. & Goldstein, D. B. Next Generation Disparities in Human Genomics: Concerns and Remedies. *Trends in Genetics* **25,** 489–494 (2009).

42. Bustamante, C. D., De La Vega, F. M. & Burchard, E. G. Genomics for the World. *Nature* **475,** 163–165 (7355 2011).

43. Popejoy, A. B. & Fullerton, S. M. Genomics Is Failing on Diversity. *Nature* **538,** 161–164 (2016).

44. Martin, A. R. *et al.* Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations. *The American Journal of Human Genetics* **100,** 635–649 (2017).

45. Mills, M. C. & Rahal, C. The GWAS Diversity Monitor Tracks Diversity by Disease in Real Time. *Nature Genetics* **52,** 242–243 (2020).

46. Petrovski, S. & Goldstein, D. B. Unequal Representation of Genetic Variation across Ancestry Groups Creates Healthcare Inequality in the Application of Precision Medicine. *Genome Biology* **17,** 157 (2016).

47. Martin, A. R. *et al.* Clinical Use of Current Polygenic Risk Scores May Exacerbate Health Disparities. *Nature Genetics* **51,** 584–591 (2019).

48. Duncan, L. *et al.* Analysis of Polygenic Risk Score Usage and Performance in Diverse Human Populations. *Nature Communications* **10,** 3328 (2019).

49. Kachuri, L. *et al.* Principles and Methods for Transferring Polygenic Risk Scores across Global Populations. *Nature Reviews Genetics,* 1–18 (2023).

50. Phillips, P. C. The Language of Gene Interaction. *Genetics* **149,** 1167–1171 (1998).

51. Bateson, W. *Mendel's Principles of Heredity* (Cambridge University Press, Cambridge, 1909).

52. Cuénot, L. La Loi de Mendel et l'Hérédité de La Pigmentation Chez Les Souris. *Archives de zoologie expérimentale et générale* **10,** 27 (1902).

53. Phillips, P. C. Epistasis — the Essential Role of Gene Interactions in the Structure and Evolution of Genetic Systems. *Nature Reviews Genetics* **9,** 855–867 (2008).

54. Miko, I. Epistasis: Gene Interaction and Phenotype Effects. *Nature Education* **1,** 197 (2008).

55. Fisher, R. A. The Correlation between Relatives on the Supposition of Mendelian Inheritance. *Transactions of the Royal Society of Edinburgh* **52,** 399–433 (1918).

56. Mackay, T. F. C. Epistasis and Quantitative Traits: Using Model Organisms to Study Gene–Gene Interactions. *Nature Reviews Genetics* **15,** 22–33 (2014).

57. Manolio, T. A. *et al.* Finding the Missing Heritability of Complex Diseases. *Nature* **461,** 747–753 (7265 2009).

58. Sadee, W. *et al.* Missing Heritability of Common Diseases and Treatments Outside the Protein-Coding Exome. *Human Genetics* **133,** 1199–1215 (2014).

59. Wang, H., Bennett, D. A., De Jager, P. L., Zhang, Q.-Y. & Zhang, H.-Y. Genome-Wide Epistasis Analysis for Alzheimer's Disease and Implications for Genetic Risk Prediction. *Alzheimer's Research & Therapy* **13,** 55 (2021).

60. Hill, W. G., Goddard, M. E. & Visscher, P. M. Data and Theory Point to Mainly Additive Genetic Variance for Complex Traits. *PLOS Genetics* **4,** e1000008 (2008).

61. Sackton, T. B. & Hartl, D. L. Genotypic Context and Epistasis in Individuals and Populations. *Cell* **166,** 279–287 (2016).

62. Hivert, V. *et al.* Estimation of Non-Additive Genetic Variance in Human Complex Traits from a Large Sample of Unrelated Individuals. *The American Journal of Human Genetics* **108,** 786–798 (2021).

63. Huang, W. *et al.* Epistasis Dominates the Genetic Architecture of *Drosophila* Quantitative Traits. *Proceedings of the National Academy of Sciences* **109,** 15553–15559 (2012).

64. Bloom, J. S., Ehrenreich, I. M., Loo, W. T., Lite, T.-L. V. & Kruglyak, L. Finding the Sources of Missing Heritability in a Yeast Cross. *Nature* **494,** 234–237 (7436 2013).

65. Costanzo, M. *et al.* A Global Genetic Interaction Network Maps a Wiring Diagram of Cellular Function. *Science* **353,** aaf1420 (2016).

66. Hou, K. *et al.* Causal Effects on Complex Traits Are Similar for Common Variants across Segments of Different Continental Ancestries within Admixed Individuals. *Nature Genetics* **55,** 549–558 (4 2023).

67. Hu, S. *et al. Leveraging Fine-Scale Population Structure Reveals Conservation in Genetic Effect Sizes between Human Populations across a Range of Human Phenotypes* `https://www.biorxiv.org/content/10.1101/2023.08.08.552281v1` (2023). preprint.

68. Domingo, J., Baeza-Centurion, P. & Lehner, B. The Causes and Consequences of Genetic Interactions (Epistasis). *Annual Review of Genomics and Human Genetics* **20,** 433–460 (2019).

69. Presgraves, D. C. Speciation Genetics: Epistasis, Conflict and the Origin of Species. *Current Biology* **17,** R125–R127 (2007).

70. De Visser, J. A. G. M., Cooper, T. F. & Elena, S. F. The Causes of Epistasis. *Proceedings of the Royal Society B: Biological Sciences* **278,** 3617–3624 (2011).

71. Cordell, H. J. Detecting Gene–Gene Interactions That Underlie Human Diseases. *Nature Reviews Genetics* **10,** 392–404 (6 2009).

72. Wei, W.-H., Hemani, G. & Haley, C. S. Detecting Epistasis in Human Complex Traits. *Nature Reviews Genetics* **15,** 722–733 (11 2014).

73. Niel, C., Sinoquet, C., Dina, C. & Rocheleau, G. A Survey about Methods Dedicated to Epistasis Detection. *Frontiers in Genetics* **6** (2015).

74. Upton, A., Trelles, O., Cornejo-García, J. A. & Perkins, J. R. Review: High-performance Computing to Detect Epistasis in Genome Scale Data Sets. *Briefings in Bioinformatics* **17,** 368–379 (2016).

75. Schüpbach, T., Xenarios, I., Bergmann, S. & Kapur, K. FastEpistasis: A High Performance Computing Solution for Quantitative Trait Epistasis. *Bioinformatics* **26,** 1468–1469 (2010).

76. Hu, X. *et al.* SHEsisEpi, a GPU-enhanced Genome-Wide SNP-SNP Interaction Scanning Algorithm, Efficiently Reveals the Risk Genetic Epistasis in Bipolar Disorder. *Cell Research* **20,** 854–857 (7 2010).

77. Hemani, G., Theocharidis, A., Wei, W. & Haley, C. EpiGPU: Exhaustive Pairwise Epistasis Scans Parallelized on Consumer Level Graphics Cards. *Bioinformatics* **27,** 1462–1465 (2011).

78. Kam-Thong, T. *et al.* GLIDE: GPU-Based Linear Regression for Detection of Epistasis. *Human Heredity* **73,** 220–236 (2012).

79. González-Domínguez, J. *et al.* Parallelizing Epistasis Detection in GWAS on FPGA and GPU-Accelerated Computing Systems. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **12,** 982–994 (2015).

80. Wan, X. *et al.* BOOST: A Fast Approach to Detecting Gene-Gene Interactions in Genome-wide Case-Control Studies. *The American Journal of Human Genetics* **87,** 325–340 (2010).

81. Gyenesei, A., Moody, J., Semple, C. A., Haley, C. S. & Wei, W.-H. High-Throughput Analysis of Epistasis in Genome-Wide Association Studies with BiForce. *Bioinformatics* **28,** 1957–1964 (2012).

82. Kam-Thong, T., Pütz, B., Karbalai, N., Müller–Myhsok, B. & Borgwardt, K. Epistasis Detection on Quantitative Phenotypes by Exhaustive Enumeration Using GPUs. *Bioinformatics* **27,** i214–i221 (2011).

83. Zhao, J., Jin, L. & Xiong, M. Test for Interaction between Two Unlinked Loci. *The American Journal of Human Genetics* **79,** 831–845 (2006).

84. Prabhu, S. & Pe'er, I. Ultrafast Genome-Wide Scan for SNP–SNP Interactions in Common Complex Disease. *Genome Research* **22,** 2230–2240 (2012).

85. Ueki, M. & Cordell, H. J. Improved Statistics for Genome-Wide Interaction Analysis. *PLOS Genetics* **8,** e1002625 (2012).

86. Sun, X. *et al.* Analysis Pipeline for the Epistasis Search – Statistical versus Biological Filtering. *Frontiers in Genetics* **5** (2014).

87. Genetic Analysis of Psoriasis Consortium & the Wellcome Trust Case Control Consortium 2. A Genome-Wide Association Study Identifies New Psoriasis Susceptibility Loci and an Interaction between HLA-C and ERAP1. *Nature Genetics* **42,** 985–990 (2010).

88. Marchini, J., Donnelly, P. & Cardon, L. R. Genome-Wide Strategies for Detecting Multiple Loci That Influence Complex Diseases. *Nature Genetics* **37,** 413–417 (4 2005).

89. Evans, D. M., Marchini, J., Morris, A. P. & Cardon, L. R. Two-Stage Two-Locus Models in Genome-Wide Association. *PLOS Genetics* **2,** e157 (2006).

90. Ueki, M. & Tamiya, G. Ultrahigh-Dimensional Variable Selection Method for Whole-Genome Gene-Gene Interaction Analysis. *BMC Bioinformatics* **13,** 72 (2012).

91. Paré, G., Cook, N. R., Ridker, P. M. & Chasman, D. I. On the Use of Variance per Genotype as a Tool to Identify Quantitative Trait Interaction Effects: A Report from the Women's Genome Health Study. *PLOS Genetics* **6,** e1000981 (2010).

92. Forsberg, S. K. G. & Carlborg, Ö. On the Relationship between Epistasis and Genetic Variance Heterogeneity. *Journal of Experimental Botany* **68,** 5431–5438 (2017).

93. Struchalin, M. V., Dehghan, A., Witteman, J. C., van Duijn, C. & Aulchenko, Y. S. Variance Heterogeneity Analysis for Detection of Potentially Interacting Genetic Loci: Method and Its Limitations. *BMC Genetics* **11,** 92 (2010).

94. Young, A. I., Wauthier, F. L. & Donnelly, P. Identifying Loci Affecting Trait Variability and Detecting Interactions in Genome-Wide Association Studies. *Nature Genetics* **50,** 1608–1614 (2018).

95. Snaebjarnarson, A. S. *et al.* Complex Effects of Sequence Variants on Lipid Levels and Coronary Artery Disease. *Cell* **186,** 4085–4099.e15 (2023).

96. Crawford, L., Zeng, P., Mukherjee, S. & Zhou, X. Detecting Epistasis with the Marginal Epistasis Test in Genetic Mapping Studies of Quantitative Traits. *PLOS Genetics* **13,** e1006869 (2017).

97. Crawford, L. & Zhou, X. *Genome-Wide Marginal Epistatic Association Mapping in Case-Control Studies* `https://www.biorxiv.org/content/10.1101/374983v1` (2022). preprint.

98. Fu, B. *et al. A Biobank-Scale Test of Marginal Epistasis Reveals Genome-Wide Signals of Polygenic Epistasis* `https://www.biorxiv.org/content/10.1101/2023.09.10.557084v1` (2023). preprint.

99. Ma, L., Clark, A. G. & Keinan, A. Gene-Based Testing of Interactions in Association Studies of Quantitative Traits. *PLOS Genetics* **9,** e1003321 (2013).

100. Wu, M. C. *et al.* Powerful SNP-Set Analysis for Case-Control Genome-wide Association Studies. *The American Journal of Human Genetics* **86,** 929–942 (2010).

101. Ritchie, M. D. *et al.* Multifactor-Dimensionality Reduction Reveals High-Order Interactions among Estrogen-Metabolism Genes in Sporadic Breast Cancer. *The American Journal of Human Genetics* **69,** 138–147 (2001).

102. Sinnott-Armstrong, N. A., Greene, C. S., Cancare, F. & Moore, J. H. Accelerating Epistasis Analysis in Human Genetics with Consumer Graphics Hardware. *BMC Research Notes* **2,** 149 (2009).

103. Greene, C. S. *et al.* Multifactor Dimensionality Reduction for Graphics Processing Units Enables Genome-Wide Testing of Epistasis in Sporadic ALS. *Bioinformatics* **26,** 694–695 (2010).

104. Culverhouse, R. C. A Comparison of Methods Sensitive to Interactions with Small Main Effects. *Genetic Epidemiology* **36,** 303–311 (2012).

105. Schwarz, D. F., König, I. R. & Ziegler, A. On Safari to Random Jungle: A Fast Implementation of Random Forests for High-Dimensional Data. *Bioinformatics* **26,** 1752–1758 (2010).

106. De Lobel, L. *et al.* A Screening Methodology Based on Random Forests to Improve the Detection of Gene–Gene Interactions. *European Journal of Human Genetics* **18,** 1127–1132 (10 2010).

107. Molinaro, A. M. *et al.* Power of Data Mining Methods to Detect Genetic Associations and Interactions. *Human Heredity* **72,** 85–97 (2011).

108. Yoshida, M. & Koike, A. SNPInterForest: A New Method for Detecting Epistatic Interactions. *BMC Bioinformatics* **12,** 469 (2011).

109. Li, J., Horstman, B. & Chen, Y. Detecting Epistatic Effects in Association Studies at a Genomic Level Based on an Ensemble Approach. *Bioinformatics* **27,** i222–i229 (2011).

110. Crawford, L., Flaxman, S. R., Runcie, D. E. & West, M. Variable Prioritization in Nonlinear Black Box Methods: A Genetic Association Case Study. *The Annals of Applied Statistics* **13,** 958–989 (2019).

111. Fu, B. *et al.* Fast Kernel-Based Association Testing of Non-Linear Genetic Effects for Biobank-Scale Data. *Nature Communications* **14,** 4936 (1 2023).

112. Chanda, P. *et al.* Information-Theoretic Metrics for Visualizing Gene-Environment Interactions. *The American Journal of Human Genetics* **81,** 939–963 (2007).

113. Dong, C. *et al.* Exploration of Gene–Gene Interaction Effects Using Entropy-Based Methods. *European Journal of Human Genetics* **16,** 229–235 (2 2008).

114. Kang, G. *et al.* An Entropy-Based Approach for Testing Genetic Epistasis Underlying Complex Diseases. *Journal of Theoretical Biology* **250,** 362–374 (2008).

115. Hu, T. *et al.* An Information-Gain Approach to Detecting Three-Way Epistatic Interactions in Genetic Association Studies. *Journal of the American Medical Informatics Association* **20,** 630–636 (2013).

116. Knights, J., Yang, J., Chanda, P., Zhang, A. & Ramanathan, M. SYMPHONY, an Information-Theoretic Method for Gene–Gene and Gene–Environment Interaction Analysis of Disease Syndromes. *Heredity* **110,** 548–559 (6 2013).

117. Ignac, T. M., Skupin, A., Sakhanenko, N. A. & Galas, D. J. Discovering Pair-Wise Genetic Interactions: An Information Theory-Based Approach. *PLOS ONE* **9,** e92310 (2014).

118. Zwick, M. Reconstructability Analysis of Epistasis. *Annals of Human Genetics* **75,** 157–171 (2011).

119. Wang, Y., Liu, X., Robbins, K. & Rekaya, R. AntEpiSeeker: Detecting Epistatic Interactions for Case-Control Studies Using a Two-Stage Ant Colony Optimization Algorithm. *BMC Research Notes* **3,** 117 (2010).

120. Shang, J. *et al.* A Review of Ant Colony Optimization Based Methods for Detecting Epistatic Interactions. *IEEE Access* **7,** 13497–13509 (2019).

121. Falconer, D. S. The Inheritance of Liability to Certain Diseases, Estimated from the Incidence among Relatives. *Annals of Human Genetics* **29,** 51–76 (1965).

122. Bycroft, C. *et al.* The UK Biobank Resource with Deep Phenotyping and Genomic Data. *Nature* **562,** 203–209 (2018).

123. Bulik-Sullivan, B. K. *et al.* LD Score Regression Distinguishes Confounding from Polygenicity in Genome-Wide Association Studies. *Nature Genetics* **47,** 291–295 (2015).

124. Neale Lab. *Heritability of >4,000 Traits & Disorders in UK Biobank* `https://nealelab.github.io/UKBB_ldsc/` (2019).

125. Manichaikul, A. *et al.* Robust Relationship Inference in Genome-Wide Association Studies. *Bioinformatics* **26,** 2867–2873 (2010).

126. Anderson, C. A. *et al.* Data Quality Control in Genetic Case-Control Association Studies. *Nature Protocols* **5,** 1564–1573 (2010).

127. Marees, A. T. *et al.* A Tutorial on Conducting Genome-Wide Association Studies: Quality Control and Statistical Analysis. *International Journal of Methods in Psychiatric Research* **27,** e1608 (2018).

128. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biology* **17,** 122 (2016).

129. Hunter, D. J. Gene–Environment Interactions in Human Diseases. *Nature Reviews Genetics* **6,** 287–298 (4 2005).

130. Halldorsdottir, T. & Binder, E. B. Gene × Environment Interactions: From Molecular Mechanisms to Behavior. *Annual Review of Psychology* **68,** 215–241 (2017).

131. Virolainen, S. J., VonHandorf, A., Viel, K. C. M. F., Weirauch, M. T. & Kottyan, L. C. Gene–Environment Interactions and Their Impact on Human Health. *Genes & Immunity* **24,** 1–11 (1 2023).

132. Hayashi, F. *Econometrics* (Princeton University Press, Princeton, N.J., 2000).

133. Wasserman, L. *All of Statistics: A Concise Course in Statistical Inference* (Springer, New York, N.Y., 2004).

134. Abbott, M. G. *OLS Estimation of the Multiple (Three-Variable) Linear Regression Model* Lecture Notes. 2009.

135. Snedecor, G. W. & Cochran, W. G. *Statistical Methods* 8th ed. (Iowa State University Press, Ames, Iowa, 1989).

136. Fuller, W. A. *Measurement Error Models* (John Wiley, Chichester, N.Y., 1987).

137. Greene, W. H. *Econometric Analysis* 8th, global ed. (Pearson, Harlow, England, 2020).

138. Zschocke, J., Byers, P. H. & Wilkie, A. O. M. Gregor Mendel and the Concepts of Dominance and Recessiveness. *Nature Reviews Genetics* **23,** 387–388 (7 2022).

139. Visscher, P. M. & Goddard, M. E. From R.A. Fisher's 1918 Paper to GWAS a Century Later. *Genetics* **211,** 1125–1130 (2019).

140. Palmer, D. S. *et al.* Analysis of Genetic Dominance in the UK Biobank. *Science* **379,** 1341–1348 (2023).

141. Otwinowski, J., McCandlish, D. M. & Plotkin, J. B. Inferring the Shape of Global Epistasis. *Proceedings of the National Academy of Sciences* **115,** E7550–E7558 (2018).

142. Tukey, J. W. & McLaughlin, D. H. Less Vulnerable Confidence and Significance Procedures for Location Based on a Single Sample: Trimming/Winsorization 1. *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002)* **25,** 331–352 (1963).

143. Wilcox, R. R. *Fundamentals of Modern Statistical Methods: Substantially Improving Power and Accuracy* (Springer, New York, N.Y., 2010).

144. Beasley, T. M., Erickson, S. & Allison, D. B. Rank-Based Inverse Normal Transformations Are Increasingly Used, But Are They Merited? *Behavior Genetics* **39,** 580–595 (2009).

145. McCaw, Z. R., Lane, J. M., Saxena, R., Redline, S. & Lin, X. Operating Characteristics of the Rank-Based Inverse Normal Transformation for Quantitative Trait Analysis in Genome-Wide Association Studies. *Biometrics* **76,** 1262–1272 (2020).

146. Peterson, R. A. & Cavanaugh, J. E. Ordered Quantile Normalization: A Semiparametric Transformation Built for the Cross-Validation Era. *Journal of Applied Statistics* **47,** 2312–2327 (2020).

