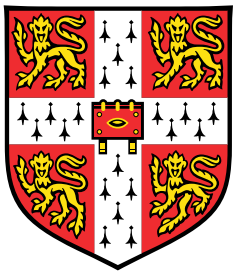


Modelling human complex traits with regression and neural-network based methods

Marton Kelemen

Wellcome Sanger Institute
University of Cambridge

This dissertation is submitted for the degree of
Doctor of Philosophy

Darwin College

November 2020

Declaration

This thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text. It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. I further state that no substantial part of my thesis has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. It does not exceed the prescribed word limit for the relevant Degree Committee. This dissertation contains fewer than the limit of 60,000 words set by the Biology Degree Committee.

Marton Kelemen
November 2020

Modelling human complex traits with regression and neural-network based methods

Marton Kelemen

Identifying how epistasis, non-linear genetic effects, contribute to phenotypic variance in humans has been an enduring challenge. So far neither the computational resources that could accommodate higher-order interactions at scale nor the large-scale population cohorts with adequate statistical power were available up until recently. With the advent of graphics processing unit computing farms and neural-network based methods, together with large biobank-scale data sets, such as the UK Biobank which offers a sample size of ~500K, this has been changing. These developments offer opportunities for the development of novel approaches that could provide insights into the genetic underpinnings of complex disease risk and trait variation.

After reviewing the necessary background material, this work consists of three research chapters. The organising theme of these is the building of genotype-phenotype maps, which grow from the simple additive, through the two-way interactions, up to higher-order interactions in the last chapter.

I begin by covering the common quality control steps and basic additive association analyses I carried out that explored the information boundaries of my data which serves as the foundation for the rest of my work. I managed to recover primary association signals described in the literature for my cohorts confirming the validity of my data processing steps. I also describe a novel method that exploits shared genetic effects to improve risk prediction for related traits. Relative to baselines, this improved squared correlations between observed and predicted sub-phenotypes by ~25% and ~19% for ulcerative colitis and Crohn's disease, respectively.

Building on the previously prepared data sets, I searched for two-way interactions using standard statistical methods belonging to the regression framework. In the UK Biobank cohort I pursued a hypothesis-free approach to consider interactions both within and between the genomic domains of SNP, transcription and protein derived predictors. For the much smaller inflammatory bowel disease studies, I followed a hypothesis driven strategy to reduce search space which only considered haplotype-specific interactions between biologically plausible loci to increase power. I found that the results from both of these approaches were consistent with the null hypothesis of no significant contribution to phenotypic variance from non-linear genetic effects.

Parallel to my search for epistasis using regression based models, I also considered the neural-network framework to find indirect evidence for non-linear effects contributing to

phenotypic variance. I confirmed via a large-scale simulation study the potential of neural-networks to be able to identify interactions at a higher accuracy than standard regression based methods. In the real datasets, I searched for individual epistatic interactions using both experimental approaches from the literature, together with methods that I developed for this purpose. However, I was unable to find convincing evidence for statistical interactions contributing to complex trait variance.

In summary, I found that despite the large cohorts I had access to and the modern non-linear methods I deployed, evidence for non-linear genetic effects contributing to complex human trait variance remained elusive.

Acknowledgements

I thank my supervisors Carl Anderson and Chris Wallace, without whom this work would not exist. Thank you for putting up with my stubbornness, I hope that my future work will make you proud. I am immensely grateful for the opportunity to pursue a research topic of my choice, which was made possible by the Wellcome Trust who supported my work financially, and the Mathematical Genomics and Medicine administrators who organised this exceptional programme. I also thank Nadav Brandes for kindly agreeing to share the data and provided help to navigate the PWAS method.

This thesis is dedicated to my mother, Judith Dimitrova, and my father, Lajos Kelemen.

Table of contents

List of figures

List of tables

Nomenclature

Acronyms / Abbreviations

AUC	Area under the curve
BMI	Body mass index
CD	Crohn's disease
CNN	Convolutional neural-network
FDR	False discovery rate
FIS	Fluid intelligence score
FNN	Fully-connected neural-network
GPU	Graphics processing unit
GWAS	Genome-wide association study
HLA	Human leukocyte antigen
HMM	Hidden Markov model
HWE	Hardy–Weinberg equilibrium
IBD	Inflammatory bowel disease
KRR	Kernel-ridge regression
LD	Linkage disequilibrium
LMM	Linear mixed-effects model
MAF	Minor allele frequency

NN Neural-network

OLS Ordinary least squares

PRS Polygenic risk score

QC Quality control

REML Restricted maximum likelihood

RKHS Reproducing kernel Hilbert space

ROC Receiver operating characteristic

RR Ridge regression

SNP Single-nucleotide polymorphism

T1D Type 1 diabetes

TF Transcription factor

TWAS Transcriptome-wide association study

UC Ulcerative colitis

UKBB UK Biobank

WGS Whole-genome sequencing

Chapter 1

Introduction

1.1 Non-linear encoding of genetic information

The human genome contains over three billion base pairs, each carrying one of four possible nucleobases. Although this may appear like a substantial amount of data, if the information was encoded linearly, as in a book where the text is read from left to right, then the number of instructions that could be stored would be very limited. Given that the genetic component in the variation of complex traits and organs is substantial (for example, the structure of the brain is ~80% heritable (Jansen et al., 2015)), how could all that information be encoded in a way that fits within our genome's capacity?

It was probably Wright (1932), who first speculated on the potentials of generating a virtually unlimited variety of phenotypic responses from a limited number of genes through their interactions. Today, we take the complex inter-dependency of the genome and the non-linear hierarchies of the resulting biological systems it encodes for granted. At the same time, we have not yet been able to precisely quantify this non-linearity neither within the framework of genetic association studies nor exploit it within the framework of genetic prediction. In this work, my principal concern will be to investigate if there is evidence for this non-linear encoding of information that affects phenotypic variance.

In this chapter I will review the necessary background material and concepts that are relevant to my work. The next sections will cover epistasis, heritability, genome, transcriptome and protein based association studies, genetic prediction, and finally, neural-network based methods.

1.1.1 Epistasis

Epistasis is the term used to describe the aforementioned non-linear encoding of functionality in the genome. At the most basic level, it means that a genetic effect is encoded by the joint action of more than one loci. However, the exact definition itself, and what precise phenomenon it applies to, have been the subject of much debate over the years (Clayton, 2009; Cordell, 2002; Moore and Williams, 2005; Phillips, 2008). There are three different definitions of epistasis in circulation, which are functional (Phillips, 2008), compositionional (Bateson, 1906) and statistical (Fisher, 1918). I will define each form in turn and also highlight the properties that are most relevant to my work.