147. Peterson, R. A. Finding Optimal Normalizing Transformations via bestNormalize. *The R Journal* **13,** 310–329 (2021).

148. Wood, A. R. *et al.* Another Explanation for Apparent Epistasis. *Nature* **514,** E3–E5 (7520 2014).

149. De los Campos, G., Sorensen, D. A. & Toro, M. A. Imperfect Linkage Disequilibrium Generates Phantom Epistasis (& Perils of Big Data). *G3: Genes|Genomes|Genetics* **9,** 1429–1436 (2019).

150. Hemani, G. *et al.* Phantom Epistasis between Unlinked Loci. *Nature* **596,** E1–E3 (7871 2021).

151. Privé, F., Vilhjálmsson, B. J., Aschard, H. & Blum, M. G. B. Making the Most of Clumping and Thresholding for Polygenic Scores. *The American Journal of Human Genetics* **105,** 1213–1221 (2019).

152. Schapire, R. E. The Strength of Weak Learnability. *Machine Learning* **5,** 197–227 (1990).

153. Freund, Y. Boosting a Weak Learning Algorithm by Majority. *Information and Computation* **121,** 256–285 (1995).

154. Friedman, J. H. Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics* **29,** 1189–1232 (2001).

155. Freund, R. M., Grigas, P. & Mazumder, R. A New Perspective on Boosting in Linear Regression via Subgradient Optimization and Relatives. *The Annals of Statistics* **45,** 2328–2364 (2017).

156. Mölder, F. *et al.* Sustainable Data Analysis with Snakemake. *F1000Research* **10** (2021).

157. R Core Team. *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna, Austria, 2022).

158. Purcell, S. & Chang, C. *PLINK 2.00 Alpha*

159. Chang, C. C. *et al.* Second-Generation PLINK: Rising to the Challenge of Larger and Richer Datasets. *GigaScience* **4,** s13742-015-0047–8 (2015).

160. Band, G. & Marchini, J. *BGEN: A Binary File Format for Imputed Genotype and Haplotype Data* `https://www.biorxiv.org/content/10.1101/308296v2` (2023). preprint.

161. Privé, F., Aschard, H., Ziyatdinov, A. & Blum, M. G. B. Efficient Analysis of Large-Scale Genome-Wide Data with Two R Packages: Bigstatsr and Bigsnpr. *Bioinformatics* **34,** 2781–2787 (2018).

162. Hutchinson, A., Asimit, J. & Wallace, C. Fine-Mapping Genetic Associations. *Human Molecular Genetics* **29,** R81–R88 (2020).

163. Schaid, D. J., Chen, W. & Larson, N. B. From Genome-Wide Associations to Candidate Causal Variants by Statistical Fine-Mapping. *Nature Reviews Genetics* **19,** 491–504 (2018).

164. Visscher, P. M., Hill, W. G. & Wray, N. R. Heritability in the Genomics Era — Concepts and Misconceptions. *Nature Reviews Genetics* **9,** 255–266 (4 2008).

165. Wray, N. R. & Visscher, P. M. Estimating Trait Heritability. *Nature Education* **1,** 29 (2008).

166. Evans, L. M. *et al.* Comparison of Methods That Use Whole Genome Data to Estimate the Heritability and Genetic Architecture of Complex Traits. *Nature Genetics* **50,** 737–745 (5 2018).

167. Barry, C.-J. S. *et al.* How to Estimate Heritability: A Guide for Genetic Epidemiologists. *International Journal of Epidemiology* **52,** 624–632 (2023).

168. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: A Tool for Genome-wide Complex Trait Analysis. *The American Journal of Human Genetics* **88,** 76–82 (2011).

169. Loh, P.-R. *et al.* Efficient Bayesian Mixed-Model Analysis Increases Association Power in Large Cohorts. *Nature Genetics* **47,** 284–290 (2015).

170. Mbatchou, J. *et al.* Computationally Efficient Whole-Genome Regression for Quantitative and Binary Traits. *Nature Genetics* **53,** 1097–1103 (7 2021).

171. Speed, D., Cai, N., Johnson, M. R., Nejentsev, S. & Balding, D. J. Reevaluation of SNP Heritability in Complex Human Traits. *Nature Genetics* **49,** 986–992 (7 2017).

172. Lowe, D., Sanvictores, T., Zubair, M. & John, S. *Alkaline Phosphatase* in *StatPearls* (StatPearls Publishing, Treasure Island, Fla., 2023).

173. Rader, B. A. Alkaline Phosphatase, an Unconventional Immune Protein. *Frontiers in Immunology* **8** (2017).

174. Vimalraj, S. Alkaline Phosphatase: Structure, Expression and Its Function in Bone Mineralization. *Gene* **754,** 144855 (2020).

175. Pratt, D. S. & Kaplan, M. M. Evaluation of Abnormal Liver-Enzyme Results in Asymptomatic Patients. *New England Journal of Medicine* **342,** 1266–1271 (2000).

176. Goltzman, D. & Miao, D. *Alkaline Phosphatase* in *Encyclopedia of Endocrine Diseases* (ed Martini, L.) 164–169 (Elsevier, New York, N.Y., 2004).

177. Shaban, S. M., Byeok Jo, S., Hafez, E., Ho Cho, J. & Kim, D.-H. A Comprehensive Overview on Alkaline Phosphatase Targeting and Reporting Assays. *Coordination Chemistry Reviews* **465,** 214567 (2022).

178. Bamford, K. F., Harris, H., Luffman, J. E., Robson, E. B. & Cleghorn, T. E. Serum-Alkaline-Phosphatase and the ABO Blood-Groups. *The Lancet* **285,** 530–531 (1965).

179. Langman, M. J. S., Constantinopoulos, A. & Bouchier, I. a. D. ABO Blood Groups, Secretor Status, and Intestinal Mucosal Concentrations of Alkaline Phosphatase. *Nature* **217,** 863–865 (5131 1968).

180. Bayer, P. M., Hotschek, H. & Knoth, E. Intestinal Alkaline Phosphatase and the Abo Blood Group System — a New Aspect. *Clinica Chimica Acta* **108,** 81–87 (1980).

181. Domar, U., Hirano, K. & Stigbrand, T. Serum Levels of Human Alkaline Phosphatase Isozymes in Relation to Blood Groups. *Clinica Chimica Acta* **203,** 305–313 (1991).

182. Nakano, T. *et al.* Involvement of Intestinal Alkaline Phosphatase in Serum Apolipoprotein B-48 Level and Its Association with ABO and Secretor Blood Group Types. *Biochemical and Biophysical Research Communications* **341,** 33–38 (2006).

183. Domar, U., Karpe, F., Hamsten, A., Stigbrand, T. & Olivecrona, T. Human Intestinal Alkaline Phosphatase—Release to the Blood Is Linked to Lipid Absorption, but Removal from the Blood Is Not Linked to Lipoprotein Clearance. *European Journal of Clinical Investigation* **23,** 753–760 (1993).

184. Masuda, M., Okuda, K., Ikeda, D. D., Hishigaki, H. & Fujiwara, T. Interaction of Genetic Markers Associated with Serum Alkaline Phosphatase Levels in the Japanese Population. *Human Genome Variation* **2,** 1–6 (1 2015).

185. Lyon, M. S., Millard, L. A. C., Smith, G. D., Gaunt, T. R. & Tilling, K. *Hypothesis-Free Detection of Gene-Interaction Effects on Biomarker Concentration in UK Biobank Using*

Variance Prioritisation https://www.medrxiv.org/content/10.1101/2022.01.05.21268406v1 (2023). preprint.

186. Yuan, X. *et al.* Population-Based Genome-wide Association Studies Reveal Six Loci Influencing Plasma Levels of Liver Enzymes. *The American Journal of Human Genetics* **83,** 520–528 (2008).

187. Kamatani, Y. *et al.* Genome-Wide Association Study of Hematological and Biochemical Traits in a Japanese Population. *Nature Genetics* **42,** 210–215 (3 2010).

188. Chambers, J. C. *et al.* Genome-Wide Association Study Identifies Loci Influencing Concentrations of Liver Enzymes in Plasma. *Nature Genetics* **43,** 1131–1138 (11 2011).

189. Li, J. *et al.* Genome-Wide Association Study on Serum Alkaline Phosphatase Levels in a Chinese Population. *BMC Genomics* **14,** 684 (2013).

190. Sinnott-Armstrong, N. *et al.* Genetics of 35 Blood and Urine Biomarkers in the UK Biobank. *Nature Genetics* **53,** 185–194 (2021).

191. Backman, J. D. *et al.* Exome Sequencing and Analysis of 454,787 UK Biobank Participants. *Nature* **599,** 628–634 (7886 2021).

192. Arvanitis, M. *et al.* Genome-Wide Association and Multi-Omic Analyses Reveal ACTN2 as a Gene Linked to Heart Failure. *Nature Communications* **11,** 1122 (1 2020).

193. Lin, X. *et al.* Genome-Wide Association Study Identifies Novel Loci Associated with Serum Level of Vitamin B12 in Chinese Men. *Human Molecular Genetics* **21,** 2610–2617 (2012).

194. Soejima, M. & Koda, Y. Detection of the Weak-Secretor Rs1047781 (385A>T) Single Nucleotide Polymorphism Using an Unlabeled Probe High-Resolution Melting-Based Method. *Electrophoresis* **42,** 1362–1365 (2021).

195. Parmar, A. S. *et al.* Association Study of FUT2 (Rs601338) with Celiac Disease and Inflammatory Bowel Disease in the Finnish Population. *Tissue Antigens* **80,** 488–493 (2012).

196. Velkova, A. *et al.* The FUT2 Secretor Variant p.Trp154Ter Influences Serum Vitamin B12 Concentration via Holo-Haptocorrin, but Not Holo-Transcobalamin, and Is Associated with Haptocorrin Glycosylation. *Human Molecular Genetics* **26,** 4975–4988 (2017).

197. Wang, M., Gao, J., Liu, J., Zhao, X. & Lei, Y. Genomic Association vs. Serological Determination of ABO Blood Types in a Chinese Cohort, with Application in Mendelian Randomization. *Genes* **12,** 959 (7 2021).

198. Kelly, R. J., Rouquier, S., Giorgi, D., Lennon, G. G. & Lowe, J. B. Sequence and Expression of a Candidate for the Human Secretor Blood Group α(1,2)Fucosyltransferase Gene (FUT2). *Journal of Biological Chemistry* **270,** 4640–4649 (1995).

199. Hazra, A. *et al.* Common Variants of FUT2 Are Associated with Plasma Vitamin B12 Levels. *Nature Genetics* **40,** 1160–1162 (10 2008).

200. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: Functional Annotation of Genetic Variants from High-Throughput Sequencing Data. *Nucleic Acids Research* **38,** e164 (2010).

201. O'Leary, N. A. *et al.* Reference Sequence (RefSeq) Database at NCBI: Current Status, Taxonomic Expansion, and Functional Annotation. *Nucleic Acids Research* **44,** D733–D745 (2016).

202. Kanai, M. *et al.* Genetic Analysis of Quantitative Traits in the Japanese Population Links Cell Types to Complex Human Diseases. *Nature Genetics* **50,** 390–400 (3 2018).

203. Mori, H. *et al.* Catalytic Reaction Mechanism Based on α-Secondary Deuterium Isotope Effects in Hydrolysis of Trehalose by European Honeybee Trehalase. *Bioscience, Biotechnology, and Biochemistry* **73,** 2466–2473 (2009).

204. Belinky, F. *et al.* PathCards: Multi-Source Consolidation of Human Biological Pathways. *Database* **2015,** bav006 (2015).

205. Ghoussaini, M. *et al.* Open Targets Genetics: Systematic Identification of Trait-Associated Genes Using Large-Scale Genetics and Functional Genomics. *Nucleic Acids Research* **49,** D1311–D1320 (2021).

206. Mountjoy, E. *et al.* An Open Approach to Systematically Prioritize Causal Variants and Genes at All Published Human GWAS Trait-Associated Loci. *Nature Genetics* **53,** 1527–1533 (11 2021).

207. Murakami, Y. *et al.* Mechanism for Release of Alkaline Phosphatase Caused by Glycosylphosphatidylinositol Deficiency in Patients with Hyperphosphatasia Mental Retardation Syndrome. *Journal of Biological Chemistry* **287,** 6318–6325 (2012).

208. Kinoshita, T. & Fujita, M. Thematic Review Series: Glycosylphosphatidylinositol (GPI) Anchors: Biochemistry and Cell Biology Biosynthesis of GPI-anchored Proteins: Special Emphasis on GPI Lipid Remodeling. *Journal of Lipid Research* **57,** 6–24 (2016).

209. Tremblay-Laganière, C. *et al.* PIGH Deficiency Can Be Associated with Severe Neurodevelopmental and Skeletal Manifestations. *Clinical Genetics* **99,** 313–317 (2021).

210. Chiyonobu, T., Inoue, N., Morimoto, M., Kinoshita, T. & Murakami, Y. Glycosylphosphatidylinositol (GPI) Anchor Deficiency Caused by Mutations in PIGW Is Associated with West Syndrome and Hyperphosphatasia with Mental Retardation Syndrome. *Journal of Medical Genetics* **51,** 203–207 (2014).

211. Nioi, P. *et al.* Variant ASGR1 Associated with a Reduced Risk of Coronary Artery Disease. *New England Journal of Medicine* **374,** 2131–2141 (2016).

212. Ensembl. *Ensembl Variation - Calculated Variant Consequences* `https://www.ensembl.org/info/genome/variation/prediction/predicted_data.html` (2023).

213. Mann, H. B. & Whitney, D. R. On a Test of Whether One of Two Random Variables Is Stochastically Larger than the Other. *The Annals of Mathematical Statistics* **18,** 50–60 (1947).

214. Wilcoxon, F. Individual Comparisons by Ranking Methods. *Biometrics Bulletin* **1,** 80–83 (1945).

215. Karch, J. D. Psychologists Should Use Brunner-Munzel's Instead of Mann-Whitney's U Test as the Default Nonparametric Procedure. *Advances in Methods and Practices in Psychological Science* **4,** 2515245921999602 (2021).

216. Brunner, E. & Munzel, U. The Nonparametric Behrens-Fisher Problem: Asymptotic Theory and a Small-Sample Approximation. *Biometrical Journal* **42,** 17–25 (2000).

217. Ara, T. *Brunnermunzel: (Permuted) Brunner-Munzel Test* version 2.0. 2022.

218. Shameer, K. *et al.* A Genome- and Phenome-Wide Association Study to Identify Genetic Variants Influencing Platelet Count and Volume and Their Pleiotropic Effects. *Human Genetics* **133,** 95–109 (2014).

219. Astle, W. J. *et al.* The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease. *Cell* **167,** 1415–1429.e19 (2016).

220. Serbanovic-Canic, J. *et al.* Silencing of RhoA Nucleotide Exchange Factor, ARHGEF3, Reveals Its Unexpected Role in Iron Uptake. *Blood* **118,** 4967–4976 (2011).

221. Chen, M.-H. *et al.* Trans-Ethnic and Ancestry-Specific Blood-Cell Genetics in 746,667 Individuals from 5 Global Populations. *Cell* **182,** 1198–1213.e14 (2020).

222. Krutein, M. C. *et al.* Restoring RUNX1 Deficiency in RUNX1 Familial Platelet Disorder by Inhibiting Its Degradation. *Blood Advances* **5,** 687–699 (2021).

223. Vuckovic, D. *et al.* The Polygenic and Monogenic Basis of Blood Traits and Diseases. *Cell* **182,** 1214–1231.e11 (2020).

224. Miyata, J. *et al.* 12/15-Lipoxygenase Regulates IL-33-Induced Eosinophilic Airway Inflammation in Mice. *Frontiers in Immunology* **12** (2021).

225. York, M. J. *Chapter 14 - Clinical Pathology* in *A Comprehensive Guide to Toxicology in Nonclinical Drug Development* (ed Faqi, A. S.) 2nd ed., 325–374 (Academic Press, Boston, Mass., 2017).

226. Ozer, J., Ratner, M., Shaw, M., Bailey, W. & Schomaker, S. The Current State of Serum Biomarkers of Hepatotoxicity. *Toxicology* **245,** 194–205 (2008).

227. McCormick, S. P. A. Lipoprotein(a): Biology and Clinical Importance. *The Clinical Biochemist Reviews* **25,** 69–80 (2004).

228. Trinder, M., Uddin, M. M., Finneran, P., Aragam, K. G. & Natarajan, P. Clinical Utility of Lipoprotein(a) and LPA Genetic Risk Score in Risk Prediction of Incident Atherosclerotic Cardiovascular Disease. *JAMA Cardiology* **6,** 287–295 (2021).

229. Feingold, K. R. *Introduction to Lipids and Lipoproteins* in *Endotext* (eds Feingold, K. R. *et al.*) (MDText.com, Inc., South Dartmouth, Mass., 2000).

230. Orsó, E. & Schmitz, G. Lipoprotein(a) and Its Role in Inflammation, Atherosclerosis and Malignancies. *Clinical Research in Cardiology Supplements* **12,** 31–37 (Suppl 1 2017).

231. Wilson, D. P. *et al.* Use of Lipoprotein(a) in Clinical Practice: A Biomarker Whose Time Has Come. A Scientific Statement from the National Lipid Association. *Journal of Clinical Lipidology* **13,** 374–392 (2019).

232. Kulminski, A. M. *et al.* Genetic and Regulatory Architecture of Alzheimer's Disease in the APOE Region. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring* **12,** e12008 (2020).

233. Bassendine, M. F., Taylor-Robinson, S. D., Fertleman, M., Khan, M. & Neely, D. Is Alzheimer's Disease a Liver Disease of the Brain? *Journal of Alzheimer's Disease* **75,** 1–14 (2020).

234. Prins, B. P. *et al.* Genome-Wide Analysis of Health-Related Biomarkers in the UK Household Longitudinal Study Reveals Novel Associations. *Scientific Reports* **7,** 11008 (1 2017).

235. Westerman, K. E. *et al.* Variance-Quantitative Trait Loci Enable Systematic Discovery of Gene-Environment Interactions for Cardiometabolic Serum Biomarkers. *Nature Communications* **13,** 3993 (1 2022).

236. Innes, H. *et al.* Genome-Wide Association Study for Alcohol-Related Cirrhosis Identifies Risk Loci in MARC1 and HNRNPUL1. *Gastroenterology* **159,** 1276–1289.e7 (2020).

237. Tabassum, R. *et al.* Genetic Architecture of Human Plasma Lipidome and Its Link to Cardiovascular Disease. *Nature Communications* **10,** 4329 (1 2019).

238. Klimentidis, Y. C. *et al.* Phenotypic and Genetic Characterization of Lower LDL Cholesterol and Increased Type 2 Diabetes Risk in the UK Biobank. *Diabetes* **69,** 2194–2205 (2020).

239. Kathiresan, S. *et al.* Six New Loci Associated with Blood Low-Density Lipoprotein Cholesterol, High-Density Lipoprotein Cholesterol or Triglycerides in Humans. *Nature Genetics* **40,** 189–197 (2 2008).

240. Barton, A. R., Sherman, M. A., Mukamel, R. E. & Loh, P.-R. Whole-Exome Imputation within UK Biobank Powers Rare Coding Variant Association and Fine-Mapping Analyses. *Nature Genetics* **53,** 1260–1269 (2021).

241. Surakka, I. *et al.* The Impact of Low-Frequency and Rare Variants on Lipid Levels. *Nature Genetics* **47,** 589–597 (2015).

242. Enrich, C., Rentero, C., Hierro, A. & Grewal, T. Role of Cholesterol in SNARE-mediated Trafficking on Intracellular Membranes. *Journal of Cell Science* **128,** 1071–1081 (2015).

243. Mahley, R. W., Huang, Y. & Weisgraber, K. H. Putting Cholesterol in Its Place: apoE and Reverse Cholesterol Transport. *The Journal of Clinical Investigation* **116,** 1226–1229 (2006).

244. Trowsdale, J. & Knight, J. C. Major Histocompatibility Complex Genomics and Human Disease. *Annual Review of Genomics and Human Genetics* **14,** 301–323 (2013).

245. Kulski, J. K., Suzuki, S. & Shiina, T. Human Leukocyte Antigen Super-Locus: Nexus of Genomic Supergenes, SNPs, Indels, Transcripts, and Haplotypes. *Human Genome Variation* **9,** 1–15 (1 2022).

246. Sakaue, S. *et al.* Tutorial: A Statistical Genetics Guide to Identifying HLA Alleles Driving Complex Disease. *Nature Protocols* **18,** 2625–2641 (9 2023).

247. Dendrou, C. A., Petersen, J., Rossjohn, J. & Fugger, L. HLA Variation and Disease. *Nature Reviews Immunology* **18,** 325–339 (5 2018).

248. Ryan, S. O. & Cobb, B. A. Roles for Major Histocompatibility Complex Glycosylation in Immune Function. *Seminars in Immunopathology* **34,** 425–441 (2012).

249. Carrington, M. & O'Brien, S. J. The Influence of HLA Genotype on AIDS. *Annual Review of Medicine* **54,** 535–551 (2003).

250. Hurley, C. K. Naming HLA Diversity: A Review of HLA Nomenclature. *Human Immunology* **82,** 457–465 (2021).

251. Yunis, E. *et al.* Inheritable Variable Sizes of DNA Stretches in the Human MHC: Conserved Extended Haplotypes and Their Fragments or Blocks. *Tissue Antigens* **62,** 1–20 (2003).

252. Dawkins, R. *et al.* Genomics of the Major Histocompatibility Complex: Haplotypes, Duplication, Retroviruses and Disease. *Immunological Reviews* **167,** 275–304 (1999).

253. Debebe, B. J. *et al.* Identifying the Immune Interactions Underlying HLA Class I Disease Associations. *eLife* **9,** e54558 (2020).

254. Fernando, M. M. A. *et al.* Defining the Role of the MHC in Autoimmunity: A Review and Pooled Analysis. *PLOS Genetics* **4,** e1000024 (2008).

255. Shiina, T., Hosomichi, K., Inoko, H. & Kulski, J. K. The HLA Genomic Loci Map: Expression, Interaction, Diversity and Disease. *Journal of Human Genetics* **54,** 15–39 (1 2009).

256. Amiel, J. L. *Study of the Leukocyte Phenotypes in Hodkin's Disease* in *Histocompatibility Testing* (eds Curtoni, E. S., Mattiuz, P. L. & Tosi, R. M.) 79–81 (Munksgaard, Copenhagen, 1967).

257. Grumet, F. C., Coukell, A., Bodmer, J. G., Bodmer, W. F. & McDevitt, H. O. Histocompatibility (HL-A) Antigens Associated with Systemic Lupus Erythematosus. *New England Journal of Medicine* **285,** 193–196 (1971).

258. Matzaraki, V., Kumar, V., Wijmenga, C. & Zhernakova, A. The MHC Locus and Genetic Susceptibility to Autoimmune and Infectious Diseases. *Genome Biology* **18,** 76 (2017).

259. Mulder, D. J. HL-A Antigens and Cœliac Disease. *The Lancet* **304,** 727 (1974).

260. Bevan, S. *et al.* Contribution of the MHC Region to the Familial Risk of Coeliac Disease. *Journal of Medical Genetics* **36,** 687–690 (1999).

261. De Menthon, M., LaValley, M. P., Maldini, C., Guillevin, L. & Mahr, A. HLA–B51/B5 and the Risk of Behçet's Disease: A Systematic Review and Meta-Analysis of Case–Control Genetic Association Studies. *Arthritis Care & Research* **61,** 1287–1296 (2009).

262. Howson, J. M. M., Walker, N. M., Clayton, D., Todd, J. A. & Diabetes, G. C. Confirmation of HLA Class II Independent Type 1 Diabetes Associations in the Major Histocompatibility Complex Including HLA-B and HLA-A. *Diabetes, Obesity and Metabolism* **11,** 31–45 (2009).

263. Todd, J. A., Bell, J. I. & McDevitt, H. O. HLA-DQβ Gene Contributes to Susceptibility and Resistance to Insulin-Dependent Diabetes Mellitus. *Nature* **329,** 599–604 (6140 1987).

264. Hu, X. *et al.* Additive and Interaction Effects at Three Amino Acid Positions in HLA-DQ and HLA-DR Molecules Drive Type 1 Diabetes Risk. *Nature Genetics* **47,** 898–905 (8 2015).

265. Cucca, F. *et al.* A Correlation between the Relative Predisposition of MHC Class II Alleles to Type 1 Diabetes and the Structure of Their Proteins. *Human Molecular Genetics* **10,** 2025–2037 (2001).

266. Raychaudhuri, S. *et al.* Five Amino Acids in Three HLA Proteins Explain Most of the Association between MHC and Seropositive Rheumatoid Arthritis. *Nature Genetics* **44,** 291–296 (3 2012).

267. Jersild, C. *et al.* Histocompatibility Determinants in Multiple Sclerosis, with Special Reference to Clinical Course. *The Lancet* **302,** 1221–1225 (1973).

268. Fogdell, A., Hillert, J., Sachs, C. & Olerup, O. The Multiple Sclerosis- and Narcolepsy-Associated HLA Class II Haplotype Includes the DRB5*0101 Allele. *Tissue Antigens* **46,** 333–336 (1995).

269. Ramagopalan, S. V. & Ebers, G. C. Multiple Sclerosis: Major Histocompatibility Complexity and Antigen Presentation. *Genome Medicine* **1,** 105 (2009).

270. The International Multiple Sclerosis Genetics Consortium. Class II HLA Interactions Modulate Genetic Risk for Multiple Sclerosis. *Nature Genetics* **47,** 1107–1113 (2015).

271. Kular, L. *et al.* DNA Methylation as a Mediator of HLA-DRB1*15:01 and a Protective Variant in Multiple Sclerosis. *Nature Communications* **9,** 2397 (1 2018).

272. Carrington, M. *et al.* HLA and HIV-1: Heterozygote Advantage and B*35-Cw*04 Disadvantage. *Science* **283,** 1748–1752 (1999).

273. The International HIV Controllers Study. The Major Genetic Determinants of HIV-1 Control Affect HLA Class I Peptide Presentation. *Science* **330,** 1551–1557 (2010).

274. Augusto, D. G. *et al.* A Common Allele of HLA Is Associated with Asymptomatic SARS-CoV-2 Infection. *Nature* **620,** 128–136 (7972 2023).

275. The Australo-Anglo-American Spondyloarthritis Consortium (TASC) *et al.* Interaction between ERAP1 and HLA-B27 in Ankylosing Spondylitis Implicates Peptide Handling in the Mechanism for HLA-B27 in Disease Susceptibility. *Nature Genetics* **43,** 761–767 (2011).

276. Cortes, A. *et al.* Major Histocompatibility Complex Associations of Ankylosing Spondylitis Are Complex and Involve Further Epistasis with ERAP1. *Nature Communications* **6,** 7146 (1 2015).

277. Kirino, Y. *et al.* Genome-Wide Association Analysis Identifies New Susceptibility Loci for Behçet's Disease and Epistasis between HLA-B*51 and ERAP1. *Nature Genetics* **45,** 202–207 (2 2013).

278. Martin, M. P. *et al.* Epistatic Interaction between KIR3DS1 and HLA-B Delays the Progression to AIDS. *Nature Genetics* **31,** 429–434 (4 2002).

279. Nelson, G. W. *et al.* Cutting Edge: Heterozygote Advantage in Autoimmune Disease: Hierarchy of Protection/Susceptibility Conferred by HLA and Killer Ig-Like Receptor Combinations in Psoriatic Arthritis1. *The Journal of Immunology* **173,** 4273–4276 (2004).

280. Van der Slik, A. R. *et al.* KIR in Type 1 Diabetes: Disparate Distribution of Activating and Inhibitory Natural Killer Cell Receptors in Patients Versus HLA-Matched Control Subjects. *Diabetes* **52,** 2639–2642 (2003).

281. Hughes, T. *et al.* Evidence for Gene–Gene Epistatic Interactions among Susceptibility Loci for Systemic Lupus Erythematosus. *Arthritis & Rheumatism* **64,** 485–492 (2012).

282. Oliveira-Caramez, M. L. de *et al.* Evidence for Epistatic Interaction between HLA-G and LILRB1 in the Pathogenesis of Nonsegmental Vitiligo. *Cells* **12,** 630 (4 2023).

283. Martel, P. *et al.* Epistasis between DSG1 and HLA Class II Genes in Pemphigus Foliaceus. *Genes & Immunity* **3,** 205–210 (4 2002).

284. Gregersen, J. W. *et al.* Functional Epistasis on a Common MHC Haplotype Associated with Multiple Sclerosis. *Nature* **443,** 574–577 (7111 2006).

285. Lincoln, M. R. *et al.* Epistasis among HLA-DRB1, HLA-DQA1, and HLA-DQB1 Loci Determines Multiple Sclerosis Susceptibility. *Proceedings of the National Academy of Sciences* **106,** 7542–7547 (2009).

286. Alcina, A. *et al.* Multiple Sclerosis Risk Variant HLA-DRB1*1501 Associates with High Expression of DRB1 Gene in Different Human Populations. *PLOS ONE* **7,** e29819 (2012).

287. Koeleman, B. P. C. *et al.* Genotype Effects and Epistasis in Type 1 Diabetes and HLA-DQ Trans Dimer Associations with Disease. *Genes & Immunity* **5,** 381–388 (5 2004).

288. Lenz, T. L. *et al.* Widespread Non-Additive and Interaction Effects within HLA Loci Modulate the Risk of Autoimmune Diseases. *Nature Genetics* **47,** 1085–1090 (9 2015).

289. Zhang, C. *et al.* Non-Additive and Epistatic Effects of HLA Polymorphisms Contributing to Risk of Adult Glioma. *Journal of Neuro-Oncology* **135,** 237–244 (2017).

290. Dilthey, A. T., Moutsianas, L., Leslie, S. & McVean, G. HLA*IMP—an Integrated Framework for Imputing Classical HLA Alleles from SNP Genotypes. *Bioinformatics* **27,** 968–972 (2011).

291. Leslie, S., Donnelly, P. & McVean, G. A Statistical Method for Predicting Classical HLA Alleles from SNP Data. *The American Journal of Human Genetics* **82,** 48–56 (2008).

292. Jia, X. *et al.* Imputing Amino Acid Polymorphisms in Human Leukocyte Antigens. *PLOS ONE* **8,** e64683 (2013).

293. Zheng, X. *et al.* HIBAG—HLA Genotype Imputation with Attribute Bagging. *The Pharmacogenomics Journal* **14,** 192–200 (2 2014).

294. Luo, Y. *et al.* A High-Resolution HLA Reference Panel Capturing Global Population Diversity Enables Multi-Ancestry Fine-Mapping in HIV Host Response. *Nature Genetics* **53,** 1504–1516 (10 2021).

295. Das, S. *et al.* Next-Generation Genotype Imputation Service and Methods. *Nature Genetics* **48,** 1284–1287 (10 2016).

296. Loh, P.-R. *et al.* Reference-Based Phasing Using the Haplotype Reference Consortium Panel. *Nature Genetics* **48,** 1443–1448 (11 2016).

297. Enczmann, J. *et al.* Next Generation Sequencing Identifies the HLA-DQA1*03:03 Allele in the Type 1 Diabetes Risk-Associated HLA-DQ8 Serotype. *Genes* **12,** 1879 (12 2021).

298. Jeffcoate, S. L. Diabetes Control and Complications: The Role of Glycated Haemoglobin, 25 Years On. *Diabetic Medicine* **21,** 657–665 (2004).

299. Noble, J. A. *et al.* The HLA Class I A Locus Affects Susceptibility to Type 1 Diabetes. *Human Immunology* **63,** 657–664 (2002).

300. Hollenbach, J. A. & Oksenberg, J. R. The Immunogenetics of Multiple Sclerosis: A Comprehensive Review. *Journal of Autoimmunity* **64,** 13–25 (2015).

301. Reveille, J. D. *et al.* HLA Class I and II Alleles in Susceptibility to Ankylosing Spondylitis. *Annals of the Rheumatic Diseases* **78,** 66–73 (2019).

302. Huang, J. *et al.* The Associations of HLA-A*02:01 and DRB1*11:01 with Hepatitis C Virus Spontaneous Clearance Are Independent of IL28B in the Chinese Population. *Scientific Reports* **6,** 31485 (1 2016).

303. Patsopoulos, N. A. *et al.* Fine-Mapping the Genetic Association of the Major Histocompatibility Complex in Multiple Sclerosis: HLA and Non-HLA Effects. *PLOS Genetics* **9,** e1003926 (2013).

304. Weiner Lachmi, K. *et al.* DQB1*06:02 Allele-Specific Expression Varies by Allelic Dosage, Not Narcolepsy Status. *Human Immunology* **73,** 405–410 (2012).

305. Fernando, M. M. A. *et al.* Identification of Two Independent Risk Factors for Lupus within the MHC in United Kingdom Families. *PLOS Genetics* **3,** e192 (2007).

306. Morris, D. L. *et al.* Unraveling Multiple MHC Gene Associations with Systemic Lupus Erythematosus: Model Choice Indicates a Role for HLA Alleles and Non-HLA Genes in Europeans. *The American Journal of Human Genetics* **91,** 778–793 (2012).

307. Gross, D. M. *et al.* Identification of LFA-1 as a Candidate Autoantigen in Treatment-Resistant Lyme Arthritis. *Science* **281,** 703–706 (1998).

308. Fernandez, C. A. *et al.* HLA-DRB1*07:01 Is Associated with a Higher Risk of Asparaginase Allergies. *Blood* **124,** 1266–1276 (2014).

309. Menegatti, J., Schub, D., Schäfer, M., Grässer, F. A. & Ruprecht, K. HLA-DRB1*15:01 Is a Co-Receptor for Epstein–Barr Virus, Linking Genetic and Environmental Risk Factors for Multiple Sclerosis. *European Journal of Immunology* **51,** 2348–2350 (2021).

310. Peterson, K. *et al.* Risk of Autoimmunity in EoE and Families: A Population-Based Cohort Study. *American Journal of Gastroenterology* **111,** 926 (2016).

311. Kurki, M. I. *et al.* FinnGen Provides Genetic Insights from a Well-Phenotyped Isolated Population. *Nature* **613,** 508–518 (7944 2023).

312. The All of Us Research Program Investigators. The "All of Us" Research Program. *New England Journal of Medicine* **381,** 668–676 (2019).

A Methods: proofs

A.1 PROOF OF PROPOSITION 3.4.2

Proof. We begin by plugging in the true model of y_i given in equation (3.11) into the expression for the OLS estimator given in equation (3.13):

$$\hat{\alpha}_2 = \frac{\sum_i S_i^2 \sum_i x_{ij} S_i (S_i + \lambda_j x_{ij} R_i + \varepsilon_i)}{\sum_i S_i^2 \sum_i x_{ij}^2 S_i^2 - \left(\sum_i S_i^2 x_{ij}\right)^2} - \frac{\sum_i S_i^2 x_{ij} \sum_i S_i (S_i + \lambda_j x_{ij} R_i + \varepsilon_i)}{\sum_i S_i^2 \sum_i x_{ij}^2 S_i^2 - \left(\sum_i S_i^2 x_{ij}\right)^2}.$$

Replacing R_i with $S_i - Q_i$ following equation (3.4), we obtain:

$$\begin{aligned}\hat{\alpha}_2 = &\frac{\sum_i S_i^2 \sum_i x_{ij} S_i [S_i + \lambda_j x_{ij} (S_i - Q_i) + \varepsilon_i]}{\sum_i S_i^2 \sum_i x_{ij}^2 S_i^2 - \left(\sum_i S_i^2 x_{ij}\right)^2} \\ &- \frac{\sum_i S_i^2 x_{ij} \sum_i S_i [S_i + \lambda_j x_{ij} (S_i - Q_i) + \varepsilon_i]}{\sum_i S_i^2 \sum_i x_{ij}^2 S_i^2 - \left(\sum_i S_i^2 x_{ij}\right)^2}.\end{aligned}$$

We now isolate λ_j:

$$\begin{aligned}\hat{\alpha}_2 = \lambda_j &\left[\frac{\sum_i S_i^2 \sum_i x_{ij}^2 S_i^2 - \sum_i S_i^2 \sum_i x_{ij}^2 S_i Q_i}{\sum_i S_i^2 \sum_i x_{ij}^2 S_i^2 - \left(\sum_i S_i^2 x_{ij}\right)^2} \right. \\ &\left. - \frac{\sum_i S_i^2 x_{ij} \sum_i S_i^2 x_{ij} - \sum_i S_i^2 x_{ij} \sum_i S_i x_{ij} Q_i}{\sum_i S_i^2 \sum_i x_{ij}^2 S_i^2 - \left(\sum_i S_i^2 x_{ij}\right)^2} \right] \\ &+ \frac{\sum_i S_i^2 \sum_i x_{ij} S_i (S_i + \varepsilon_i)}{\sum_i S_i^2 \sum_i x_{ij}^2 S_i^2 - \left(\sum_i S_i^2 x_{ij}\right)^2} \\ &- \frac{\sum_i S_i^2 x_{ij} \sum_i S_i (S_i + \varepsilon_i)}{\sum_i S_i^2 \sum_i x_{ij}^2 S_i^2 - \left(\sum_i S_i^2 x_{ij}\right)^2}.\end{aligned}$$

Rearranging the first two fractions and cancelling terms in the second two, we get:

$$\hat{\alpha}_2 = \lambda_j \left[1 - \frac{\sum_i S_i^2 \sum_i x_{ij}^2 S_i Q_i - \sum_i S_i^2 x_{ij} \sum_i S_i x_{ij} Q_i}{\sum_i S_i^2 \sum_i x_{ij}^2 S_i^2 - \left(\sum_i S_i^2 x_{ij}\right)^2} \right] + \frac{\sum_i S_i^2 \sum_i x_{ij} S_i \varepsilon_i - \sum_i S_i^2 x_{ij} \sum_i S_i \varepsilon_i}{\sum_i S_i^2 \sum_i x_{ij}^2 S_i^2 - \left(\sum_i S_i^2 x_{ij}\right)^2}.$$

We can now use the Weak Law of Large Numbers (since our samples are assumed to be i.i.d.) and Slutsky's Theorem to examine the estimator's behaviour as the sample size becomes large:

$$\hat{\alpha}_2 \xrightarrow{\text{p}} \lambda_j \left[1 - \frac{\mathbb{E}(S_i^2)\mathbb{E}(x_{ij}^2 S_i Q_i) - \mathbb{E}(S_i^2 x_{ij})\mathbb{E}(S_i x_{ij} Q_i)}{\mathbb{E}(S_i^2)\mathbb{E}(x_{ij}^2 S_i^2) - \left[\mathbb{E}(S_i^2 x_{ij})\right]^2} \right] + \frac{\mathbb{E}(S_i^2)\mathbb{E}(x_{ij} S_i \varepsilon_i) - \mathbb{E}(S_i^2 x_{ij})\mathbb{E}(S_i \varepsilon_i)}{\mathbb{E}(S_i^2)\mathbb{E}(x_{ij}^2 S_i^2) - \left[\mathbb{E}(S_i^2 x_{ij})\right]^2}.$$

Since, by assumption, x_{ij} is uncorrelated with S_i and S_i^2, and ε_i is independent from x_{ij} and S_i, and given that x_{ij} is also assumed to have mean zero, we have:

$$\mathbb{E}(S_i^2 x_{ij}) = \mathbb{E}(x_{ij} S_i \varepsilon_i) = 0.$$

Moreover, we assumed $\mathbb{E}(x_{ij}^2) = 1$ and that x_{ij}^2 is uncorrelated with S_i, Q_i and S_i^2, which gives:

$$\mathbb{E}(x_{ij}^2 S_i Q_i) = \mathbb{E}(x_{ij}^2)\mathbb{E}(S_i Q_i) = \mathbb{E}(S_i Q_i), \text{ and}$$
$$\mathbb{E}(x_{ij}^2 S_i^2) = \mathbb{E}(x_{ij}^2)\mathbb{E}(S_i^2) = \mathbb{E}(S_i^2).$$