Functional epistasis encompasses all inter-dependency of functionality between areas of the genome that encode different aspects of the whole system. This is the property that describes the non-linear encoding of information in the genetic code. A key attribute of functional epistasis is that it does not assume any inter-individual genetic variation in the population; rather, it may be understood as the the genome's interaction against itself. It merely describes a static property of the genome which may be common to all individuals (Phillips, 2008).

Compositional epistasis is the phenomenon where the expected phenotype of one locus is masked by genes at other loci, as observed by departures from Mendelian ratios in dihybrid crosses. This is the original definition of epistasis by Bateson (1906), and is also the way many textbooks introduce the concept (Guénet et al., 2015). While this also describes a biological function, this definition also suggests that there would have to exist genetic variation at all involved loci in order for such phenomenon to be observable in the first place.

Statistical epistasis describes deviations from additivity in a statistical model which describes how genetic variation affects phenotypic variation in a population. This is Fisher's definition of epistasis (Fisher, 1918). Here, the emphasis is entirely on the impact on phenotypic variance, which requires that all involved loci must be polymorphic in a population of samples, otherwise their effects would not be possible to estimate in a statistical framework.

These above descriptions are based on the definitions of epistasis put forward by Phillips (2008). It is necessary to further clarify these concepts, their relationships to each other, and under what circumstances they overlap or differ from each other.

Functional epistasis is the broadest category of the three, as with a few exceptions that I will cover later, all statistical epistasis also requires functional epistasis as well. It is possible to have functional epistasis without the presence of statistical epistasis, as all loci within the genome that are non-polymorphic but depend on functionality elsewhere, are in fact engaged in functional epistasis.

Statistical epistasis may only exist without functional epistasis under a few special circumstances. Such 'technical' statistical epistasis, which does not arise from underlying biological processes, originates from the physical properties of the DNA molecule or imperfect recordings of the phenotype. I will discuss this topic in depth in sections 1.1.6.1 and 1.1.6.2.

In a study of a population of individuals, as is the case in most association studies, compositional epistasis simply describes a snapshot of the mechanism by which statistical epistasis manifests itself. Compositional epistasis also qualifies as functional epistasis as it can only exist due to the functional relationships between different loci. Thus, as this form of epistasis may be defined as the intersection of the other two, treating it as a separate entity would not contribute to my work here. Therefore, I will not be considering compositional epistasis further from this point onward.

The real conceptual difference lies between statistical and functional epistasis. While functional epistasis is possible without allelic substitutions that would result in changes in phenotypic variance, statistical epistasis is not possible under such circumstances. One particular area where one may be tempted to expect statistical epistasis is where an interaction takes place between a single polymorphic locus (such as a SNP) and non-polymorphic loci. However, unless the phenotypic effect depends on the joint action of at least another variant, this interaction can only be classed as functional epistasis. Another difference between statistical and functional epistasis is the number of opportunities for them to occur. As approximately there is only one SNP per a 1,000 base pairs (Marth et al., 1999), this would suggest that there are many times more opportunities for loci to be involved in functional epistasis than in statistical epistasis. Therefore, the latter may be expected to be a correspondingly rarer phenomenon.

Moving forward, if we accept that compositional epistasis is not a distinct category, that leaves only two forms of epistasis to consider, functional and statistical. These require two different approaches to study which I will describe next.

1.1.2 The two main forms of epistasis

As functional epistasis can arise from variation within genomes, and statistical epistasis arises from variation between individuals, these two forms of epistasis may be studied in frameworks that are conceptually orthogonal. Consider the genotype matrix of a hypothetical population in the figure below:

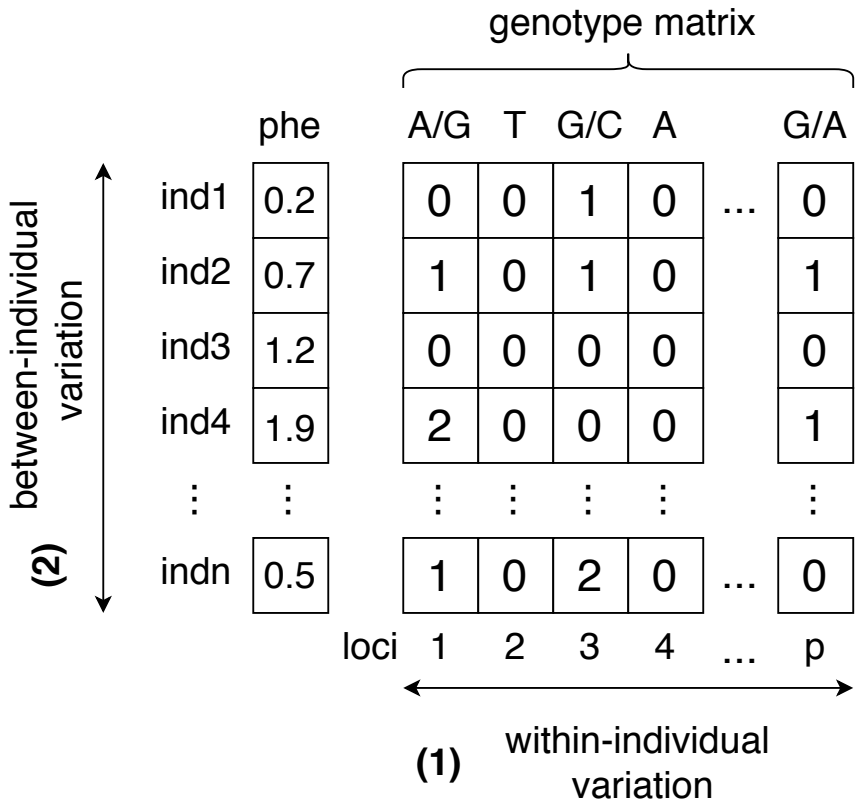


Fig. 1.1 **Hypothetical genotype matrix of n individuals at p loci.** Functional epistasis can take place between any of the p loci. However, statistical epistasis can only take place between loci 1, 3 and p, which are SNPs. The two loci, 2 and 4, do not vary in a population, therefore interactions between them, or even between these and SNPs, cannot contribute to phenotypic variance. Figure and terminology adapted from Angermueller et al. (2016).