Removing null terms and substituting in these simplified expressions, we obtain:

$$\hat{\alpha}_2 \xrightarrow{\text{p}} \lambda_j \left[1 - \frac{\mathbb{E}(S_i^2)\mathbb{E}(S_i Q_i)}{[\mathbb{E}(S_i^2)]^2} \right] \xrightarrow{\text{p}} \lambda_j \left[1 - \frac{\mathbb{E}(S_i Q_i)}{\mathbb{E}(S_i^2)} \right].$$

Finally, since $S_i = R_i + Q_i$ and all variables have mean zero, we can write:

$$\begin{aligned} 1 - \frac{\mathbb{E}(S_i Q_i)}{\mathbb{E}(S_i^2)} &= \frac{\mathbb{E}(S_i(S_i - Q_i))}{\mathbb{E}(S_i^2)} \\ &= \frac{\mathrm{Cov}(R_i, S_i)}{\mathrm{Var}(S_i)}, \end{aligned}$$

which gives the desired result. □

A.2 Proof of Proposition 3.4.3

Proof. We plug the true model of y_i given in equation (3.14) into the expression for the OLS estimator given in equation (3.16):

$$\hat{\alpha}_2 = \frac{\sum_i \hat{S}_i^2 \sum_i x_{ij} \hat{S}_i (S_i + \lambda_j x_{ij} S_i + \varepsilon_i)}{\sum_i \hat{S}_i^2 \sum_i x_{ij}^2 \hat{S}_i^2 - \left(\sum_i \hat{S}_i^2 x_{ij}\right)^2} - \frac{\sum_i \hat{S}_i^2 x_{ij} \sum_i \hat{S}_i (S_i + \lambda_j x_{ij} S_i + \varepsilon_i)}{\sum_i \hat{S}_i^2 \sum_i x_{ij}^2 \hat{S}_i^2 - \left(\sum_i \hat{S}_i^2 x_{ij}\right)^2}.$$

Since S_i is now observed with error, we use equation (3.7) to replace it:

$$\begin{aligned} \hat{\alpha}_2 = &\frac{\sum_i \hat{S}_i^2 \sum_i x_{ij} \hat{S}_i [\hat{S}_i - \zeta_i + \lambda_j x_{ij} (\hat{S}_i - \zeta_i) + \varepsilon_i]}{\sum_i \hat{S}_i^2 \sum_i x_{ij}^2 \hat{S}_i^2 - \left(\sum_i \hat{S}_i^2 x_{ij}\right)^2} \\ &- \frac{\sum_i \hat{S}_i^2 x_{ij} \sum_i \hat{S}_i [\hat{S}_i - \zeta_i + \lambda_j x_{ij} (\hat{S}_i - \zeta_i) + \varepsilon_i]}{\sum_i \hat{S}_i^2 \sum_i x_{ij}^2 \hat{S}_i^2 - \left(\sum_i \hat{S}_i^2 x_{ij}\right)^2}. \end{aligned}$$

Isolating λ_j:

$$\begin{aligned} \hat{\alpha}_2 = \lambda_j &\left[\frac{\sum_i \hat{S}_i^2 \sum_i x_{ij}^2 \hat{S}_i^2 - \sum_i \hat{S}_i^2 \sum_i x_{ij}^2 \hat{S}_i \zeta_i}{\sum_i \hat{S}_i^2 \sum_i x_{ij}^2 \hat{S}_i^2 - \left(\sum_i \hat{S}_i^2 x_{ij}\right)^2} \right. \\ &\left. - \frac{\sum_i \hat{S}_i^2 x_{ij} \sum_i \hat{S}_i^2 x_{ij} - \sum_i \hat{S}_i^2 x_{ij} \sum_i \hat{S}_i x_{ij} \zeta_i}{\sum_i \hat{S}_i^2 \sum_i x_{ij}^2 \hat{S}_i^2 - \left(\sum_i \hat{S}_i^2 x_{ij}\right)^2} \right] \\ &+ \frac{\sum_i \hat{S}_i^2 \sum_i x_{ij} \hat{S}_i (\hat{S}_i - \zeta_i + \varepsilon_i)}{\sum_i \hat{S}_i^2 \sum_i x_{ij}^2 \hat{S}_i^2 - \left(\sum_i \hat{S}_i^2 x_{ij}\right)^2} \\ &- \frac{\sum_i \hat{S}_i^2 x_{ij} \sum_i \hat{S}_i (\hat{S}_i - \zeta_i + \varepsilon_i)}{\sum_i \hat{S}_i^2 \sum_i x_{ij}^2 \hat{S}_i^2 - \left(\sum_i \hat{S}_i^2 x_{ij}\right)^2}. \end{aligned}$$

Rearranging the first two fractions and cancelling terms in the second two, we get:

$$\hat{\alpha}_2 = \lambda_j \left[1 - \frac{\sum_i \hat{S}_i^2 \sum_i x_{ij}^2 \hat{S}_i \zeta_i - \sum_i \hat{S}_i^2 x_{ij} \sum_i \hat{S}_i x_{ij} \zeta_i}{\sum_i \hat{S}_i^2 \sum_i x_{ij}^2 \hat{S}_i^2 - \left(\sum_i \hat{S}_i^2 x_{ij}\right)^2} \right] + \frac{\sum_i \hat{S}_i^2 x_{ij} \sum_i \hat{S}_i (\zeta_i - \varepsilon_i) - \sum_i \hat{S}_i^2 \sum_i x_{ij} \hat{S}_i (\zeta_i - \varepsilon_i)}{\sum_i \hat{S}_i^2 \sum_i x_{ij}^2 \hat{S}_i^2 - \left(\sum_i \hat{S}_i^2 x_{ij}\right)^2}.$$

Using the Weak Law of Large Numbers and Slutsky's Theorem as before:

$$\hat{\alpha}_2 \xrightarrow{\text{p}} \lambda_j \left[1 - \frac{\mathbb{E}(\hat{S}_i^2)\mathbb{E}(x_{ij}^2 \hat{S}_i \zeta_i) - \mathbb{E}(\hat{S}_i^2 x_{ij})\mathbb{E}(\hat{S}_i x_{ij} \zeta_i)}{\mathbb{E}(\hat{S}_i^2)\mathbb{E}(x_{ij}^2 \hat{S}_i^2) - [\mathbb{E}(\hat{S}_i^2 x_{ij})]^2} \right] + \frac{\mathbb{E}(\hat{S}_i^2 x_{ij})\mathbb{E}(\hat{S}_i(\zeta_i - \varepsilon_i)) - \mathbb{E}(\hat{S}_i^2)\mathbb{E}(x_{ij} \hat{S}_i (\zeta_i - \varepsilon_i))}{\mathbb{E}(\hat{S}_i^2)\mathbb{E}(x_{ij}^2 \hat{S}_i^2) - [\mathbb{E}(\hat{S}_i^2 x_{ij})]^2}.$$

Since, by assumption, x_{ij} is uncorrelated with $\hat{S}_i$, $\hat{S}_i^2$, ζ_i and ε_i, and x_{ij}^2 is uncorrelated with $\hat{S}_i$, $\hat{S}_i^2$ and ζ_i, the expression above simplifies to:

$$\hat{\alpha}_2 \xrightarrow{\text{p}} \lambda_j \left[1 - \frac{\mathbb{E}(\hat{S}_i, \zeta_i)}{\mathbb{E}(\hat{S}_i^2)} \right].$$

And given that $\hat{S}_i = S_i + \zeta_i$ by construction, we can write:

$$\hat{\alpha}_2 \xrightarrow{\text{p}} \lambda_j \frac{\text{Cov}(\hat{S}_i, S_i)}{\text{Var}(\hat{S}_i)}.$$

□

A.3 Derivation of main result for Case 4

We plug the true model of y_i from equation (3.17) into the expression for the estimator given in equation (3.16):

$$\hat{\alpha}_2 = \frac{\sum_i \hat{S}_i^2 \sum_i x_{ij} \hat{S}_i (S_i + \lambda_j x_{ij} R_i + \varepsilon_i)}{\sum_i \hat{S}_i^2 \sum_i x_{ij}^2 \hat{S}_i^2 - \left(\sum_i \hat{S}_i^2 x_{ij}\right)^2} - \frac{\sum_i \hat{S}_i^2 x_{ij} \sum_i \hat{S}_i (S_i + \lambda_j x_{ij} R_i + \varepsilon_i)}{\sum_i \hat{S}_i^2 \sum_i x_{ij}^2 \hat{S}_i^2 - \left(\sum_i \hat{S}_i^2 x_{ij}\right)^2}.$$

We now use equation (3.8) to replace R_i:

$$\hat{\alpha}_2 = \frac{\sum_i \hat{S}_i^2 \sum_i x_{ij}\hat{S}_i[\hat{S}_i - \zeta_i + \lambda_j x_{ij}(\hat{S}_i - \zeta_i - Q_i) + \varepsilon_i]}{\sum_i \hat{S}_i^2 \sum_i x_{ij}^2 \hat{S}_i^2 - \left(\sum_i \hat{S}_i^2 x_{ij}\right)^2} - \frac{\sum_i \hat{S}_i^2 x_{ij} \sum_i \hat{S}_i[\hat{S}_i - \zeta_i + \lambda_j x_{ij}(\hat{S}_i - \zeta_i - Q_i) + \varepsilon_i]}{\sum_i \hat{S}_i^2 \sum_i x_{ij}^2 \hat{S}_i^2 - \left(\sum_i \hat{S}_i^2 x_{ij}\right)^2}.$$

Isolating λ_j:

$$\begin{aligned}\hat{\alpha}_2 = \lambda_j &\left[\frac{\sum_i \hat{S}_i^2 \sum_i x_{ij}^2 \hat{S}_i^2 - \sum_i \hat{S}_i^2 \sum_i x_{ij}^2 \hat{S}_i \zeta_i - \sum_i \hat{S}_i^2 \sum_i x_{ij}^2 \hat{S}_i Q_i}{\sum_i \hat{S}_i^2 \sum_i x_{ij}^2 \hat{S}_i^2 - \left(\sum_i \hat{S}_i^2 x_{ij}\right)^2}\right. \\ &\left. - \frac{\sum_i \hat{S}_i^2 x_{ij} \sum_i \hat{S}_i^2 x_{ij} - \sum_i \hat{S}_i^2 x_{ij} \sum_i \hat{S}_i x_{ij} \zeta_i - \sum_i \hat{S}_i^2 x_{ij} \sum_i \hat{S}_i x_{ij} Q_i}{\sum_i \hat{S}_i^2 \sum_i x_{ij}^2 \hat{S}_i^2 - \left(\sum_i \hat{S}_i^2 x_{ij}\right)^2}\right] \\ &+ \frac{\sum_i \hat{S}_i^2 \sum_i x_{ij}\hat{S}_i(\hat{S}_i - \zeta_i + \varepsilon_i)}{\sum_i \hat{S}_i^2 \sum_i x_{ij}^2 \hat{S}_i^2 - \left(\sum_i \hat{S}_i^2 x_{ij}\right)^2} \\ &- \frac{\sum_i \hat{S}_i^2 x_{ij} \sum_i \hat{S}_i(\hat{S}_i - \zeta_i + \varepsilon_i)}{\sum_i \hat{S}_i^2 \sum_i x_{ij}^2 \hat{S}_i^2 - \left(\sum_i \hat{S}_i^2 x_{ij}\right)^2}.\end{aligned}$$

Rearranging and cancelling terms, we obtain:

$$\begin{aligned}\hat{\alpha}_2 = \lambda_j &\left[1 - \frac{\sum_i \hat{S}_i^2 \sum_i x_{ij}^2 \hat{S}_i(\zeta_i + Q_i) - \sum_i \hat{S}_i^2 x_{ij} \sum_i \hat{S}_i x_{ij}(\zeta_i + Q_i)}{\sum_i \hat{S}_i^2 \sum_i x_{ij}^2 \hat{S}_i^2 - \left(\sum_i \hat{S}_i^2 x_{ij}\right)^2}\right] \\ &- \frac{\sum_i \hat{S}_i^2 \sum_i x_{ij}\hat{S}_i\zeta_i - \sum_i \hat{S}_i^2 x_{ij} \sum_i \hat{S}_i \zeta_i}{\sum_i \hat{S}_i^2 \sum_i x_{ij}^2 \hat{S}_i^2 - \left(\sum_i \hat{S}_i^2 x_{ij}\right)^2} \\ &+ \frac{\sum_i \hat{S}_i^2 \sum_i x_{ij}\hat{S}_i\varepsilon_i - \sum_i \hat{S}_i^2 x_{ij} \sum_i \hat{S}_i \varepsilon_i}{\sum_i \hat{S}_i^2 \sum_i x_{ij}^2 \hat{S}_i^2 - \left(\sum_i \hat{S}_i^2 x_{ij}\right)^2}.\end{aligned}$$

Finally we compute the probability limit:

$$\begin{aligned}\hat{\alpha}_2 \xrightarrow{\text{p}} \lambda_j \Bigg[1 - \frac{\mathbb{E}(\hat{S}_i^2)\mathbb{E}(x_{ij}^2 \hat{S}_i \zeta_i) - \mathbb{E}(\hat{S}_i^2 x_{ij})\mathbb{E}(\hat{S}_i x_{ij} \zeta_i)}{\mathbb{E}(\hat{S}_i^2)\mathbb{E}(x_{ij}^2 \hat{S}_i^2) - \left[\mathbb{E}(\hat{S}_i^2 x_{ij})\right]^2} \\ - \frac{\mathbb{E}(\hat{S}_i^2)\mathbb{E}(x_{ij}^2 \hat{S}_i Q_i) - \mathbb{E}(\hat{S}_i^2 x_{ij})\mathbb{E}(\hat{S}_i x_{ij} Q_i)}{\mathbb{E}(\hat{S}_i^2)\mathbb{E}(x_{ij}^2 \hat{S}_i^2) - \left[\mathbb{E}(\hat{S}_i^2 x_{ij})\right]^2} \Bigg] \\ - \frac{\mathbb{E}(\hat{S}_i^2)\mathbb{E}(x_{ij} \hat{S}_i \zeta_i) - \mathbb{E}(\hat{S}_i^2 x_{ij})\mathbb{E}(\hat{S}_i \zeta_i)}{\mathbb{E}(\hat{S}_i^2)\mathbb{E}(x_{ij}^2 \hat{S}_i^2) - \left[\mathbb{E}(\hat{S}_i^2 x_{ij})\right]^2} \\ + \frac{\mathbb{E}(\hat{S}_i^2)\mathbb{E}(x_{ij} \hat{S}_i \varepsilon_i) - \mathbb{E}(\hat{S}_i^2 x_{ij})\mathbb{E}(\hat{S}_i \varepsilon_i)}{\mathbb{E}(\hat{S}_i^2)\mathbb{E}(x_{ij}^2 \hat{S}_i^2) - \left[\mathbb{E}(\hat{S}_i^2 x_{ij})\right]^2}.\end{aligned}$$

Since ε_i is assumed to be independent of x_{ij} and $\hat{S}_i$, the fraction in the last line is equal to zero. Rearranging terms gives the desired expression.

B Data: list of traits

Table B.1: List of 97 quantitative traits analysed.
(continued on the next two pages)

Trait category	Description
Body size measures	Waist circumference
Body size measures	Hip circumference
Body size measures	Waist-to-hip ratio
Body size measures	Standing height
Body size measures	Sitting height
Body size measures	Sitting height, relative
Body size measures	Leg length
Body size measures	Leg length, relative
Body size measures	Body mass index (BMI)
Body size measures	Weight
Impedance measures	Body fat percentage
Impedance measures	Whole body fat mass
Impedance measures	Whole body fat-free mass
Impedance measures	Whole body water mass
Impedance measures	Basal metabolic rate
Impedance measures	Leg fat percentage (left)
Impedance measures	Leg fat mass (left)
Impedance measures	Leg fat-free mass (left)
Impedance measures	Leg predicted mass (left)
Impedance measures	Arm fat percentage (left)
Impedance measures	Arm fat mass (left)
Impedance measures	Arm fat-free mass (left)
Impedance measures	Arm predicted mass (left)
Impedance measures	Trunk fat percentage
Impedance measures	Trunk fat mass
Impedance measures	Trunk fat-free mass
Impedance measures	Trunk predicted mass

Table B.1: *(continued from the previous page)*

Trait category	Description
Blood pressure	Pulse rate
Blood pressure	Diastolic blood pressure
Blood pressure	Systolic blood pressure
Urine assays	Creatinine (enzymatic) in urine
Urine assays	Sodium in urine
Blood count	White blood cell (leukocyte) count
Blood count	Red blood cell (erythrocyte) count
Blood count	Haemoglobin concentration
Blood count	Haematocrit percentage
Blood count	Mean corpuscular volume
Blood count	Mean corpuscular haemoglobin
Blood count	Mean corpuscular haemoglobin concentration
Blood count	Red blood cell (erythrocyte) distribution width
Blood count	Platelet count
Blood count	Platelet crit
Blood count	Mean platelet (thrombocyte) volume
Blood count	Platelet distribution width
Blood count	Lymphocyte count
Blood count	Monocyte count
Blood count	Neutrophil count
Blood count	Eosinophil count
Blood count	Lymphocyte percentage
Blood count	Monocyte percentage
Blood count	Neutrophil percentage
Blood count	Eosinophil percentage
Blood count	Basophil percentage
Blood count	Reticulocyte percentage
Blood count	Reticulocyte count
Blood count	Mean reticulocyte volume
Blood count	Mean sphered cell volume
Blood count	Immature reticulocyte fraction
Blood count	High light scatter reticulocyte percentage
Blood count	High light scatter reticulocyte count
Blood biochemistry	Albumin (g/L)
Blood biochemistry	Alkaline phosphatase (U/L)
Blood biochemistry	Alanine aminotransferase (U/L)
Blood biochemistry	Apoliprotein A (g/L)
Blood biochemistry	Apoliprotein B (g/L)

Table B.1: *(continued from the previous page)*

Blood biochemistry	Aspartate aminotransferase (U/L)
Blood biochemistry	Direct bilirubin (umol/L)
Blood biochemistry	Urea (mmol/L)
Blood biochemistry	Calcium (mmol/L)
Blood biochemistry	Cholesterol (mmol/L)
Blood biochemistry	Creatinine (umol/L)
Blood biochemistry	C-reactive protein (mg/L)
Blood biochemistry	Cystatin C (mg/L)
Blood biochemistry	Gamma glutamyltransferase (U/L)
Blood biochemistry	Glucose (mmol/L)
Blood biochemistry	Glycated haemoglobin (mmol/mol)
Blood biochemistry	HDL cholesterol (mmol/L)
Blood biochemistry	IGF-1 (nmol/L)
Blood biochemistry	LDL direct (mmol/L)
Blood biochemistry	Lipoprotein A (nmol/L)
Blood biochemistry	Phosphate (mmol/L)
Blood biochemistry	SHBG (nmol/L)
Blood biochemistry	Total bilirubin (umol/L)
Blood biochemistry	Testosterone (nmol/L)
Blood biochemistry	Total protein (g/L)
Blood biochemistry	Triglycerides (mmol/L)
Blood biochemistry	Urate (umol/L)
Blood biochemistry	Vitamin D (nmol/L)
Other traits	Hand grip strength (left)
Other traits	Heel bone mineral density (BMD) T-score
Other traits	Number of incorrect matches in round
Other traits	Ankle spacing width
Other traits	Fluid intelligence score
Other traits	Mean time to correctly identify matches
Other traits	Forced expiratory volume in 1-second (FEV1)
Other traits	Forced vital capacity (FVC), Best measure
Other traits	FEV1/FVC ratio

C Simulations: supplementary results

Table C.1: False positives in null simulations with $\alpha = 0$, a heritability of 0.6 and QN residuals.

SNPs	Arc.	Tr.	LOCO PGS			Full PGS		
			GWS	FPR	Ind.	GWS	FPR	Ind.
100	Reg.	–	0	0	0	0	0	0
100	Reg.	Bef.	0	0	0	0	0	0
100	Reg.	Aft.	0	0	0	0	0	0
100	Cl.	–	0	0	0	0	0	0
100	Cl.	Bef.	0	0	0	0	0	0
100	Cl.	Aft.	1	7.88×10^{-8}	1	0	0	0
1K	Reg.	–	1	7.88×10^{-8}	1	0	0	0
1K	Reg.	Bef.	0	0	0	0	0	0
1K	Reg.	Aft.	1	7.88×10^{-8}	1	0	0	0
1K	Cl.	–	0	0	0	0	0	0
1K	Cl.	Bef.	0	0	0	5	3.94×10^{-7}	1
1K	Cl.	Aft.	0	0	0	0	0	0
10K	Reg.	–	1	7.88×10^{-8}	1	1	7.88×10^{-8}	0
10K	Reg.	Bef.	1	7.88×10^{-8}	1	0	0	0
10K	Reg.	Aft.	0	0	0	0	0	0
10K	Cl.	–	0	0	0	0	0	0
10K	Cl.	Bef.	0	0	0	0	0	0
10K	Cl.	Aft.	0	0	0	0	0	0
Combined FPR				2.19×10^{-8}			2.63×10^{-8}	

Table C.2: False positives in null simulations with $\alpha = -0.5$, a heritability of 0.3 and QN residuals.

SNPs	Arc.	Tr.	LOCO PGS			Full PGS		
			GWS	FPR	Ind.	GWS	FPR	Ind.
100	Reg.	–	1	7.88×10^{-8}	1	1	7.88×10^{-8}	0
100	Reg.	Bef.	2	1.58×10^{-7}	1	2	1.58×10^{-7}	0
100	Reg.	Aft.	0	0	0	0	0	0
100	Cl.	–	0	0	0	1	7.88×10^{-8}	1
100	Cl.	Bef.	0	0	0	0	0	0
100	Cl.	Aft.	0	0	0	0	0	0
1K	Reg.	–	0	0	0	0	0	0
1K	Reg.	Bef.	1	7.88×10^{-8}	1	1	7.88×10^{-8}	1
1K	Reg.	Aft.	0	0	0	0	0	0
1K	Cl.	–	0	0	0	0	0	0
1K	Cl.	Bef.	0	0	0	0	0	0
1K	Cl.	Aft.	0	0	0	0	0	0
10K	Reg.	–	0	0	0	0	0	0
10K	Reg.	Bef.	1	7.88×10^{-8}	1	0	0	0
10K	Reg.	Aft.	1	7.88×10^{-8}	1	0	0	0
10K	Cl.	–	0	0	0	1	7.88×10^{-8}	1
10K	Cl.	Bef.	0	0	0	0	0	0
10K	Cl.	Aft.	0	0	0	0	0	0
Combined FPR				2.63×10^{-8}			2.63×10^{-8}	

Table C.3: False positives in null simulations with $\alpha = -0.5$, a heritability of 0.6 and QN residuals.

SNPs	Arc.	Tr.	LOCO PGS			Full PGS		
			GWS	FPR	Ind.	GWS	FPR	Ind.
100	Reg.	–	0	0	0	0	0	0
100	Reg.	Bef.	0	0	0	0	0	0
100	Reg.	Aft.	0	0	0	0	0	0
100	Cl.	–	4	3.15×10^{-7}	1	0	0	0
100	Cl.	Bef.	3	2.36×10^{-7}	2	38	2.99×10^{-6}	2
100	Cl.	Aft.	0	0	0	0	0	0
1K	Reg.	–	0	0	0	0	0	0
1K	Reg.	Bef.	0	0	0	0	0	0
1K	Reg.	Aft.	0	0	0	0	0	0
1K	Cl.	–	0	0	0	1	7.88×10^{-8}	1
1K	Cl.	Bef.	1	7.88×10^{-8}	1	1	7.88×10^{-8}	0
1K	Cl.	Aft.	1	7.88×10^{-8}	1	0	0	0
10K	Reg.	–	0	0	0	0	0	0
10K	Reg.	Bef.	3	2.36×10^{-7}	2	1	7.88×10^{-8}	0
10K	Reg.	Aft.	0	0	0	0	0	0
10K	Cl.	–	0	0	0	0	0	0
10K	Cl.	Bef.	0	0	0	2	1.58×10^{-7}	1
10K	Cl.	Aft.	0	0	0	0	0	0
Combined FPR				5.25×10^{-8}			1.88×10^{-7}	

Table C.4: False positives in null simulations with $\alpha = 0$, a heritability of 0.3 and Winsorised residuals.

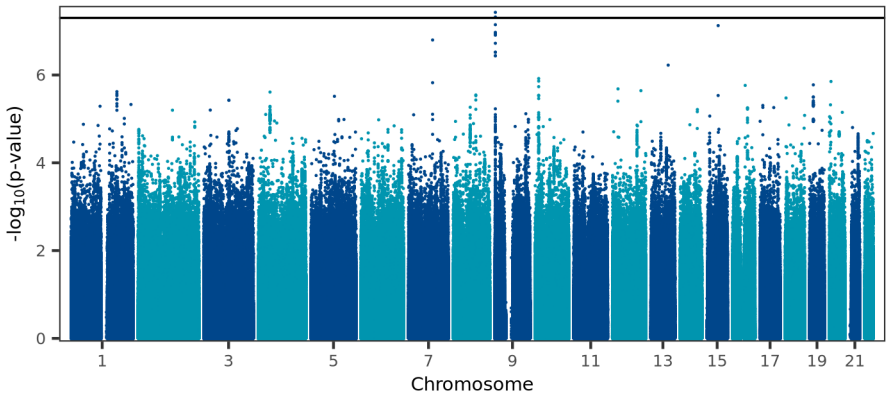
SNPs	Arc.	Tr.	LOCO PGS			Full PGS		
			GWS	FPR	Ind.	GWS	FPR	Ind.
100	Reg.	–	0	0	0	0	0	0
100	Reg.	Bef.	2	1.58×10^{-7}	1	0	0	0
100	Reg.	Aft.	0	0	0	0	0	0
100	Cl.	–	0	0	0	0	0	0
100	Cl.	Bef.	0	0	0	0	0	0
100	Cl.	Aft.	0	0	0	0	0	0
1K	Reg.	–	0	0	0	0	0	0
1K	Reg.	Bef.	0	0	0	0	0	0
1K	Reg.	Aft.	0	0	0	0	0	0
1K	Cl.	–	0	0	0	0	0	0
1K	Cl.	Bef.	0	0	0	0	0	0
1K	Cl.	Aft.	0	0	0	0	0	0
10K	Reg.	–	0	0	0	0	0	0
10K	Reg.	Bef.	0	0	0	0	0	0
10K	Reg.	Aft.	0	0	0	0	0	0
10K	Cl.	–	0	0	0	0	0	0
10K	Cl.	Bef.	0	0	0	0	0	0
10K	Cl.	Aft.	0	0	0	0	0	0
Combined FPR				8.76×10^{-9}			0	

Table C.5: False positives in null simulations with $\alpha = 0$, a heritability of 0.6 and Winsorised residuals.

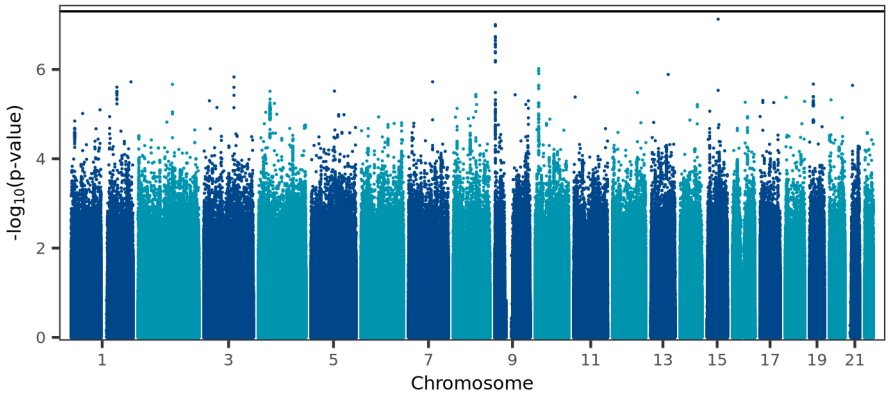
SNPs	Arc.	Tr.	LOCO PGS			Full PGS		
			GWS	FPR	Ind.	GWS	FPR	Ind.
100	Reg.	–	0	0	0	0	0	0
100	Reg.	Bef.	0	0	0	0	0	0
100	Reg.	Aft.	0	0	0	0	0	0
100	Cl.	–	0	0	0	0	0	0
100	Cl.	Bef.	311	2.45×10^{-5}	6	494	3.89×10^{-5}	1
100	Cl.	Aft.	0	0	0	0	0	0
1K	Reg.	–	0	0	0	0	0	0
1K	Reg.	Bef.	0	0	0	0	0	0
1K	Reg.	Aft.	0	0	0	0	0	0
1K	Cl.	–	0	0	0	0	0	0
1K	Cl.	Bef.	0	0	0	0	0	0
1K	Cl.	Aft.	0	0	0	0	0	0
10K	Reg.	–	0	0	0	0	0	0
10K	Reg.	Bef.	0	0	0	0	0	0
10K	Reg.	Aft.	0	0	0	0	0	0
10K	Cl.	–	0	0	0	0	0	0
10K	Cl.	Bef.	0	0	0	0	0	0
10K	Cl.	Aft.	0	0	0	0	0	0
Combined FPR				1.36×10^{-6}			2.16×10^{-6}	

Table C.6: False positives in null simulations with $\alpha = -0.5$, a heritability of 0.3 and Winsorised residuals.

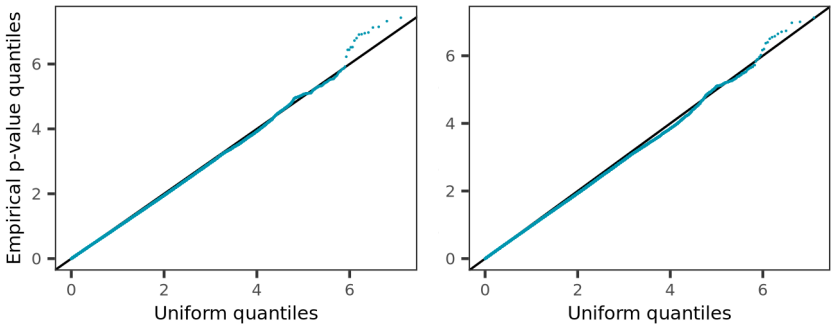
SNPs	Arc.	Tr.	LOCO PGS			Full PGS		
			GWS	FPR	Ind.	GWS	FPR	Ind.
100	Reg.	–	0	0	0	0	0	0
100	Reg.	Bef.	2	1.58×10^{-7}	1	0	0	0
100	Reg.	Aft.	0	0	0	0	0	0
100	Cl.	–	0	0	0	1	7.88×10^{-8}	1
100	Cl.	Bef.	7	5.52×10^{-7}	1	7	5.52×10^{-7}	0
100	Cl.	Aft.	0	0	0	0	0	0
1K	Reg.	–	0	0	0	0	0	0
1K	Reg.	Bef.	0	0	0	1	7.88×10^{-8}	0
1K	Reg.	Aft.	0	0	0	0	0	0
1K	Cl.	–	0	0	0	0	0	0
1K	Cl.	Bef.	0	0	0	0	0	0
1K	Cl.	Aft.	0	0	0	0	0	0
10K	Reg.	–	0	0	0	0	0	0
10K	Reg.	Bef.	1	7.88×10^{-8}	1	0	0	0
10K	Reg.	Aft.	0	0	0	0	0	0
10K	Cl.	–	0	0	0	0	0	0
10K	Cl.	Bef.	0	0	0	0	0	0
10K	Cl.	Aft.	0	0	0	0	0	0
Combined FPR				4.38×10^{-8}			3.94×10^{-8}	



(a) LOCO PGS

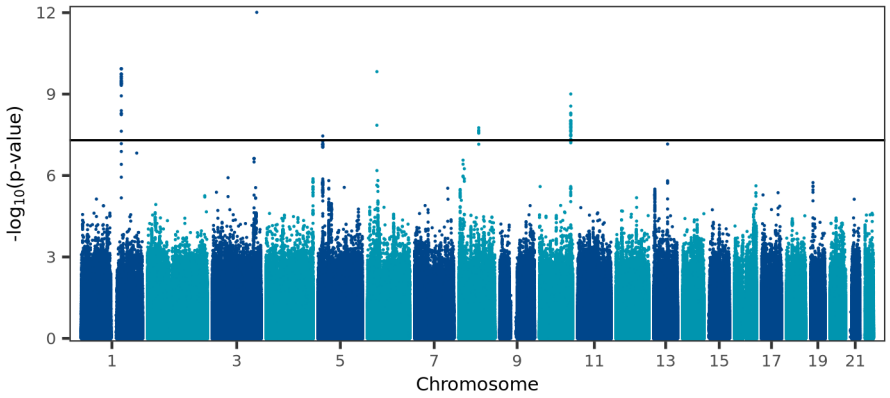


(b) Full PGS

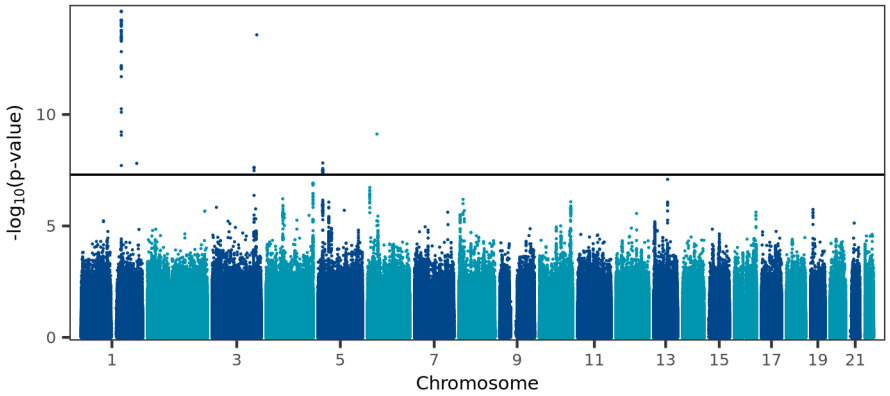


(c) LOCO PGS (d) Full PGS

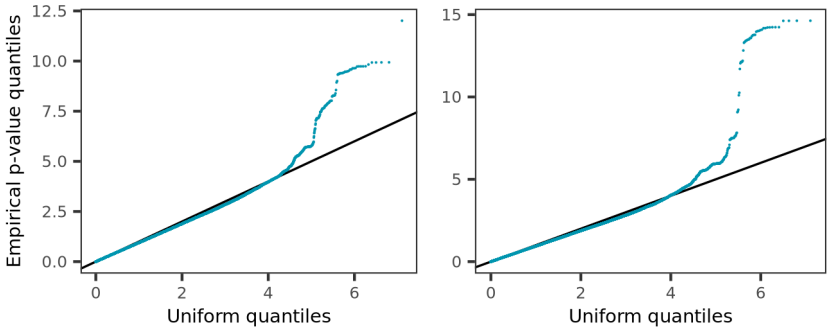
Figure C.1: Manhattan and QQ plots of results from the SNP × PGS interaction GWASs with the LOCO or full PGSs: null simulation with 100 causal loci, a regular architecture, a sigmoid transformation applied before adding noise, an α parameter of 0, a heritability of 0.3 and Winsorised residuals.



(a) LOCO PGS

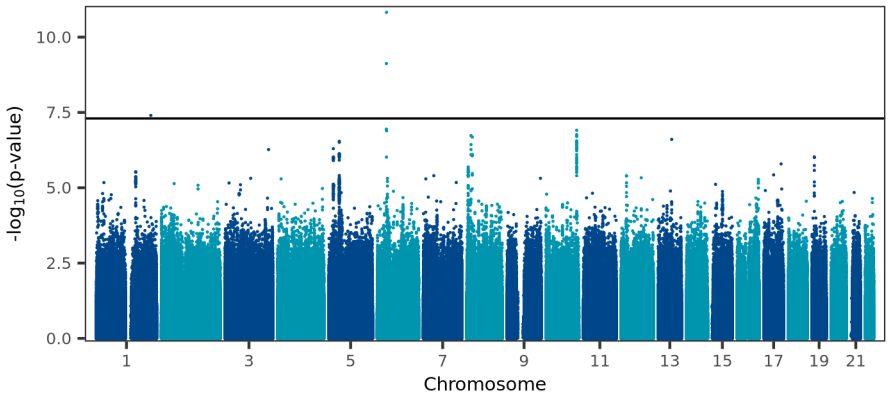


(b) Full PGS

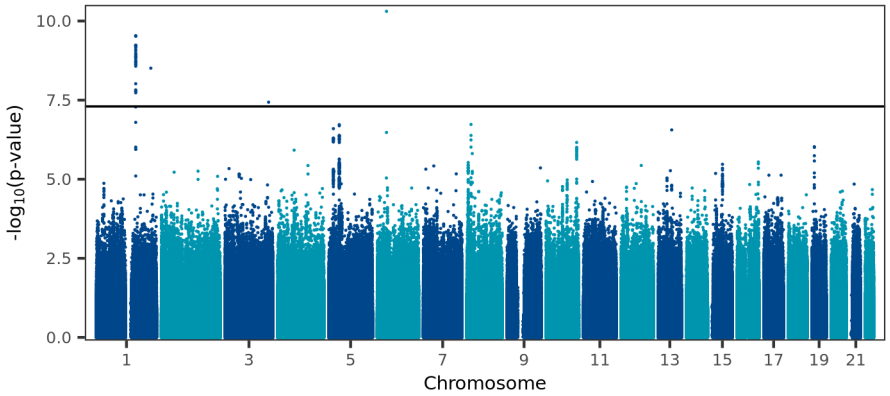


(c) LOCO PGS (d) Full PGS

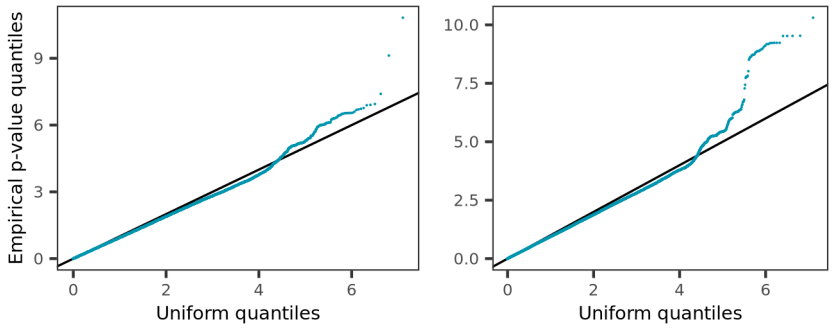
Figure C.2: Manhattan and QQ plots of results from the SNP × PGS interaction GWASs with the LOCO or full PGSs: null simulation with 100 causal loci, a clustered architecture, a sigmoid transformation applied before adding noise, an α parameter of -0.5, a heritability of 0.6 and Winsorised residuals.



(a) LOCO PGS



(b) Full PGS



(c) LOCO PGS (d) Full PGS

Figure C.3: Manhattan and QQ plots of results from the SNP × PGS interaction GWASs with the LOCO or full PGSs: null simulation with 100 causal loci, a clustered architecture, a sigmoid transformation applied before adding noise, an α parameter of -0.5, a heritability of 0.6 and QN residuals.

Table C.7: Log-transformed p-values with sign of corresponding coefficient obtained from SNP × PGS interaction testing for simulated traits with $\alpha = 0$ and Winsorised residuals.