Studies whose goal is to investigate **(1) within-individual variation**, compare different areas of the (reference) genome to discover its mechanism of effect. As these studies have a different objective than the work in this thesis, I will only provide a brief overview of their purpose. Here, the training data are regions of the base sequence, such as those captured by FASTA files, which are then related to properties of sequence features. These features include regulatory motifs such as transcription factor (TF) binding sites, enhancers, promoters or their combinatorial relationships that control the function of genes, collectively known as the cis-regulatory code. Recent successful examples include the de novo prediction of the function of non-coding variants from 1000bp contexts (Zhou and Troyanskaya, 2015), and learning aspects of 3D genome folding from 1Mb sequence contexts (Fudenberg et al., 2019).

On the other hand, studies with the objective to explain **(2) between-individual variation**, seek to relate individual-level phenotypic variance to genetic variation in a population. The training data is genetic variation in a population, as captured on microarrays or WGS

data and phenotype labels. The genetic data and phenotypes are then related to each other to provide inference about either individual loci or their aggregate effects on phenotypic variance. Example applications for the former include all GWAS (Tam et al., 2019), and studies that build genetic risk prediction models via polygenic scores may serve as illustrative examples for the latter (Khera et al., 2018).

In summary, **(1)** aims to investigate features of the genome common to all individuals, and **(2)** seeks to reveal what makes us phenotypically different.

1.1.3 The importance of epistasis in understanding biology

Beyond the general insight into understanding how information is encoded in the genome, functional epistasis may aid the interpretation of individual SNP effects. It is self-evident that a SNP, which is a single molecular change, cannot effect an organism-level phenotype directly. The SNP's function may only be understood through its interactions with the complex cascade of downstream systems which it is a part of. As a hypothetical example, consider a single nucleotide change that exerts its influence by knocking out a TF binding site, which subsequently would affect protein-protein interactions in a cascade of downstream events ultimately leading to a phenotypic change. Note that other than the SNP itself, none of the other elements need to be polymorphic. Thus, identifying functional epistasis may provide insights on the mechanism of effect behind GWAS associations (Gallagher and Chen-Plotkin, 2018).

On the other hand, statistical epistasis may reveal the mechanism of joint effects of multiple SNPs. For example, if both variants are required to increase risk, this may suggest pathway redundancy (Xie et al., 2018). Alternatively, if risk does not increase further with the presence of both risk variants, this in turn may suggest that both markers are on the same pathway, and just one is sufficient to impair its functionality (Castillejo-López et al., 2012).

1.1.4 Examples of statistical epistasis in humans

The dramatic impact of epistatic interactions observed in model organisms and populations created via artificial selection, such as the coat colour of Labrador retrievers (Everts et al., 2000), appear to be absent in humans. In fact, there have been very few confirmed cases of statistical epistasis, with much more subtle effects.

Initially, hypothesis-free, exhaustive searches for pairwise interactions by Wan et al. (2010) and Lippert et al. (2013) on traits in the Wellcome Trust Case–Control Consortium studies appeared to find signal in the HLA region. However, no subsequent efforts were made to validate these findings or commission follow-up studies to replicate their results. In

2014, another large scale study appeared to find evidence for widespread statistical epistasis affecting gene expression in whole blood data (Hemani et al., 2014). However, in a follow-up study by Wood et al. (2014), it was found that the previously identified interactions could also be explained by artefacts caused by LD. I will cover the details of how LD and haplotype effects may generate false statistical epistasis under section 1.1.6.2. Another example that illustrates the lack of reliable results was a study on Alzheimer's disease that appeared to find SNP-SNP interactions between variants in *RNF219* and *APOE4* (Rhinn et al., 2013). This study was subsequently retracted by the authors, due to problems caused by sample processing errors (Rhinn et al., 2015).

Most confirmed examples for statistical epistasis with reliable evidence come from hypothesis driven studies. Here, much about the biological mechanism was already known, and the researchers were investigating specific candidate loci to confirm what was already suspected. One such example was a study of rheumatoid arthritis by Génin et al. (2013). Here, working on known risk loci, the authors only needed to perform epistasis tests on two genes, *BANK1* and *BLK* (which are on different chromosomes), and were able to identify a SNP-SNP interaction. In a more recent hypothesis driven study, Belbin et al. (2019) found a statistical interaction between loci on genes *BDNF* and *DBH* (also on different chromosomes) that increased risk of Alzheimer's disease.

1.1.5 Variance component analyses of epistasis

Variance component analyses aim to decompose the phenotypic variance into its constituent components of genetic, environmental and noise terms. Assuming no interaction effects between genetic and environmental factors, in this framework the total genetic variance (σ_g^2) is given by the sum of three orthogonal components:

$$\sigma_g^2 = \sigma_a^2 + \sigma_d^2 + \sigma_i^2, \tag{1.1}$$

where σ_a^2, σ_d^2 and σ_i^2 denote the additive, dominance and epistatic variance components, respectively. These components capture the aggregate effects on phenotypic variance from genetic effects in which alleles contribute additively, by masking the effects of other alleles and via statistical interactions, for the additive, dominance and epistatic variance components, respectively. A key advantage of considering sources of variance as components is that, while it cannot identify individual variants or their combinations, it is a powerful approach to make general statements about the sources of variation. Thus, variance components may be used to infer the existence of epistasis, even in the absence of adequate statistical power to confirm the identities of any particular loci involved.

There have been a number of recent well powered studies that aimed to find evidence for statistical interactions affecting phenotypic variance relying on either pedigree or genetic data. Relying on pedigree-based meta-analysis involving over 14 million twin pairs, Polderman et al. (2015) performed variance component analyses to dissect the genetic architecture of complex traits. They found that for the majority of traits a parsimonious model, which only relied on the environmental and additive genetic variance components, proved sufficient to explain phenotypic variation. A similar picture has been emerging from studies involving molecular genetic data. A recent study that relied on WGS data and quantitative physiological traits by Wainschtein et al. (2019), found that the additive genetic variance component explained virtually all of the phenotypic variance attributable to heritability.

1.1.5.1 An alternative explanation to the apparent lack importance of epistatic variance

As the evidence from both association and variance component studies suggest that epistasis contributes little to phenotypic variance, many investigators concluded that statistical epistasis is irrelevant to complex trait genetics (Crow, 2010; Mäki-Tanila and Hill, 2014). However, additive genetic variance may appear to be adequate for accounting for the genetic contribution to phenotypic variance due to reasons other than the obvious explanation, that additive genetic action would be all that matters.

An alternative explanation to the apparent lack importance of epistatic variance was put forward by Huang and Mackay (2016), where the authors argued that this may be due to an artefact of the way classical variance component models are parameterised. They argued that depending on the order in which the variance components are accounted for different, equally convincing models may be constructed. To illustrate this, consider the simple case of a single locus model. Here, the traditional variance component model is fit by first maximising the variance explained by the additive component, and the epistatic component is only considered as a residual. This is equivalent to the least squares solution where the line of best fit captures the additive variance, and deviations from this line capture the non-linear variance. The authors showed, that if the order in which the components explain phenotypic variance is reversed, this could also reverse their relative importance as well. For example, they showed that even in the absence of genuine non-linear effects, if the model is fit to explain the non-linear variance first, then the non-linear component could appear to be more important than the additive component. They reasoned, that as there is no natural correspondence between biological gene action and the variance components, the ordering of the fit is arbitrary. Thus, the authors concluded that there is no intrinsic justification for giving priority to the additive variance component over the non-linear component.

Hansen (2013) also argued that variance components are inadequate to capture the importance of epistasis, by pointing out that epistatic genetic processes contribute to the additive variance component as well. Therefore, while the estimated variance components are statistically orthogonal, they are not 'biologically orthogonal', as gene actions that contribute to the variance components also overlap. For instance, the additive variance component will receive contributions from epistatic (and dominance) effects as well, unless all involved variants have maximum variance (a MAF of 0.5 for all loci (Huang and Mackay, 2016)). Even in the presence of a genuine two-way interaction, unless both alleles have a MAF of 0.5, the additive variance component will appear to dominate. In a situation where one allele is very rare, irrespective of the MAF at the other loci, almost all of the variance will appear additive, as there will be very little epistatic variance generated (this is similar to the situation where rare SNPs with additive effects generate little additive variance). An intuitive explanation for this phenomenon is that this situation approximates the scenario where there is a functional epistatic interaction between a SNP and a non-polymorphic loci I described in section 1.1.1.

1.1.5.2 Is non-linear population genetic variance needed for non-linear information encoding?

One apparent paradox is that if the variance of a population's genotype-phenotype map is substantially additive at any given time, then how did the non-linear information encoding in the genome occur in the first place? Also, how did the simpler genomes of single-cell organisms, that were our evolutionary ancestors, change into the genomes of humans, a transformation that is altogether non-linear? This paradox appears to be particularly puzzling, given that the raw material evolution works with are mutations that typically arise via a linear process, one at a time. I see two potential explanations. It is possible, that in order to encode information in a non-linear manner, non-linear genetic variance arising from statistical epistasis is simply not required. An alternative explanation is that non-linear genetic variance is required, but it only occurs under particular circumstances and it may be a transient phenomenon. These two theories are expanded upon in the following paragraphs.

Fisher proposed that selection operates as an adaptive process with a single global optimum. Under this model, the selection coefficient of an allele is determined by its dependency on a constant genetic background of already fixed loci (Fisher, 1930). In other words, the probability of the increase or decrease of a new allele's frequency depends on functional epistasis with the rest of the genome without generating any statistical epistasis. This way, non-linear information may be encoded from additive changes, one substitution at a time. Therefore, natural selection could operate via a process that requires only additive genetic

variance to build up the non-linear genome information structure. From this perspective, the previously described problem is only a paradox upon first consideration. To further illustrate this explanation, one may compare this process to how an artist may draw a work of arbitrary complexity. Her pencil will only ever need to touch the canvas at a single point at any given time, but the probability of the pencil leaving a mark there is always conditioned upon what she has drawn so far.

The alternative explanation is that non-additive variance does contribute to natural selection, and thus to the non-linear information encoding into the genome. Models that allow for statistical epistasis to exist permanently originate from Wright's Shifting Balance Theory (Wright, 1932). Here, instead of Fisher's single global fitness optimum that would drive all new alleles to fixation or extinction, an adaptive landscape of optima exist. These multiple fitness optima would then permit the existence of statistical epistasis, which could then play a role in facilitating the movement of populations between these adaptive peaks. However, as nature tends to prefer parsimonious solutions over elaborate ones, in practice, Wright's theory found little empirical support in natural populations (Coyne et al., 1997).

1.1.6 Challenges of statistical epistasis detection

There are a number of challenges facing researchers interested in finding evidence for statistical epistasis. These challenges include statistical/computational considerations, LD and artefacts arising from the thresholding of phenotypes. In the following sections I will consider each challenge in turn.

1.1.6.1 Statistical and computational challenges

An exhaustive search for all two-way interactions from p markers will generate $p(p-1)/2$ association tests. Therefore, the computational demands for performing an exponentially increasing number of tests may become a challenge, especially if p was large to begin with, due to a dense marker panel for example. However, with the advent of large computing cluster farms and GPUs, the computational demands are seen as less of a burden today than they were in previous years (Ponte-Fernández et al., 2020).

The statistical challenges originate from the substantial multiple testing burden that also arises from performing a great number of interaction tests (Van Steen, 2012b; Wei et al., 2014b). This issue is exacerbated if one is interested in estimating all possible types of interaction effects, including the three terms involving dominance effects as:

$$\begin{aligned} Y = a + \beta_1 G1 + \beta_2 G2 + \beta_{1,2} G1 * G2 + \beta_{1D} G1_D + \beta_{2D} G2_D \\ + \beta_{1,2D} G1 * G2_D + \beta_{1D,2} G1_D * G2 + \beta_{1D,2D} G1_D * G2_D + e, \end{aligned} \tag{1.2}$$

where Y, G, a and e denote the phenotype, the SNPs, the intercept and noise terms, respectively. The βs denote coefficients for each term. Gs take values {0,1,2} which represent the dosages of the alternative allele, and G_Ds take values {0,1} which represent dominance effects. Relative to a main effects only model, fitting the above requires the estimation of an additional three parameters. This increase in the number of terms consumes an additional three degrees of freedom, which results in a corresponding decrease of power to detect any of the individual terms' effects (Wei et al., 2014b).

1.1.6.2 Linkage Disequilibrium

The correlations between different sites of the genome is known as Linkage Disequilibrium (LD) (Nordborg and Tavaré, 2002). While this definition may refer to linkage between loci with an arbitrarily long distance between them (Koch et al., 2013), in practice, this is usually employed in reference to shorter distance ($< 500kb$) relationships. Such dependencies arise through the tendency of short haplotype blocks to be passed along intact without recombination due to their physical proximity. Unless otherwise stated, from here onward I will be using LD to mean such short-distance dependencies.