MAF	Coeff.	Sign	Type	# SNPs affected	*Trait 1* Effect: 1%		*Trait 2* Effect: 10%		*Trait 3* Effect: 50%		*Trait 4* Effect: 100%	
					LOCO	Full	LOCO	Full	LOCO	Full	LOCO	Full
0.38	0.7	-	Direct effect	1	-0.1	0	-0.4	-0.5	-1.5	-2	-2	-2.3
0.42	-0.1	+	Direct effect	1	-0.3	-0.3	0	0.1	0.7	0.9	1.1	1.1
0.28	0.84	-	Direct effect	2%	0	-0.1	0.3	0.3	-0.9	-1	-6.9	-7.2
0.46	1.09	-	Direct effect	2%	-0.3	-0.2	-0.9	-1.3	-4.4	-4	**-8.4**	**-8.2**
0.13	-1.04	+	Direct effect	10%	3	3.3	4.2	4.3	**63.3**	**65.3**	**217.2**	**220.2**
0.34	-1.86	+	Direct effect	10%	0.3	0.4	1.1	0.8	**38**	**39.7**	**134.7**	**132.3**
0.12	-0.18	-	Direct effect	100%	-1	-0.7	**-135.7**	**-127.1**	**-3788.7**	**-3574.5**	**-2326.1**	**-2200.5**
0.42	-0.01	+	Direct effect	100%	2.2	2.8	**301.3**	**293.6**	**8754.7**	**8474.6**	**5445.6**	**5279.4**
0.11	0	-	No direct effect	1	-0.1	-0.1	0.1	0	-1	-1	0	0
0.45	0	+	No direct effect	1	1.3	1.1	-0.1	-0.2	1.4	1.4	2.2	2
0.08	0	-	No direct effect	2%	-0.3	-0.3	-0.7	-0.9	-4.6	-4.1	-6.2	-6.2
0.29	0	-	No direct effect	2%	-1.4	-1.2	-1.1	-1.1	-3.5	-3.8	**-22.2**	**-22.9**
0.25	0	+	No direct effect	10%	0.2	0.4	1.4	1.6	**36.4**	**37.5**	**151.7**	**155.6**
0.39	0	-	No direct effect	10%	0.4	0.5	-2	-1.8	**-39.7**	**-38.4**	**-158.3**	**-153.4**
0.08	0	+	No direct effect	100%	3.6	2.9	**93**	**90.8**	**2293.3**	**2222.2**	**1448.9**	**1400.8**
0.29	0	+	No direct effect	100%	4.5	3.9	**282.1**	**258.3**	**6887.3**	**6436.9**	**4335.4**	**4047.1**
0.05	0	-	Biased inter. sign	1	0.4	0.5	-0.9	-1.2	-2.3	-2.4	-1.5	-1.6
0.1	0	-	Biased inter. sign	1	0	0.1	-0.1	-0.2	0.1	0.1	-0.3	-0.2
0.08	0	+	Biased inter. sign	2%	0.2	0.2	0.3	0.4	4	4.5	**9.6**	**10.3**
0.08	0	+	Biased inter. sign	2%	0.4	0.4	0.8	0.6	0.5	0.6	1.8	1.8
0.14	0	+	Biased inter. sign	10%	-0.2	-0.1	2.8	3.4	**36.4**	**35.6**	**150.3**	**141.4**
0.3	0	+	Biased inter. sign	10%	0.8	0.7	1.5	1.3	**39.1**	**26.6**	**199**	**134.3**

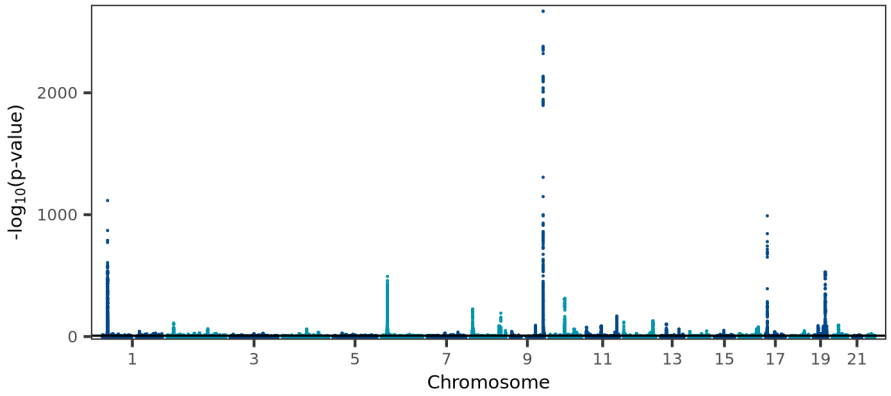
Table C.8: Log-transformed p-values with sign of corresponding coefficient obtained from SNP × PGS interaction testing for simulated traits with $\alpha = -0.5$ and QN residuals.

MAF	Coeff.	Sign	Type	# SNPs affected	*Trait 1* Effect: 1%		*Trait 2* Effect: 10%		*Trait 3* Effect: 50%		*Trait 4* Effect: 100%	
					LOCO	Full	LOCO	Full	LOCO	Full	LOCO	Full
0.38	0.7	-	Direct effect	1	0	-0.1	0.6	0.6	-1.7	-1.7	-1.8	-2
0.42	-0.1	+	Direct effect	1	0.2	0.3	0.3	0.3	0.9	1	0	0.1
0.28	0.84	-	Direct effect	2%	0.2	0.3	-0.5	-0.6	-3.1	-3.1	**-11.5**	**-11.3**
0.46	1.09	-	Direct effect	2%	0.2	0.1	-0.5	-0.3	-4.8	-4.6	**-8.6**	**-8.3**
0.13	-1.04	+	Direct effect	10%	1.5	1.3	3.6	3.6	**46.5**	**48**	**199**	**201**
0.34	-1.86	+	Direct effect	10%	-0.2	-0.1	2.8	3.2	**32.4**	**34**	**156.7**	**157.8**
0.12	-0.18	-	Direct effect	100%	-4.1	-3.2	**-161.1**	**-152.8**	**-3651.2**	**-3434.4**	**-2220.7**	**-2099.1**
0.42	-0.01	+	Direct effect	100%	5	4.9	**358.9**	**351.7**	**8104.4**	**7856.3**	**5108.2**	**4973.7**
0.11	0	-	No direct effect	1	-0.2	-0.2	0.5	0.4	-0.5	-0.5	-0.1	0
0.45	0	+	No direct effect	1	0.2	0.2	-0.6	-0.7	1.4	1.5	1.1	1.2
0.08	0	-	No direct effect	2%	-0.1	-0.1	-0.3	-0.2	-4.1	-3.8	**-7.7**	-7.1
0.29	0	-	No direct effect	2%	-0.4	-0.4	-1.2	-1.3	-3.1	-3.3	**-15.2**	**-16.3**
0.25	0	+	No direct effect	10%	0	0.1	1.7	1.9	**36.3**	**37.2**	**143.2**	**147**
0.39	0	-	No direct effect	10%	-0.1	-0.2	-4.1	-4.7	**-45.5**	**-42.1**	**-168.7**	**-159.8**
0.08	0	+	No direct effect	100%	3.7	3.5	**120.7**	**115.9**	**2172.1**	**2094.4**	**1401.9**	**1355**
0.29	0	+	No direct effect	100%	4.5	3.5	**299.4**	**289.9**	**6441.1**	**6050.5**	**4067.1**	**3833.2**
0.05	0	-	Biased inter. sign	1	-0.2	-0.2	0.1	0	-0.8	-1.1	-0.8	-0.9
0.1	0	-	Biased inter. sign	1	-0.6	-0.6	-0.4	-0.3	0.3	0.3	-0.2	-0.2
0.08	0	+	Biased inter. sign	2%	0	0.2	0.7	0.6	2.1	2.2	6	6.6
0.08	0	+	Biased inter. sign	2%	-0.4	-0.4	-0.1	-0.1	1.1	1.3	2.5	2.5
0.14	0	+	Biased inter. sign	10%	0.4	0.4	0.8	0.7	**34.1**	**32.3**	**158**	**149.6**
0.3	0	+	Biased inter. sign	10%	-0.7	-0.6	1.7	1.3	**40.5**	**32.4**	**194.4**	**148.4**

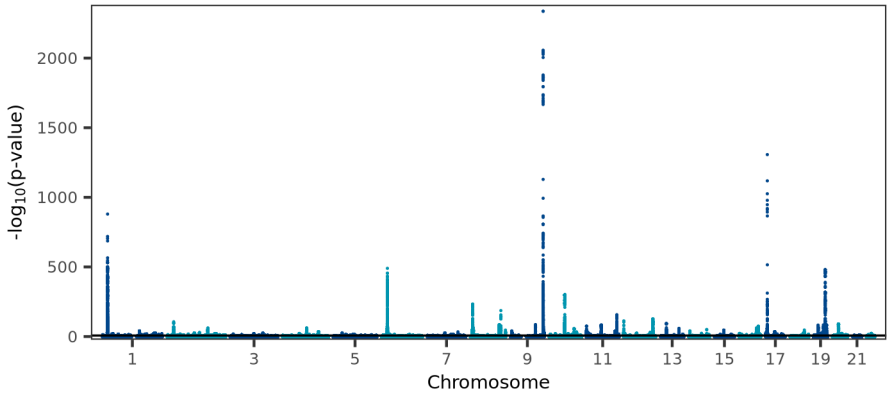
Table C.9: Log-transformed p-values with sign of corresponding coefficient obtained from SNP × PGS interaction testing for simulated traits with $\alpha = -0.5$ and Winsorised residuals.

MAF	Coeff.	Sign	Type	# SNPs affected	*Trait 1* Effect: 1%		*Trait 2* Effect: 10%		*Trait 3* Effect: 50%		*Trait 4* Effect: 100%	
					LOCO	Full	LOCO	Full	LOCO	Full	LOCO	Full
0.38	0.7	-	Direct effect	1	-0.3	-0.4	0.4	0.3	-1.4	-1.4	-1.9	-2.1
0.42	-0.1	+	Direct effect	1	0.1	0.2	0.2	0.2	0.8	0.9	0	0.1
0.28	0.84	-	Direct effect	2%	0.4	0.4	-0.3	-0.3	-3.2	-3	**-10.3**	**-10.4**
0.46	1.09	-	Direct effect	2%	0.2	0.1	-0.3	-0.2	-5.8	-5.7	**-9.3**	**-9.1**
0.13	-1.04	+	Direct effect	10%	1.3	1	2.8	2.8	**48.6**	**50.2**	**204.8**	**206.4**
0.34	-1.86	+	Direct effect	10%	-0.2	-0.1	2.6	2.9	**36.4**	**37.3**	**166.5**	**166.7**
0.12	-0.18	-	Direct effect	100%	-3.1	-2.4	**-137.9**	**-130.9**	**-3548.4**	**-3344.1**	**-2307.7**	**-2180.4**
0.42	-0.01	+	Direct effect	100%	4.1	4	**294.4**	**290**	**8251.9**	**8019**	**5320.1**	**5176.1**
0.11	0	-	No direct effect	1	-0.1	-0.2	0.4	0.3	-0.3	-0.3	-0.1	0
0.45	0	+	No direct effect	1	0.2	0.3	-0.8	-0.8	2.3	2.3	1.5	1.6
0.08	0	-	No direct effect	2%	0	0	-0.3	-0.2	-3.9	-3.6	-6.5	-6
0.29	0	-	No direct effect	2%	-0.5	-0.4	-1.5	-1.6	-3	-3.2	**-15.7**	**-16.4**
0.25	0	+	No direct effect	10%	0.3	0.5	1.8	1.9	**35**	**36**	**143.8**	**148.1**
0.39	0	-	No direct effect	10%	0	-0.1	-3.3	-3.8	**-46**	**-42.6**	**-179.8**	**-168.7**
0.08	0	+	No direct effect	100%	3.8	3.8	**100.8**	**97.2**	**2164.3**	**2085.3**	**1420.2**	**1372.1**
0.29	0	+	No direct effect	100%	4.4	3.3	**247**	**239.5**	**6625**	**6248.2**	**4239.1**	**3989.8**
0.05	0	-	Biased inter. sign	1	-0.4	-0.4	0	0	-0.7	-1	-0.6	-0.8
0.1	0	-	Biased inter. sign	1	-0.5	-0.5	-0.5	-0.3	0.1	0.1	-0.3	-0.3
0.08	0	+	Biased inter. sign	2%	-0.1	0.1	1	0.8	2.3	2.1	6.6	7.1
0.08	0	+	Biased inter. sign	2%	-0.5	-0.4	-0.3	-0.3	0.8	0.9	2.2	2.2
0.14	0	+	Biased inter. sign	10%	0.3	0.3	0.4	0.3	**35.1**	**33.5**	**161.3**	**153.5**
0.3	0	+	Biased inter. sign	10%	-1	-1.1	1.8	1.3	**40.4**	**33.2**	**203**	**154.7**

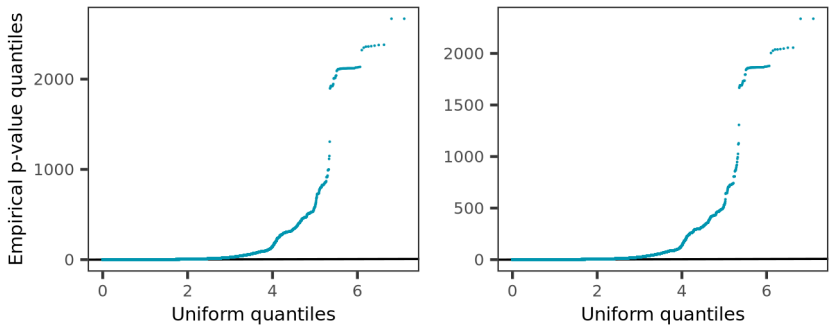
D Alkaline phosphatase: supplementary results



(a) QN residuals



(b) Winsorised residuals



(c) QN residuals (d) Winsorised residuals

Figure D.1: Manhattan and QQ plots of results from original GWAS: ALP with QN or Winsorised residuals.

E List of independent SNP × PGS interaction hits

Table E.1: Independent 'LHS' hits with signed log-transformed p-values from SNP × PGS GWASs with QN residuals for all traits. *(continued on the next six pages)*

The ID column shows a 'positional' identifier for each SNP (CHR:POS:REF:ALT), the A1 column the minor allele whose effect is measured in the GWAS and the Consequence and Gene columns functional information taken from the VEP dataset. The p-values shown in the final two columns are log-transformed and have the same sign as the corresponding coefficient.

Trait	rsID	ID	A1	MAF	Consequence	Gene	SNP × PGS	Orig.
Waist circumference	rs6691427	1:203510048:G:C	C	0.489	Intergenic var.	–	7.6	1.2
Waist circumference	rs35608615	6:27040402:G:C	C	0.136	Intergenic var.	–	9.7	0.3
Hip circumference	rs13219354	6:27185664:T:C	C	0.137	Intergenic var.	–	8.2	5.7
Waist-to-hip ratio	rs5835988	2:165501927:T:TG	TG	0.406	Intergenic var.	–	8.0	-22.2
BMI	rs112132062	3:131655987:A:C	C	0.035	Intron var.	*CPNE4*	7.3	0.7
BMI	rs13191326	6:27031029:A:G	G	0.136	Upstream gene var.	*VN1R13P*	11.1	2.1
Body fat %	rs34927463	1:219652467:T:TG	T	0.332	Regulatory region var.	–	9.7	22.5
Body fat %	rs35526527	6:27042287:C:A	A	0.136	Intergenic var.	–	7.9	4.1
Body fat %	rs56094641	16:53806453:A:G	G	0.404	Intron var.	*FTO*	8.1	97.9
Whole body fat mass	rs67457459	6:27198343:T:G	G	0.137	Intergenic var.	–	8.7	4.4
Whole body fat mass	rs35924649	17:73181927:G:GT	GT	0.050	Upstream gene var.	*SUMO2*	7.5	1.1
Whole body fat-free mass	rs72934653	6:97654656:A:G	G	0.010	Intron var.	*MMS22L*	7.7	-0.3
Whole body water mass	rs72934653	6:97654656:A:G	G	0.010	Intron var.	*MMS22L*	7.4	-0.3
Basal metabolic rate	rs59373875	6:149509343:T:G	G	0.075	Intergenic var.	–	7.7	1.3
Leg fat %	rs34927463	1:219652467:T:TG	T	0.332	Regulatory region var.	–	8.7	11.8
Leg fat %	rs10195252	2:165513091:T:C	C	0.407	Intron var.	*COBLL1*	7.4	15.3
Leg fat mass	rs779470261	3:131646163:CTCTAA:C	C	0.263	Intron var.	*CPNE4*	7.8	12.0
Leg fat mass	rs67457459	6:27198343:T:G	G	0.137	Intergenic var.	–	9.7	3.3

Table E.1: *(continued from the previous page)*

Trait	rsID	ID	A1	MAF	Consequence	Gene	SNP × PGS	Orig.
Arm fat %	rs13191326	6:27031029:A:G	G	0.136	Upstream gene var.	*VN1R13P*	9.8	2.2
Arm fat %	rs1421085	16:53800954:T:C	C	0.402	Intron var.	*FTO*	8.5	120.7
Arm fat mass	rs6801798	3:131589085:T:G	G	0.278	Intron var.	*CPNE4*	7.6	14.2
Arm fat mass	rs35608615	6:27040402:G:C	C	0.136	Intergenic var.	–	10.6	2.6
Arm fat mass	rs1421085	16:53800954:T:C	C	0.402	Intron var.	*FTO*	7.8	152.4
Arm predicted mass	rs72934653	6:97654656:A:G	G	0.010	Intron var.	*MMS22L*	7.6	-0.2
Trunk fat %	rs2605111	1:219656650:C:G	C	0.328	Intergenic var.	–	8.5	29.4
Trunk fat mass	rs13212921	6:27205422:C:T	T	0.137	Intergenic var.	–	8.0	5.0
Pulse rate	rs116372144	5:137287666:G:A	A	0.127	Intron var.	*FAM13B*	7.5	-19.6
Pulse rate	rs236352	6:36817113:A:G	A	0.339	Intergenic var.	–	8.2	-21.0
Pulse rate	rs11047527	12:24762501:A:C	C	0.151	Intergenic var.	–	16.6	-84.2
Diastolic blood pressure	rs12321528	12:642276:A:G	G	0.007	Intron var.	*B4GALNT3*	-7.6	0.0
White blood cell count	rs145164263	18:38541168:G:A	A	0.012	Intergenic var.	–	7.8	0.2
Red blood cell count	rs1435547	2:11738091:T:C	T	0.002	Synonymous var.	*GREB1*	7.4	0.2
Mean corpuscular haemoglobin	rs570013781	16:149539:G:A	A	0.002	Intron var.	*NPRL3*	14.8	-172.8
Red blood cell distribution width	rs3811444	1:248039451:C:T	T	0.334	Missense var.	*TRIM58*	18.2	-169.9
Red blood cell distribution width	rs3812049	5:127418850:C:G	C	0.259	Noncoding transcript exon var.	*CTC-228N24.3*	13.7	-342.2
Red blood cell distribution width	rs2596546	6:31329386:A:G	G	0.323	Upstream gene var.	*HLA-B*	7.4	11.0
Platelet count	rs35350651	12:111907431:A:AC	A	0.496	Intron var.	*ATXN2*	7.9	315.0
Platelet crit	rs7310615	12:111865049:C:G	C	0.482	Intron var.	*SH2B3*	11.2	447.1
Mean platelet volume	rs1354034	3:56849749:T:C	T	0.399	Intron var.	*ARHGEF3*	-29.1	1862.8

Table E.1: *(continued from the previous page)*

Trait	rsID	ID	A1	MAF	Consequence	Gene	SNP × PGS	Orig.
Mean platelet volume	rs415064	20:57597971:G:C	C	0.047	Missense var.	*TUBB1*	12.8	-91.4
Platelet distribution width	rs1354034	3:56849749:T:C	T	0.399	Intron var.	*ARHGEF3*	21.1	-0.1
Platelet distribution width	rs4841097	8:8946163:G:A	G	0.222	Intron var.	*ERI1*	-9.0	122.4
Platelet distribution width	rs55966801	11:200343:T:C	C	0.231	Intron var.	*BET1L*	-14.4	19.0
Platelet distribution width	–	17:27833678:AGATT:A	A	0.450	Intron var.	*TAOK1*	-9.7	-2.4
Lymphocyte count	rs200932087	9:77077999:T:TA	TA	0.005	Intergenic var.	–	-7.5	1.3
Eosinophil count	rs9271681	6:32592918:C:T	T	0.321	Upstream gene var.	*HLA-DQA1*	-9.5	-36.5
Eosinophil count	rs2024049	7:75496097:C:T	C	0.499	Intron var.	*RHBDD2*	-10.2	2.7
Eosinophil count	rs928413	9:6213387:G:A	G	0.259	Upstream gene var.	*IL33*	9.3	247.5
Monocyte %	rs6429432	1:236107241:A:C	A	0.108	Upstream gene var.	*RP5-940F7.2*	-10.5	33.2
Neutrophil %	rs115866005	4:27925743:T:C	C	0.017	Intergenic var.	–	-7.4	0.4
Eosinophil %	rs2239704	6:31540141:A:C	A	0.383	5'-UTR var.	*LTA*	11.5	-5.6
Eosinophil %	rs2024050	7:75460393:A:G	A	0.104	Upstream gene var.	*AC005102.1*	-9.1	49.1
Eosinophil %	rs2381416	9:6193455:C:A	C	0.264	Upstream gene var.	*RP11-575C20.1*	10.4	258.1
Eosinophil %	rs34108746	11:57146225:A:G	G	0.078	Missense var.	*PRG3*	15.4	-40.7
Eosinophil %	rs4766897	12:112179471:T:C	T	0.340	Intron var.	*ACAD10*	-7.4	107.1
Reticulocyte count	rs192425119	6:103929691:G:A	A	0.001	Intergenic var.	–	8.2	-0.6
Alkaline phosphatase	rs77569513	1:21449335:G:A	A	0.009	Intron var.	*EIF4G3*	12.4	-3.3
Alkaline phosphatase	rs10465542	1:21859507:A:G	G	0.222	Intron var.	*ALPL*	-22.8	-27.7
Alkaline phosphatase	rs149356924	1:21875506:T:C	C	0.004	Intron var.	*ALPL*	15.7	-46.4
Alkaline phosphatase	rs573303250	1:22138578:G:A	A	0.001	Upstream gene var.	*LDLRAD2*	10.8	-19.1