LD is measured by either squared correlation (r^2) between alleles, or D', which is a normalised version of the difference between observed and expected (assuming independence) haplotype frequencies (Lewontin, 1964). As the definition of both LD and statistical epistasis require the co-occurrence of alleles (Hansen, 2013), these two phenomena perfectly overlap (Wang et al., 2011). This overlap may then cause pure haplotype effects to be mistaken for epistasis. In a study by Hemani et al. (2014), this pattern resulted in the detection of apparent epistasis that was later explained as a haplotype effect by Wood et al. (2014), who found that all apparent statistical interactions lost significance once previously unaccounted variants were added into the model. To illustrate with a specific pattern the technical circumstances where this may occur, consider the following example. Two SNPs, which are in apparent statistical epistasis, are imperfectly tagging a third variant which is the true causal signal. If the causal variant is not in the model this pattern could result in the detection of statistical epistasis, whereas in fact, the two SNPs are imputing a haplotype that involves the causal SNP. The schematic below demonstrates how such a pattern may occur in practice:

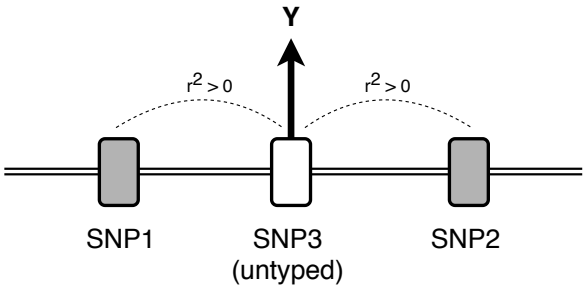


Fig. 1.2 **Illustration of the haplotype effect as an artefact generator for statistical epistasis.** The two SNPs imperfectly tag ($r^2 > 0$) the untyped SNP3, which is the causal variant affecting the phenotype Y. This pattern can arise even if the r^2 between SNP1 and SNP2 is zero. Statistical epistasis may only be generated by this pattern if the SNP1-SNP2 haplotype is a better tag for SNP3 than either SNPs on their own.

Accounting for this problem is challenging, as simply pruning the markers on LD is insufficient, and in fact, may even make the situation worse by creating more 'untyped' SNPs. Additionally, even if both variants are in perfect LE with each other, the third variant may still be in LD with both of them (Wood et al., 2014). The best way to solve this problem is to add the third marker into the model. This would cause the interaction term to loose significance, as its signal would be absorbed by the new predictor. As this solution requires perfect coverage, in practice, this is frequently not realistic as the third SNP may have been lost due to genotype or imputation QC. The only other alternative solution is to only consider markers sufficiently far apart for short-range LD not to exist.

Although inferring unobserved genetic variation satisfies the definition of statistical epistasis, in a sense that the joint effect of the SNPs deviates from their sum, these effects do not contribute to the information encoding capacity of the genome. Such effects are entirely 'flat', and merely arise due to the the physical properties of the DNA molecule; thus, such artefacts do not increase the total information storage capacity of the genome.

One final aspect of LD relevant for epistasis detection is how the underlying causal signal tagged by markers decays with distance. For additive associations, signal decays linearly with r^2 with the index variant (Vukcevic et al., 2011). However, epistatic signal declines much faster. Additive-by-additive interactions decay with r^4, additive-by-dominance with r^6 and dominance-by-dominance with r^8 (Wei et al., 2014b). Such a fast loss of signal leaves a greatly reduced power to detect epistatic interactions, especially those that involve dominance effects.

1.1.6.3 Thresholding effects for traits with a limited recorded range

Traits for which the sensitivity of measurement does not cover the full range of possibilities are susceptible to additional artefacts that may generate spurious statistical epistasis from variants with additive effects only. Two examples of this phenomenon are binary phenotypes and gene expression data.

Under the liability threshold model, individuals are thought to carry a continuous liability of genetic risk. When the risk crosses a certain threshold this results in a clinical diagnosis on the observed scale, and the underlying effect of the continuous distribution is dichotomised at the point of the threshold. However, individuals are more likely to cross this liability scale threshold and become cases if they have two copies of large effect size alleles (Wei et al., 2014b). This may give the appearance of statistical epistasis, as cases are more likely to carry combinations of these alleles. For example, individuals carrying two risk alleles of very large effect would more likely to cross the disease threshold even if on the liability scale all alleles acted only additively (Wray et al., 2018).

A similar mechanism is at work due the limited dynamic range of probes on microarrays that measure gene expression. If the combined effect of two variants with additive effects exceeds the maximum range of the probe, then their aggregate effect would be less than the sum of their individual effects. Such ceiling effects may also generate apparent statistical epistasis that arise from purely additive effects (Fish et al., 2016). It is important to not confuse this scale effect with a similar sounding problem where apparent statistical epistasis may be generated by the choice of the scale of the recorded phenotype (Wang et al., 2010). In that scenario, interactions between genotypes could have an apparent non-linear effect on the phenotype, depending on the scale of the recorded phenotype. Such spurious interactions may be eliminated via an appropriate reversible transformation of the phenotype's scale (Satagopan and Elston, 2013). However that is a problem that is qualitatively different than the previously described truncation of measurement. As in the latter case, the problem is caused by a truncation of the phenotype recording that results in an irreversible loss of information. As such, this effect cannot be reversed via any kind of transformation; thus, this artefact may only be eliminated by recording the full phenotype range in the first place.

1.1.7 General approaches to epistasis detection

In this section some of the common principles that were found to improve the success of statistical epistasis detection are reviewed. One general approach, common to most methods irrespective of the particulars, is to perform the search as a two-stage process. The first step consists of a pre-screening stage which is then followed-up by an association step.

Given the number of tests, managing the dimensionality is a necessary first step. Unless a sound biological prior is known, the most successful approach to accomplish this has been to filter on the additive main effects of each SNP (Cordell, 2009; Marchini et al., 2005; Van Steen, 2012a). Beyond the biological plausibility of independent marginal effects of the interacting loci, there are also statistical reasons why filtering on main effects may be beneficial, even in the absence of genuine marginal effects. Consider the following true model

$$Y = \beta_{1,2} G1 * G2 + e, \tag{1.3}$$

where Y, $G1$, $G2$ and e denote the phenotype, SNP1, SNP2 and the noise term, respectively. SNPs may take values of 0,1 or 2, depending on the number of copies of the alternative allele. However, the following incorrect marginal model was fit instead

$$\widehat{Y} = \widehat{\beta}_1 G1 + \widehat{\beta}_2 G2. \tag{1.4}$$

Given adequate statistical power (considering sample size, MAFs and effect sizes), both terms, $G1$ and $G2$, would be estimated as significant, with coefficients approximately equal to $\beta_{1,2}$. This also holds true even if the marginal effects are estimated in a series of univariate regressions (such as in a GWAS).

The same principle applies to third (and higher) order interactions. Consider the following true model:

$$Y = \beta_{1,2,3} G1 * G2 * G3 + e. \tag{1.5}$$

Once again, we fit a similarly incorrect model that only considers the main effects and second-order interactions:

$$\begin{aligned} \widehat{Y} = \widehat{\beta}_1 G1 + \widehat{\beta}_2 G2 + \widehat{\beta}_3 G3 + \widehat{\beta_{1,2}} G1 * G2 \\ + \widehat{\beta_{2,3}} G2 * G3 + \widehat{\beta_{1,3}} G1 * G3. \end{aligned} \tag{1.6}$$

Assuming adequate statistical power, all tested terms would be identified as significant once again. The reason behind this phenomenon is that for a Dth order interaction to exist, all $D-1$th order interactions must also exist as well (again, assuming adequate power) (Sorokina et al., 2008). This mechanism may then be used to drastically reduce the search-space for (higher-order) interactions by filtering on marginal effects (and lower-order interactions).

The second step is concerned with the identification of individual combinations of SNPs involved in interactions. Methods that accomplish this may be broadly categorised as either traditional statistical approaches that perform an exhaustive search (on the SNPs surviving the first stage), or machine learning methods that carry out non-exhaustive searches. A

notable example for the latter are neural-network based approaches that I will cover in depth in section 1.7. Here, I am only going to consider the two traditional statistical approaches that are relevant for my work.

For binary phenotypes there exist a cases-only test for interactions that consumes only a single degree of freedom (Vittinghoff and Bauer, 2006). This is a powerful method that tests for significant deviations from the expected frequencies of a contingency table conditioned on case status. This test evaluates the hypothesis that if the interaction effect is genuine, then cases carrying the interacting alleles at both loci should be over-represented, relative to what would be expected from the alleles' additive effects. The limitations of this method are that it does not permit the inclusion of covariates and that it is only applicable to binary traits.

The other approach for detecting statistical epistasis involves the regression framework. This approach provides more flexibility, as it is able to facilitate both additive and dominance modes of epistasis, together with an arbitrary number of relevant covariates. To mitigate the problems associated associated with LD (section 1.1.6.2) and the number of tests (section 1.1.6.1), instead of the full model with all terms (eq 1.2), the following model is commonly used, as it only needs to estimate the marginal effects and the additive-by-additive interaction term

$$\widehat{Y} = \widehat{\beta}_1 G1 + \widehat{\beta}_2 G2 + \widehat{\beta_{1,2}} G1 * G2. \tag{1.7}$$

In this approach, the p-value of the $\widehat{\beta_{1,2}}$ term, which may be obtained from the ratio of the coefficient and its standard error (which yields the quantile of a t-distribution), is used to evaluate the evidence for statistical epistasis.

1.2 Heritability

Heritability quantifies the total genetic effect on phenotypic variance. Assuming no interactions between genetic and environmental factors, phenotypic variance is assumed to arise as a sum of the genetic and environmental variance components as

$$\sigma_p^2 = \sigma_g^2 + \sigma_e^2, \tag{1.8}$$

where σ_p^2, σ_g^2 and σ_e^2 denote the variances of the phenotype, genotype and environment, respectively. Heritability (h^2) is then the quantity defined by ratio of the genetic to phenotypic variance components

$$h^2 = \sigma_g^2 / \sigma_p^2. \tag{1.9}$$

If σ_g^2 includes only additive genetic effects (subsequently denoted σ_a^2), then the aforementioned quantity is known as *narrow-sense heritability*. However, if it also includes non-additive effects (such as epistatic and dominance effects), then it is known as *broad-sense heritability*, denoted by H^2. σ_e^2, encompasses all contributions to the phenotype not due to base sequence variation. These contributions include random measurement error, life history and even heritable differences that do not modify the base sequence, such as epigenetic marks. In summary, heritability provides a low resolution overview of the extent that a trait depends on genetic factors without revealing any of the finer details, such as the contribution of individual variants.

There are a few additional subtleties that need to be considered to fully appreciate heritability. For example, the effect of parental genotypes that influences phenotypic variance in their children, when considered to be part of σ_e^2, may decrease h^2, even though the trait variance has not become any less 'genetic'. This effect was demonstrated in a recent study by Kong et al. (2018), where it was shown that non-transmitted parental alleles contributed to phenotypic variance in educational attainment. Additionally, as heritability is a ratio, its magnitude is relative to the environmental variance. That is, even if a trait is under strong genetic influence, h^2 may be low in the presence of an even greater environmental variance. Conversely, if all environmental variation would be eliminated, then h^2 may approximate ~100% even if only a few genetic factors contributed to the phenotype, as then those would be the only source of variance that remained.

Finally, the maximum heritability to be estimated in an analysis is limited to the extent that the genetic factors captured in the study cover all potential genetic effects that influence the phenotype. Therefore, the heritability estimated from SNPs, known as h^2_{SNP}, is typically lower than heritability measured from pedigree based studies h^2_{ped}, as the latter considers all genetic factors. Therefore, h^2_{SNP} is known to be an underestimate of the full h^2 if it does not incorporate rare variants (Wainschtein et al., 2019). On the other hand, as h^2_{SNP} is estimated from molecular data from unrelated individuals, it is less likely to be biased by shared environmental factors (Evans et al., 2018).

1.2.1 Genetic prediction and heritability

Phenotypic variation not due to genetic factors cannot be predicted from genotype data; thus, the ceiling of genetic prediction is heritability (Clayton, 2009). For binary traits the population prevalence of the disease also needs to be considered. For such phenotypes heritability is defined on two levels, liability and observed scale, with the former always being equal or lower than the latter. This liability threshold model assumes that there is an unobserved, continuous liability of risk that arises from the aggregate effect of all risk alleles,

which when cross a threshold, results in a diagnosis with the disease. If the population and sample prevalence of the trait are known, these two heritabilities may be readily converted into one another (Lee et al., 2011). The importance of this property is that for uncommon diseases (prevalence under 1%), even if all causal genetic factors were known, predictability may remain low in the general population due to the low incidence of the condition (Clayton, 2009).

1.2.2 Overview of methods that estimate variance components

To estimate heritability, the phenotypic variance is decomposed into environmental and genetic variance components. This decomposition may be accomplished either via obtaining a direct estimate from the phenotypes and relatedness of a given cohort, or alternatively it may be estimated from GWAS summary statistics.