Table E.1: *(continued from the previous page)*

Trait	rsID	ID	A1	MAF	Consequence	Gene	SNP × PGS	Orig.
Alkaline phosphatase	rs145223095	1:22361255:C:T	T	0.008	Downstream gene var.	*LINC00339*	14.4	-28.5
Alkaline phosphatase	rs635634	9:136155000:C:T	T	0.186	Upstream gene var.	*ABO*	49.1	-2348.5
Alanine aminotransferase	rs13141441	4:88212722:G:A	A	0.284	Regulatory region var.	–	-16.4	-52.1
Alanine aminotransferase	rs756350040	19:19370340:TGACA:T	T	0.074	Intron var.	*HAPLN4*	13.1	79.6
Alanine aminotransferase	rs429358	19:45411941:T:C	C	0.156	Missense var.	*APOE*	-8.8	-40.7
Alanine aminotransferase	rs738409	22:44324727:C:G	G	0.216	Missense var.	*PNPLA3*	59.6	274.1
Apoliprotein A	rs9274474	6:32633683:C:T	T	0.257	Intron var.	*HLA-DQB1*	11.1	-4.0
Aspartate aminotransferase	rs1150755	6:32038550:C:T	T	0.174	Intron var.	*TNXB*	10.2	-79.8
Aspartate aminotransferase	rs1926736	10:17891705:A:G	A	0.380	Missense var.	*MRC1L1*	11.6	-649.9
Aspartate aminotransferase	–	10:101101889:GT:G	G	0.002	Intron var.	*CNNM1*	35.1	-84.2
Aspartate aminotransferase	rs756350040	19:19370340:TGACA:T	T	0.074	Intron var.	*HAPLN4*	9.5	42.4
Aspartate aminotransferase	rs3747207	22:44324855:G:A	A	0.215	Intron var.	*PNPLA3*	21.2	277.0
Direct bilirubin	rs34983651	2:234668879:C:CAT	CAT	0.312	Intron var.	*UGT1A6*	-195.7	12889.0
Direct bilirubin	rs78833965	6:25804921:T:A	A	0.114	Intron var.	*SLC17A1*	-8.2	28.1
Direct bilirubin	rs72805692	10:71099109:A:G	G	0.116	Intron var.	*HK1*	-10.8	60.4
Direct bilirubin	rs10841648	12:20954557:C:A	C	0.339	Intergenic var.	–	-19.1	112.4
Direct bilirubin	rs7132068	12:21099892:G:C	G	0.252	Intron var.	*LST3*	-25.0	176.4
Direct bilirubin	rs11045630	12:21125521:T:C	C	0.198	Intron var.	*LST3*	-18.6	130.0
Direct bilirubin	rs11519275	12:21363713:G:C	C	0.213	Intron var.	*SLCO1B1*	-17.4	205.4
Direct bilirubin	rs57743625	12:21367633:G:A	A	0.159	Intron var.	*SLCO1B1*	-42.5	426.1
Direct bilirubin	rs4149074	12:21370682:C:T	T	0.349	Intron var.	*SLCO1B1*	-15.7	127.2

Table E.1: *(continued from the previous page)*

Trait	rsID	ID	A1	MAF	Consequence	Gene	SNP × PGS	Orig.
Cholesterol	rs2954018	8:126477153:C:T	C	0.298	Upstream gene var.	*RP11-136O12.2*	7.9	39.3
Cholesterol	–	11:116623213:TA:T	TA	0.139	Intron var.	*BUD13*	10.6	79.9
Cholesterol	rs747875948	19:11189298:CG:C	C	0.114	Intergenic var.	–	-20.7	-391.3
Cholesterol	rs814573	19:45424351:A:T	T	0.188	Downstream gene var.	*APOC1*	9.3	453.8
Cystatin C	rs144750909	18:49054586:G:A	A	0.010	Intron var.	*RP11-267C16.1*	7.4	0.0
Cystatin C	rs185932638	20:21931106:T:C	C	0.002	Intron var.	*RP5-1185K9.1*	7.7	-0.2
Cystatin C	rs2405367	20:23622880:G:A	A	0.216	Upstream gene var.	*CST3*	28.7	-4056.8
Cystatin C	rs139450906	20:23626256:C:T	T	0.017	Intergenic var.	–	8.0	-4.9
Cystatin C	rs77149469	20:24100352:C:T	T	0.002	Regulatory region var.	–	13.2	-31.5
Glycated haemoglobin	rs560887	2:169763148:T:C	T	0.299	Intron var.	*G6PC2*	8.5	-351.3
Glycated haemoglobin	rs1799945	6:26091179:C:G	G	0.151	Missense var.	*HFE*	10.7	-122.2
Glycated haemoglobin	rs1800562	6:26093141:G:A	A	0.078	Missense var.	*HFE*	8.7	-172.6
Glycated haemoglobin	rs35371668	6:32561638:C:T	T	0.182	Upstream gene var.	*HLA-DRB1*	13.5	-1.2
Glycated haemoglobin	rs34265667	8:41542093:G:A	A	0.035	Synonymous var.	*ANK1*	14.6	-114.2
Glycated haemoglobin	rs4737010	8:41630447:G:A	A	0.228	Intron var.	*ANK1*	-12.4	169.9
Glycated haemoglobin	–	9:91392686:TTC:T	T	0.054	Intergenic var.	–	7.6	-94.5
Glycated haemoglobin	rs17476364	10:71094504:T:C	C	0.110	Intron var.	*HK1*	38.9	-1615.1
Glycated haemoglobin	rs7903146	10:114758349:C:T	T	0.290	Intron var.	*TCF7L2*	30.4	180.7
Glycated haemoglobin	rs3184504	12:111884608:T:C	T	0.482	Missense var.	*SH2B3*	7.4	-29.7
Glycated haemoglobin	rs17817288	16:53807764:A:G	G	0.482	Intron var.	*FTO*	7.7	12.4
Glycated haemoglobin	rs2748424	17:76124865:C:G	G	0.190	Intron var.	*TMC6*	-9.0	182.9

Table E.1: *(continued from the previous page)*

Trait	rsID	ID	A1	MAF	Consequence	Gene	SNP × PGS	Orig.
Glycated haemoglobin	rs59616136	19:17252041:G:A	A	0.093	Intron var.	*MYO9B*	7.7	-75.8
Glycated haemoglobin	rs6000553	22:37469192:A:G	A	0.468	Intron var.	*TMPRSS6*	-11.0	135.1
HDL cholesterol	rs2740488	9:107661742:A:C	C	0.264	Intron var.	*ABCA1*	-10.6	-151.4
HDL cholesterol	rs191986139	11:39515079:T:C	C	0.005	Intergenic var.	–	-7.9	-0.2
HDL cholesterol	rs56228609	16:56987765:C:T	T	0.316	Upstream gene var.	*AC012181.1*	9.6	1813.0
HDL cholesterol	rs72836561	17:41926126:C:T	T	0.032	Missense var.	*CD300LG*	7.9	-139.1
HDL cholesterol	rs7412	19:45412079:C:T	T	0.081	Missense var.	*APOE*	12.3	60.7
IGF-1	rs1065656	16:1838836:C:G	C	0.302	3'-UTR var.	*NUBP2*	-15.2	-77.9
LDL direct	rs142158911	19:11190534:G:A	A	0.116	Intergenic var.	–	-12.3	-477.8
Lipoprotein A	rs73014164	6:160976620:C:T	T	0.187	Intron var.	*LPA*	-10.8	-363.1
Lipoprotein A	rs6938647	6:160986915:A:C	A	0.218	Intron var.	*LPA*	-15.1	-4452.1
Lipoprotein A	rs571848809	6:161005388:G:A	A	0.107	Intron var.	*LPA*	11.7	-520.2
Phosphate	rs4713663	6:33674197:G:A	A	0.429	Intron var.	*UQCC2*	-8.8	22.3
SHBG	rs191712309	5:65173182:C:T	T	0.010	Intergenic var.	–	-8.0	-0.7
SHBG	rs117986105	12:51064266:G:A	A	0.004	Intron var.	*DIP2B*	-7.6	-0.1
SHBG	rs62059839	17:7533015:C:T	T	0.261	Intron var.	*SHBG*	13.9	1499.5
SHBG	rs9282846	17:7533742:C:T	T	0.008	Synonymous var.	*SHBG*	-17.0	-107.6
SHBG	rs6258	17:7534678:C:T	T	0.007	Missense var.	*SHBG*	-10.2	-1306.2
Total bilirubin	rs1976391	2:234665983:A:G	G	0.315	Intron var.	*UGT1A6*	-201.3	21268.1
Total bilirubin	rs78273613	6:25866075:A:G	G	0.103	Intron var.	*SLC17A3*	-8.9	33.3
Total bilirubin	rs72805692	10:71099109:A:G	G	0.116	Intron var.	*HK1*	-13.5	98.2

Table E.1: *(continued from the previous page)*

Trait	rsID	ID	A1	MAF	Consequence	Gene	SNP × PGS	Orig.
Total bilirubin	rs10841648	12:20954557:C:A	C	0.339	Intergenic var.	–	-20.9	173.7
Total bilirubin	rs12371715	12:21067724:T:G	G	0.199	Intron var.	*SLCO1B3*	-11.0	125.9
Total bilirubin	rs12814270	12:21097329:A:G	G	0.221	Intron var.	*LST3*	-11.9	158.5
Total bilirubin	rs7132068	12:21099892:G:C	G	0.252	Intron var.	*LST3*	-18.6	194.8
Total bilirubin	rs11045816	12:21323272:A:G	G	0.165	Intron var.	*SLCO1B1*	10.0	-93.3
Total bilirubin	–	12:21344313:GA:G	G	0.239	Intron var.	*SLCO1B1*	24.4	-203.3
Total bilirubin	rs73067080	12:21388829:T:C	C	0.157	Intron var.	*SLCO1B1*	-54.8	603.9
Total bilirubin	rs855791	22:37462936:A:G	A	0.440	Missense var.	*TMPRSS6*	8.7	-30.8
Testosterone	rs727428	17:7537792:T:C	T	0.444	Downstream gene var.	*SHBG*	-8.5	-385.9
Triglycerides	–	11:116623213:TA:T	TA	0.139	Intron var.	*BUD13*	17.6	1134.6
Triglycerides	–	15:58679807:CAGA:C	C	0.357	Intron var.	*ALDH1A2*	-9.6	30.5
Triglycerides	rs2070895	15:58723939:G:A	A	0.219	Intron var.	*LIPC*	-9.4	45.0
Triglycerides	rs738408	22:44324730:C:T	T	0.216	Synonymous var.	*PNPLA3*	-8.0	-2.5
Urate	rs7696092	4:10025320:C:A	C	0.214	Intron var.	*SLC2A9*	14.9	-2899.7
Vitamin D	–	4:72610152:GGAGACAGAGC:G	G	0.292	Intron var.	*GC*	-8.1	-1101.1
Vitamin D	rs549227613	11:14009148:G:A	A	0.002	Intron var.	*SPON1*	15.2	16.9
Vitamin D	rs1037378	11:14852490:G:A	G	0.460	Intron var.	*PDE3B*	91.8	382.6
Vitamin D	rs117913124	11:14900931:G:A	A	0.028	Synonymous var.	*CYP2R1*	-54.0	-553.6
Vitamin D	rs10500804	11:14910273:T:G	G	0.422	Intron var.	*CYP2R1*	-88.8	-272.5
Ankle spacing width	rs540963451	1:31563102:G:A	A	0.006	Downstream gene var.	*SEPW1P*	-7.6	-0.1
Ankle spacing width	rs7746082	6:106435269:G:C	C	0.286	Intergenic var.	–	-8.3	-0.5

F List of independent pairwise interaction hits

Table F.1: Independent pairwise interaction hits with LOCO PGS and QN residuals for all traits. *(continued on the next six pages)*

The Beta and P-value columns the coefficient and associated log-transformed p-value, respectively, of the interaction term of the two SNPs in a regression in which the main effects (and the squares of their allele counts) are also present.

Trait	LHS SNP					RHS SNP					Beta	P-value
	rsID	Chr.	MAF	Consequence	Gene	rsID	Chr.	MAF	Consequence	Gene		
Hip circumference	rs13219354	6	0.137	Intergenic var.	–	rs999728	7	0.009	Intron var.	*ZNF467*	0.16	7.31
BMI	rs112132062	3	0.035	Intron var.	*CPNE4*	10:56260995:AT:A	10	0.035	Intron var.	*PCDH15*	0.16	7.70
Whole body fat-free mass	rs72934653	6	0.010	Intron var.	*MMS22L*	rs16848445	2	0.043	Downstream gene var.	*AC092578.1*	0.23	7.45
Whole body water mass	rs72934653	6	0.010	Intron var.	*MMS22L*	rs16848445	2	0.043	Downstream gene var.	*AC092578.1*	0.23	7.49
Leg fat %	rs10195252	2	0.407	Intron var.	*COBLL1*	rs566108812	9	0.003	Intron var.	*PTPRD*	0.21	7.60
Leg fat mass	rs67457459	6	0.137	Intergenic var.	–	rs540144621	3	0.002	Intron var.	*CAPN7*	0.30	7.40
Arm fat mass	rs35608615	6	0.136	Intergenic var.	–	rs146511665	11	0.017	Intergenic var.	–	0.11	7.46
Red blood cell count	rs1435547	2	0.002	Synonymous var.	*GREB1*	rs9559732	13	0.191	Intergenic var.	–	-0.31	8.42
Red blood cell count	rs1435547	2	0.002	Synonymous var.	*GREB1*	rs7257522	19	0.381	Intron var.	*CYP4F11*	-0.24	7.47
Mean corp. haemoglobin	rs570013781	16	0.002	Intron var.	*NPRL3*	rs12940324	17	0.103	Intron var.	*AC061992.2*	0.43	10.02
Red blood cell dist. width	rs3811444	1	0.334	Missense var.	*TRIM58*	rs75349470	14	0.012	Intron var.	*MDGA2*	-0.10	8.34
Red blood cell dist. width	rs3812049	5	0.259	Noncod. transc. exon var.	*CTC-228N24.3*	rs184736239	13	0.022	Intergenic var.	–	-0.08	7.34
Red blood cell dist. width	rs2596546	6	0.323	Upstream gene var.	*HLA-B*	rs561857559	19	0.002	Intergenic var.	–	-0.30	9.09
Mean platelet volume	rs1354034	3	0.399	Intron var.	*ARHGEF3*	rs78909033	2	0.135	Intron var.	*RNPEPL1*	0.03	10.47
Mean platelet volume	rs1354034	3	0.399	Intron var.	*ARHGEF3*	rs34592828	5	0.046	Missense var.	*IQGAP2*	-0.05	7.77
Mean platelet volume	rs1354034	3	0.399	Intron var.	*ARHGEF3*	rs11615667	12	0.117	Intron var.	*WDR66*	-0.06	24.65
Mean platelet volume	rs1354034	3	0.399	Intron var.	*ARHGEF3*	rs71329093	21	0.054	Intron var.	*RUNX1*	0.05	10.16
Platelet dist. width	rs1354034	3	0.399	Intron var.	*ARHGEF3*	rs79920678	6	0.019	Upstream gene var.	*RP5-1086L22.1*	0.07	7.55
Platelet dist. width	rs1354034	3	0.399	Intron var.	*ARHGEF3*	rs342263	7	0.445	Intron var.	*CTB-111H14.1*	0.03	11.80

Table F.1: *(continued from the previous page)*

Trait	LHS SNP					RHS SNP					Beta	P-value
	rsID	Chr.	MAF	Consequence	Gene	rsID	Chr.	MAF	Consequence	Gene		
Platelet dist. width	rs1354034	3	0.399	Intron var.	*ARHGEF3*	rs11604127	11	0.232	5'-UTR var.	*ODF3*	-0.05	28.18
Platelet dist. width	rs1354034	3	0.399	Intron var.	*ARHGEF3*	rs12817046	12	0.113	Intron var.	*WDR66*	-0.03	7.76
Platelet dist. width	rs55966801	11	0.231	Intron var.	*BET1L*	rs1354034	3	0.399	Intron var.	*ARHGEF3*	-0.05	28.10
Eosinophil count	rs928413	9	0.259	Upstream gene var.	*IL33*	rs34210653	17	0.021	Missense var.	*ALOX15*	-0.11	13.70
Eosinophil %	rs2381416	9	0.264	Upstream gene var.	*RP11-575C20.1*	rs34210653	17	0.021	Missense var.	*ALOX15*	-0.09	11.22
Eosinophil %	rs34108746	11	0.078	Missense var.	*PRG3*	rs74116771	1	0.033	Intergenic var.	–	-0.10	7.87
Alkaline phosphatase	rs77569513	1	0.009	Intron var.	*EIF4G3*	rs144406757	3	0.012	Intron var.	*NAALADL2*	-0.56	7.86
Alkaline phosphatase	rs77569513	1	0.009	Intron var.	*EIF4G3*	rs182976181	8	0.007	Intron var.	*UNC5D*	-0.64	7.64
Alkaline phosphatase	rs77569513	1	0.009	Intron var.	*EIF4G3*	rs7329258	13	0.403	Intron var.	*ENOX1*	0.12	7.87
Alkaline phosphatase	rs77569513	1	0.009	Intron var.	*EIF4G3*	rs1801248	13	0.042	Synonymous var.	*ATP7B*	-0.28	7.49
Alkaline phosphatase	rs77569513	1	0.009	Intron var.	*EIF4G3*	rs3918095	14	0.022	Intron var.	*CCNK*	-0.41	8.00
Alkaline phosphatase	rs149356924	1	0.004	Intron var.	*ALPL*	rs12636326	3	0.260	Regulatory region var.	–	0.19	7.60
Alkaline phosphatase	rs149356924	1	0.004	Intron var.	*ALPL*	rs62297580	4	0.065	Intron var.	*FIP1L1*	-0.34	7.68
Alkaline phosphatase	rs149356924	1	0.004	Intron var.	*ALPL*	rs7684999	4	0.057	Regulatory region var.	–	-0.35	7.55
Alkaline phosphatase	rs149356924	1	0.004	Intron var.	*ALPL*	rs17123355	12	0.062	Intergenic var.	–	0.36	8.60
Alkaline phosphatase	rs149356924	1	0.004	Intron var.	*ALPL*	rs149786005	13	0.201	Intron var.	*SGCG*	-0.20	7.72
Alkaline phosphatase	rs149356924	1	0.004	Intron var.	*ALPL*	rs147817132	13	0.034	Intergenic var.	–	-0.45	7.35
Alkaline phosphatase	rs149356924	1	0.004	Intron var.	*ALPL*	rs8065491	17	0.450	Intergenic var.	–	-0.17	7.95
Alkaline phosphatase	rs573303250	1	0.001	Upstream gene var.	*LDLRAD2*	rs76571364	5	0.057	Intergenic var.	–	-0.70	9.08
Alkaline phosphatase	rs573303250	1	0.001	Upstream gene var.	*LDLRAD2*	rs9297078	6	0.441	Intergenic var.	–	-0.29	7.43
Alkaline phosphatase	rs573303250	1	0.001	Upstream gene var.	*LDLRAD2*	rs13229792	7	0.135	Intron var.	*DOCK4*	-0.46	8.86