The first group of methods may be intuitively understood as estimating the total genetic effect on the phenotype by regressing the phenotype on genetic relatedness. Consider

$$Y \sim N(0, \mathbf{K}\sigma_a^2 + \mathbf{I}\sigma_e^2), \tag{1.10}$$

where Y is a phenotype vector and $\mathbf{K} \in \mathbb{R}^{n \times n}$ is the realised kinship matrix for the n individuals considered (Jiang and Reif, 2015). The entries in $\mathbf{K}$ represent the pairwise genetic resemblance between the individuals considered which may be obtained by

$$\mathbf{K} = \frac{\mathbf{X}\mathbf{X}^T}{\gamma}, \tag{1.11}$$

where $\mathbf{X}$ is the genotype matrix of SNPs, γ is a scaling factor proportionate to p (the number of SNPs) that may also optionally take into account MAF and other factors such as imputation quality (Speed et al., 2012).

Next, to obtain σ_a^2, the variance is decomposed via either an actual regression method, such as by Haseman-Elston (HE) regression (Sham and Purcell, 2001), or by restricted maximum likelihood (REML) based solvers (Kang et al., 2008). The former may be simply formulated as an ordinary least squares (OLS)

$$Y' = \beta_k k' + e, \tag{1.12}$$

where the terms are defined for each pair of individuals i and j, as $Y'_{ij} = (y_i - y_j)^2$ and k' is a column vector of pairwise genetic relatedness that may be obtained from the upper/lower

triangle of **K**. The coefficient of this model is then used to obtain the additive variance by

$$\sigma_a^2 = -\beta_k/2. \tag{1.13}$$

An alternative strategy, which does not require genotype level data, is to estimate heritability from GWAS summary statistics and a suitable LD reference panel matched to the ancestry of the GWAS cohort. LDScore-regression (Bulik-Sullivan et al., 2015) is a commonly used method that accomplishes this via the following equations

$$\begin{aligned} LDS_j &= \sum_k r_{jk}^2, \\ \chi_j^2 &= LDS_j * h_{Gj}^2 + n * a + 1 + e. \end{aligned}$$

where LDS_j is the '*LD-score*' of SNP j, which is defined as the sum of all r^2 between SNP_j and all k neighbouring SNPs within a 1cM region. χ_j^2 is the summary test statistic from the GWAS and $(n*a+1)$ is an intercept term scaled for the number of individuals, which also captures the potential confounders of environmental effects and population stratification. The model coefficient, h_{Gj}^2, is the expected heritability contribution of SNP j scaled by n/p (the number of individuals over the number of markers). The overall heritability arising from all markers may then be obtained by summing all individual SNP contributions. The disadvantages of this method are that it only works on common variants, and that due to the additive origins of summary statistics it also cannot produce estimates for non-additive variance components. A further issue is that the intercept term, which was meant to control for the aforementioned confounders, has been recently shown to be less robust with increasing sample sizes and heritability (Loh et al., 2018). Finally, if an appropriate LD reference panel is not be available (for example, if the target samples include individuals with divergent ancestry), then that may result in inaccurate h^2 estimates due to mismatched LD patterns. Despite these limitations, if an accurate estimate of heritability encompassing rare variants is not required, this method is considered to be a useful tool to obtain a rough estimate of the genetic signal available in a dataset.

1.3 Genome-wide association studies

1.3.1 GWAS background

The goal of genome-wide association studies (GWAS) is to identify associations between variation in allele dosages and phenotypic variance. A key advantage of the GWAS design

is that, provided adequate QC measures are taken (described in section 1.3.2.1), reverse causality is not possible as the base sequence is fixed at the moment of conception. Therefore, GWAS allow an exhaustive, hypothesis-free investigation of the genotype-phenotype map of complex diseases and traits.

The origins of GWAS may be traced back to the early 2000s when genome-wide linkage scan studies were gradually replaced by SNP based association studies that started to cover the genome at a higher and higher density. Notable efforts from this 'pre-GWAS' period include studies by Ozaki et al. (2002) (~92K SNPs) and by Klein (2005) (~116K SNPs). The real breakthrough however, came from a study organised by the Wellcome Trust Case Control Consortium (WTCCC et al., 2007), which used a ~500K SNP chip, established the current standards of modern GWAS design, including best practices for data collection, quality control and statistical considerations.

1.3.2 GWAS framework

The GWAS framework consists of two stages, a quality control (QC) and an association stage. The objective of the QC stage is to eliminate all data that could induce false positive associations. The association stage relies on the (logistic) regression framework, where the trait is regressed on the SNPs that survived the previous QC step. I will consider each stage in turn in the following two sections.

1.3.2.1 GWAS quality-control

For a GWAS to be successful, it is of crucial importance to eliminate spurious associations that could arise due to factors unrelated to the investigated genetic effect on the trait or disease. Due to the sheer number of tests (often in the millions) even a low rate of spurious associations may result in many thousands of false positives; therefore, a strict enforcement of data quality standards is necessary. The measured allele frequency in the study may be influenced by several data quality issues and population characteristics unrelated to the phenotype. Such quality issues may manifest either at the individual or at the marker level. The general protocol for common QC measures is as follows.

If no imputation is planned (which may be used to recover untyped or poor quality variants), individual QC should precede SNP QC to reduce the potential for removing SNPs due to poor quality samples. This step consists of the removal of individuals based on two criteria, indicators of overall low genotype quality or exhibiting unrepresentative allele frequencies with respect to the rest of the cohort. Data quality metrics deployed to infer the former are genetic and recorded sex discordancy, high missingness (>3-7%) and excess

(>3SD) heterozygosity (Anderson et al., 2010). The latter is defined as the proportion of heterozygous genotypes for an individual. Samples may also be excluded on the basis of high relatedness or being population outliers. Recently, there has been a shift in practices to keep more samples in the analysis belonging to this category by modelling relatedness/population stratification via the random effects term in a linear mixed effects model (Loh et al., 2018; Yang et al., 2014). The advantage of this approach is that the power of the study would be increased by a factor proportional to the additional samples allowed to remain. However, such joint analysis only corrects for genetic effects, and may not be robust against environmental confounders that may be correlated with genetic ancestry (Peterson et al., 2019).