Table F.1: *(continued from the previous page)*

Trait	LHS SNP					RHS SNP					Beta	P-value
	rsID	Chr.	MAF	Consequence	Gene	rsID	Chr.	MAF	Consequence	Gene		
Alkaline phosphatase	rs573303250	1	0.001	Upstream gene var.	*LDLRAD2*	rs34819186	8	0.473	Intron var.	*TSNARE1*	0.30	7.83
Alkaline phosphatase	rs573303250	1	0.001	Upstream gene var.	*LDLRAD2*	rs4745057	9	0.251	Intron var.	*TRPM3*	-0.34	8.38
Alkaline phosphatase	rs573303250	1	0.001	Upstream gene var.	*LDLRAD2*	rs7043546	9	0.475	Intron var.	*GLDC*	-0.31	8.14
Alkaline phosphatase	rs573303250	1	0.001	Upstream gene var.	*LDLRAD2*	rs2796866	13	0.331	Intergenic var.	–	-0.35	9.96
Alkaline phosphatase	rs573303250	1	0.001	Upstream gene var.	*LDLRAD2*	rs9540399	13	0.132	Intergenic var.	–	-0.48	9.43
Alkaline phosphatase	rs573303250	1	0.001	Upstream gene var.	*LDLRAD2*	rs12884672	14	0.282	Intron var.	*RP11-1152H15.1*	0.33	7.58
Alkaline phosphatase	rs573303250	1	0.001	Upstream gene var.	*LDLRAD2*	15:76145831:TTA:T	15	0.036	Intron var.	*UBE2Q2*	-0.85	8.04
Alkaline phosphatase	rs573303250	1	0.001	Upstream gene var.	*LDLRAD2*	rs1634419	17	0.366	Upstream gene var.	*SLC5A10*	0.33	9.04
Alkaline phosphatase	rs573303250	1	0.001	Upstream gene var.	*LDLRAD2*	rs72891768	18	0.044	Intron var.	*FHOD3*	-0.73	8.08
Alkaline phosphatase	rs573303250	1	0.001	Upstream gene var.	*LDLRAD2*	rs528777301	21	0.469	Intron var.	*SH3BGR*	0.34	9.99
Alkaline phosphatase	rs573303250	1	0.001	Upstream gene var.	*LDLRAD2*	rs1893383	21	0.241	Intergenic var.	–	-0.36	8.42
Alkaline phosphatase	rs145223095	1	0.008	Downstream gene var.	*LINC00339*	rs534037265	6	0.024	Intergenic var.	–	-0.38	7.33
Alkaline phosphatase	rs145223095	1	0.008	Downstream gene var.	*LINC00339*	rs145058027	8	0.014	Intergenic var.	–	-0.46	8.01
Alkaline phosphatase	rs635634	9	0.186	Upstream gene var.	*ABO*	rs2301455	1	0.441	Intron var.	*C1orf105*	0.03	8.05
Alkaline phosphatase	rs635634	9	0.186	Upstream gene var.	*ABO*	rs2276065	11	0.228	Missense var.	*TREH*	0.05	20.75
Alkaline phosphatase	rs635634	9	0.186	Upstream gene var.	*ABO*	rs681343	19	0.492	Synonymous var.	*FUT2*	-0.10	113.07
Alkaline phosphatase	rs635634	9	0.186	Upstream gene var.	*ABO*	rs12019136	19	0.042	Intron var.	*FUT6*	-0.08	13.56
Alkaline phosphatase	rs635634	9	0.186	Upstream gene var.	*ABO*	rs533852698	21	0.002	Intron var.	*MIS18A*	0.31	7.41
Alanine aminotransf.	rs13141441	4	0.284	Regulatory region var.	–	rs1823740	15	0.474	Intron var.	*AGBL1*	0.02	7.78
Alanine aminotransf.	rs13141441	4	0.284	Regulatory region var.	–	rs3747207	22	0.215	Intron var.	*PNPLA3*	-0.03	8.98
Alanine aminotransf.	rs429358	19	0.156	Missense var.	*APOE*	rs2176521	3	0.103	Intergenic var.	–	0.04	7.37

Table F.1: *(continued from the previous page)*

Trait	LHS SNP					RHS SNP					Beta	P-value
	rsID	Chr.	MAF	Consequence	Gene	rsID	Chr.	MAF	Consequence	Gene		
Alanine aminotransf.	rs738409	22	0.216	Missense var.	*PNPLA3*	rs13130041	4	0.284	Intron var.	*HSD17B13*	-0.03	9.23
Alanine aminotransf.	rs738409	22	0.216	Missense var.	*PNPLA3*	rs138165101	5	0.010	Intergenic var.	–	0.12	7.37
Aspartate aminotransf.	10:101101889:GT:G	10	0.002	Intron var.	*CNNM1*	rs34472802	1	0.145	Intergenic var.	–	-0.33	7.35
Aspartate aminotransf.	10:101101889:GT:G	10	0.002	Intron var.	*CNNM1*	rs200725569	2	0.269	Downstream gene var.	*AC068042.1*	-0.26	7.79
Aspartate aminotransf.	10:101101889:GT:G	10	0.002	Intron var.	*CNNM1*	rs61029446	2	0.360	Intergenic var.	–	-0.24	7.86
Aspartate aminotransf.	10:101101889:GT:G	10	0.002	Intron var.	*CNNM1*	rs6760787	2	0.422	Regulatory region var.	–	-0.23	7.49
Aspartate aminotransf.	10:101101889:GT:G	10	0.002	Intron var.	*CNNM1*	rs62170982	2	0.034	Downstream gene var.	*CBX3P6*	-0.68	7.37
Aspartate aminotransf.	10:101101889:GT:G	10	0.002	Intron var.	*CNNM1*	rs7631623	3	0.196	Downstream gene var.	*AC117444.1*	-0.31	9.27
Aspartate aminotransf.	10:101101889:GT:G	10	0.002	Intron var.	*CNNM1*	rs35655659	3	0.447	Intergenic var.	–	0.24	8.92
Aspartate aminotransf.	10:101101889:GT:G	10	0.002	Intron var.	*CNNM1*	3:26881066:AT:A	3	0.378	Intergenic var.	–	0.24	7.43
Aspartate aminotransf.	10:101101889:GT:G	10	0.002	Intron var.	*CNNM1*	rs2256056	4	0.333	Intron var.	*AF146191.4*	-0.27	9.59
Aspartate aminotransf.	10:101101889:GT:G	10	0.002	Intron var.	*CNNM1*	rs34825759	4	0.077	Intergenic var.	–	0.46	7.65
Aspartate aminotransf.	10:101101889:GT:G	10	0.002	Intron var.	*CNNM1*	rs775719730	4	0.069	Intergenic var.	–	0.43	7.31
Aspartate aminotransf.	10:101101889:GT:G	10	0.002	Intron var.	*CNNM1*	rs369295678	5	0.074	5'-UTR var.	*ZFP62*	0.50	9.53
Aspartate aminotransf.	10:101101889:GT:G	10	0.002	Intron var.	*CNNM1*	rs3828668	5	0.153	Intron var.	*MARCH3*	0.31	7.57
Aspartate aminotransf.	10:101101889:GT:G	10	0.002	Intron var.	*CNNM1*	rs750903495	5	0.125	Intron var.	*THBS4*	-0.34	8.13
Aspartate aminotransf.	10:101101889:GT:G	10	0.002	Intron var.	*CNNM1*	rs79225032	6	0.110	Intergenic var.	–	-0.46	11.66
Aspartate aminotransf.	10:101101889:GT:G	10	0.002	Intron var.	*CNNM1*	rs13192694	6	0.101	Intergenic var.	–	-0.38	7.60
Aspartate aminotransf.	10:101101889:GT:G	10	0.002	Intron var.	*CNNM1*	rs188932246	6	0.365	Intron var.	*PHACTR2*	0.25	8.23
Aspartate aminotransf.	10:101101889:GT:G	10	0.002	Intron var.	*CNNM1*	rs9389752	6	0.297	Intergenic var.	–	0.25	7.65
Aspartate aminotransf.	10:101101889:GT:G	10	0.002	Intron var.	*CNNM1*	rs36199275	7	0.195	Intergenic var.	–	-0.32	8.67

Table F.1: *(continued from the previous page)*

Trait	LHS SNP					RHS SNP					Beta	P-value
	rsID	Chr.	MAF	Consequence	Gene	rsID	Chr.	MAF	Consequence	Gene		
Aspartate aminotransf.	10:101101889:GT:G	10	0.002	Intron var.	*CNNM1*	rs60191378	7	0.305	Intron var.	*CPVL*	0.25	8.08
Aspartate aminotransf.	10:101101889:GT:G	10	0.002	Intron var.	*CNNM1*	rs501190	7	0.315	Intergenic var.	–	0.25	8.09
Aspartate aminotransf.	10:101101889:GT:G	10	0.002	Intron var.	*CNNM1*	rs13250709	8	0.399	Intron var.	*SAMD12*	0.22	7.34
Aspartate aminotransf.	10:101101889:GT:G	10	0.002	Intron var.	*CNNM1*	rs12155626	8	0.129	Intergenic var.	–	-0.35	8.84
Aspartate aminotransf.	10:101101889:GT:G	10	0.002	Intron var.	*CNNM1*	rs9325660	8	0.143	Intron var.	*RP11-145O15.3*	-0.32	7.54
Aspartate aminotransf.	10:101101889:GT:G	10	0.002	Intron var.	*CNNM1*	rs75759636	9	0.135	Intron var.	*C9orf153*	-0.41	10.88
Aspartate aminotransf.	10:101101889:GT:G	10	0.002	Intron var.	*CNNM1*	rs11604937	11	0.337	Intron var.	*RP11-35J10.5*	-0.26	8.59
Aspartate aminotransf.	10:101101889:GT:G	10	0.002	Intron var.	*CNNM1*	rs7120587	11	0.294	Intron var.	*PIWIL4*	0.27	8.82
Aspartate aminotransf.	10:101101889:GT:G	10	0.002	Intron var.	*CNNM1*	rs7296542	12	0.208	Noncod. transc. exon var.	*RP4-601P9.1*	0.28	7.41
Aspartate aminotransf.	10:101101889:GT:G	10	0.002	Intron var.	*CNNM1*	rs28522347	14	0.372	Intergenic var.	–	0.27	9.82
Aspartate aminotransf.	10:101101889:GT:G	10	0.002	Intron var.	*CNNM1*	rs210306	14	0.211	Intron var.	*AL163953.3*	0.31	9.22
Aspartate aminotransf.	10:101101889:GT:G	10	0.002	Intron var.	*CNNM1*	rs3784351	15	0.398	Intron var.	*ITGA11*	-0.26	9.25
Aspartate aminotransf.	10:101101889:GT:G	10	0.002	Intron var.	*CNNM1*	rs74851099	15	0.154	Regulatory region var.	–	0.35	9.32
Aspartate aminotransf.	10:101101889:GT:G	10	0.002	Intron var.	*CNNM1*	rs2930581	15	0.423	Intergenic var.	–	0.23	7.39
Aspartate aminotransf.	10:101101889:GT:G	10	0.002	Intron var.	*CNNM1*	rs113438694	15	0.128	Intron var.	*IL16*	0.37	7.97
Aspartate aminotransf.	10:101101889:GT:G	10	0.002	Intron var.	*CNNM1*	rs7213009	17	0.069	Intron var.	*CACNG5*	-0.48	7.78
Aspartate aminotransf.	10:101101889:GT:G	10	0.002	Intron var.	*CNNM1*	rs8096114	18	0.118	Intron var.	*ZNF521*	0.35	7.53
Aspartate aminotransf.	10:101101889:GT:G	10	0.002	Intron var.	*CNNM1*	rs72887889	18	0.201	Intron var.	*RP11-449D8.5*	-0.30	7.94
Aspartate aminotransf.	10:101101889:GT:G	10	0.002	Intron var.	*CNNM1*	rs111811849	18	0.442	Regulatory region var.	–	0.25	9.09
Aspartate aminotransf.	10:101101889:GT:G	10	0.002	Intron var.	*CNNM1*	rs6090229	20	0.456	Intron var.	*RP11-305P22.9*	-0.25	8.65
Aspartate aminotransf.	10:101101889:GT:G	10	0.002	Intron var.	*CNNM1*	rs4810671	20	0.278	Intron var.	*SULF2*	-0.27	8.12

Table F.1: *(continued from the previous page)*

Trait	LHS SNP					RHS SNP					Beta	P-value
	rsID	Chr.	MAF	Consequence	Gene	rsID	Chr.	MAF	Consequence	Gene		
Aspartate aminotransf.	10:101101889:GT:G	10	0.002	Intron var.	*CNNM1*	rs425541	21	0.368	Intergenic var.	–	0.26	9.73
Aspartate aminotransf.	rs756350040	19	0.074	Intron var.	*HAPLN4*	rs141986761	16	0.002	Intron var.	*RBFOX1*	0.43	8.38
Aspartate aminotransf.	rs3747207	22	0.215	Intron var.	*PNPLA3*	rs13150834	4	0.438	Upstream gene var.	*MIR5705*	-0.03	12.33
Direct bilirubin	rs34983651	2	0.312	Intron var.	*UGT1A6*	rs13071085	3	0.126	Intron var.	*PCCB*	0.03	7.38
Direct bilirubin	rs34983651	2	0.312	Intron var.	*UGT1A6*	rs537623375	5	0.004	Intergenic var.	–	-0.17	7.38
Direct bilirubin	rs78833965	6	0.114	Intron var.	*SLC17A1*	rs140283467	3	0.003	Regulatory region var.	–	0.34	7.49
Cholesterol	11:116623213:TA:T	11	0.139	Intron var.	*BUD13*	rs56043443	19	0.151	Intron var.	*CTC-559E9.6*	-0.04	7.68
Cholesterol	rs814573	19	0.188	Downstream gene var.	*APOC1*	rs964184	11	0.132	3'-UTR var.	*ZNF259*	-0.04	8.20
Cystatin C	rs185932638	20	0.002	Intron var.	*RP5-1185K9.1*	rs149514716	12	0.048	Intron var.	*SDS*	-0.68	10.19
Cystatin C	rs139450906	20	0.017	Intergenic var.	–	rs11652995	17	0.025	Intron var.	*TTYH2*	-0.25	7.91
Cystatin C	rs77149469	20	0.002	Regulatory region var.	–	rs77823120	2	0.033	Intron var.	*PQLC3*	-0.63	7.37
Cystatin C	rs77149469	20	0.002	Regulatory region var.	–	6:44066943:TA:T	6	0.498	Intergenic var.	–	-0.21	8.06
Cystatin C	rs77149469	20	0.002	Regulatory region var.	–	rs4872091	8	0.255	Intron var.	*RP11-1149O23.3*	0.28	9.57
Cystatin C	rs77149469	20	0.002	Regulatory region var.	–	rs199892857	11	0.046	Upstream gene var.	*ARL2*	-0.47	7.55
Glycated haemoglobin	rs1800562	6	0.078	Missense var.	*HFE*	rs574802604	9	0.004	Intron var.	*UBAP1*	-0.29	7.51
Glycated haemoglobin	rs7903146	10	0.290	Intron var.	*TCF7L2*	rs139168400	12	0.004	Intron var.	*UBE3B*	-0.19	7.44
Glycated haemoglobin	rs59616136	19	0.093	Intron var.	*MYO9B*	rs149233151	6	0.014	Regulatory region var.	–	0.15	8.13
HDL cholesterol	rs56228609	16	0.316	Upstream gene var.	*AC012181.1*	rs557467562	3	0.008	Intron var.	*AC136289.1*	0.14	8.04
HDL cholesterol	rs72836561	17	0.032	Missense var.	*CD300LG*	rs147648373	1	0.014	Intron var.	*STX6*	-0.26	7.45
HDL cholesterol	rs7412	19	0.081	Missense var.	*APOE*	rs62284708	3	0.003	Intron var.	*NAALADL2*	0.36	8.03
Lipoprotein A	rs6938647	6	0.218	Intron var.	*LPA*	rs11007996	10	0.054	Intergenic var.	–	-0.06	7.89

Table F.1: *(continued from the previous page)*

Trait	LHS SNP					RHS SNP					Beta	P-value
	rsID	Chr.	MAF	Consequence	Gene	rsID	Chr.	MAF	Consequence	Gene		
Lipoprotein A	rs571848809	6	0.107	Intron var.	*LPA*	rs7412	19	0.081	Missense var.	*APOE*	-0.10	19.32
SHBG	rs62059839	17	0.261	Intron var.	*SHBG*	rs111631648	6	0.028	Intergenic var.	–	-0.07	7.86
Total bilirubin	rs12371715	12	0.199	Intron var.	*SLCO1B3*	rs78231801	3	0.037	Intron var.	*ST6GAL1*	0.07	7.90
Total bilirubin	rs12814270	12	0.221	Intron var.	*LST3*	rs543050186	13	0.001	Intergenic var.	–	-0.42	8.41
Total bilirubin	rs7132068	12	0.252	Intron var.	*LST3*	rs527860769	8	0.002	Intergenic var.	–	-0.32	7.99
Vitamin D	rs549227613	11	0.002	Intron var.	*SPON1*	rs34890943	1	0.285	Intron var.	*GREM2*	-0.30	8.28
Vitamin D	rs549227613	11	0.002	Intron var.	*SPON1*	rs7609126	2	0.422	Intron var.	*C2orf42*	0.26	7.90
Vitamin D	rs549227613	11	0.002	Intron var.	*SPON1*	rs34115555	3	0.448	Intergenic var.	–	0.28	9.61
Vitamin D	rs549227613	11	0.002	Intron var.	*SPON1*	rs6792646	3	0.321	Intergenic var.	–	0.27	8.22
Vitamin D	rs549227613	11	0.002	Intron var.	*SPON1*	rs10063011	5	0.415	Upstream gene var.	*GRXCR2*	-0.28	9.28
Vitamin D	rs549227613	11	0.002	Intron var.	*SPON1*	rs80281693	7	0.073	Intron var.	*PDE1C*	-0.47	7.91
Vitamin D	rs549227613	11	0.002	Intron var.	*SPON1*	rs17145897	7	0.180	Intron var.	*ICA1*	-0.34	8.44
Vitamin D	rs549227613	11	0.002	Intron var.	*SPON1*	rs10046752	8	0.384	Intron var.	*CSMD1*	0.26	7.80
Vitamin D	rs549227613	11	0.002	Intron var.	*SPON1*	rs7072473	10	0.457	Downstream gene var.	*RN7SL232P*	-0.26	7.95
Vitamin D	rs549227613	11	0.002	Intron var.	*SPON1*	rs6588742	10	0.159	Intergenic var.	–	0.40	8.74
Vitamin D	rs549227613	11	0.002	Intron var.	*SPON1*	rs35443850	12	0.084	Intron var.	*ERC1*	0.45	7.62
Vitamin D	rs549227613	11	0.002	Intron var.	*SPON1*	rs10847657	12	0.181	Intron var.	*TMEM132C*	-0.32	7.43
Vitamin D	rs549227613	11	0.002	Intron var.	*SPON1*	rs3825720	14	0.474	Noncod. transc. exon var.	*KRT18P7*	-0.27	8.11
Vitamin D	rs549227613	11	0.002	Intron var.	*SPON1*	rs1687668	16	0.230	Intergenic var.	–	0.31	8.20
Vitamin D	rs549227613	11	0.002	Intron var.	*SPON1*	rs75661931	21	0.023	Intron var.	*TMPRSS15*	0.98	8.06
Vitamin D	rs10500804	11	0.422	Intron var.	*CYP2R1*	4:31131887:AC:A	4	0.034	Intron var.	*PCDH7*	0.06	7.79

COLOPHON

This thesis was typeset in Minion 3 12 pt using LaTeX 2ε. Mathematical formulas were produced with the help of the *MinionPro* package. All data plots were generated in R with *ggplot2*.