Marker QC consist of eliminating variants that are most likely to be subject to errors that would bias allele frequencies and induce false positive associations. Commonly deployed steps include filtering markers with missingness above a certain threshold, such as 5%, or missingness substantially different between cases and controls for disease studies. Very rare variants may also be removed as they are more likely to be subject to genotype calling errors. Extreme deviations at a locus, those unlikely to be caused by selection acting on a deleterious allele, may be identified by performing hypothesis tests for the Hardy-Weinberg equilibrium (HWE). HWE tests evaluate the null hypothesis of expected genotype frequencies against the observed data, which may be performed via either a χ^2 test or a Fisher's exact test, for common and rare variants, respectively. The thresholds for HWE depend on the data and must be inspected on a case-by-case basis, but commonly employed thresholds range between $5 * 10^{-12}$ - $5 * 10^{-5}$. For case-control studies, HWE tests may be performed only in the subset of controls to rule out the potential for genuine associations to cause any deviations observed in HWE (Amos et al., 2017; Anderson et al., 2010). Finally, variants whose MAF is too low for a realistic chance of obtaining valid p-values may be removed to reduce the multiple testing burden. For studies that rely on imputation, an additional round of marker QC may be performed on the newly inferred SNPs based on metrics that evaluate the quality of imputation. One often used metric of imputation is the INFO score. The INFO score may take values between zero and one, which indicate poor or high confidence imputation, respectively (Marchini and Howie, 2010). Thus, post-imputation QC includes all the previous steps, together with the removal of any newly inferred variants with a low confidence imputation, such as those with an INFO score < 0.4 (Peterson et al., 2019; Popejoy and Fullerton, 2016).

Finally, at the end of the GWAS analysis, any putative associations must be re-examined closely to avoid unnecessary replication of false positives. Post-association QC steps may include the examination of the cluster plots of directly called genotypes. In the case of imputed markers, cluster plots for the LD proxies of the target variants that were directly

genotyped may be considered instead, together with the weighing of the evidence by the imputation quality score of the target SNP.

1.3.2.2 The GWAS model and statistical considerations

The basic GWAS test consists of a univariate regression of each SNP individually against the phenotype as

$$Y = G\beta_G + \mathbf{Z}\beta_Z + e, \tag{1.14}$$

$$Y = \sigma(G\beta_G + \mathbf{Z}\beta_Z + e), \tag{1.15}$$

where Y, G, $\mathbf{Z}$ and e denote the phenotype column vector, each SNP, a matrix of covariates and a random noise term, respectively. σ is the logistic function, defined as $\sigma(x) = 1/(1+e^{-x})$, and β_G and β_Z are the coefficients for the SNP and the covariates, respectively. Quantitative traits use eq 1.14, and binary traits use eq 1.15. The SNP coefficients are interpreted as follows. Each additional allele contributes a β_G level of additive change of either units of phenotype or a multiplicative change in odds ratio for quantitative or binary traits, respectively.

The very large number of performed tests, which typically range between ~500K and ~10mil SNPs, induces a substantial multiple testing burden. However, due to the LD between markers, the number of tests is actually lower than the number of SNPs investigated. The exact number of tests to be corrected for is based on the effective number of independently varying loci in the genome. Therefore, the 'genome-wide significance threshold' (corresponding to a per-study Type I error of 5%) has been determined by permutations, and is set between $5 * 10^{-8}$ and 1-$5 * 10^{-9}$. To former threshold was established for chip GWAS of European ancestry participants (Dudbridge and Gusnanto, 2008), and the latter more stringent threshold has been used more recently for WGS GWAS that may include rarer MAF variants or when the cohort includes individuals of diverse genetic ancestries (Pulit et al., 2017; Xu et al., 2014). This recent decrease in the significance threshold is motivated by the fact that including non-European ancestry individuals or testing lower MAF variants found in WGS data increases the effective number of independent loci.

SNPs that pass all aforementioned QC and multiple testing correction criteria are considered to be genuinely associated with the phenotype by tagging the causal variants via LD. One possible step after this initial GWAS is fine-mapping analysis, where the objective is to identify the most likely causal variant(s) in an associated locus (Spain and Barrett, 2015).

1.3.3 GWAS insights and recent trends

Most GWAS up to date involved participants of a predominantly European ancestry (86% up until 2018 (Mills and Rahal, 2019)). However, expanding recruitment to include more genetically diverse populations is expected to increase power to detect rare variants that are more frequent in those populations, together with the increasing of the applicability of any potential therapeutic interventions outside of Europe. Thus, one of the major trends in recent GWAS is the move to include individuals from a wider range of genetic ancestries in either meta or joint analyses (International Multiple Sclerosis Genetics Consortium et al., 2015; Peterson et al., 2019; Wojcik et al., 2019).

The emerging picture of the genotype-phenotype map produced by a decade of GWAS is that for most complex traits, a massively polygenic component dominates heritability. There is an ongoing discussion as of the nature of this component, whether there is a strong hierarchy where a few genes play a central role (the so called 'omnigenic' model) (Boyle et al., 2017) or if the relative importance of genetic variation is more evenly distributed (Wray et al., 2018). Another recent important insight that emerged is the relative importance of rare variants. In a recent study utilising WGS data by Wainschtein et al. (2019), it was shown that over half of the heritability of height and BMI were due to rare variants (a MAF of 0.0001 - 0.1) in low LD.

Finally, there is a shift towards more functional studies, where GWAS findings are subjected to functional follow-up experiments. However, this 'post GWAS era' is unlikely to mean the end of GWAS. On the contrary, motivated by insights on the massively polygenic architecture of most traits, and the continuously decreasing costs of genotyping and sequencing, the general trend of GWAS is towards an increase of both cohort size and density of coverage (Mills and Rahal, 2019). Another recent development is the pooling of cohorts into large scale meta-analyses, where some of the largest combined sample sizes have now exceeded one million individuals. Recent representative examples are the GIANT (Yengo et al., 2018) (~700K), PGC (Lee et al., 2019) (~720K) and COGENT (Lee et al., 2018) (~1.1mil) consortia. Current state of the art biobanks number around ~500K participants (such as the UKBiobank (Bycroft et al., 2017) or the FINGEN biobanks (FinnGen, 2020)), but this is set to increase in the near future into the millions. The currently ongoing USA based *"All of Us"* biobank will include over 1 million participants (The All of Us Research Program Investigators, 2019), and the UK's next generation biobank effort, the *"5 million genomes project"*, is expected to sequence 5 million individuals by 2023 (GEL, 2020).

1.4 Transcriptome-wide association study

GWAS have been successful in identifying marker-trait signals; however, interpreting these associations remain an ongoing challenge. The transcriptome-wide association study (TWAS) design was proposed to address this limitation by replacing individual variants with gene-level predictors, which are then related to phenotypic variance. Linking expression to disease may provide insights one step closer to the mechanism of effect that may then help to identify the effector genes and relevant cell types. As the majority of disease associated SNPs are located in the regulatory genome (Hindorff et al., 2009), the TWAS approach has greater potential to provide insights on the contribution of non-coding variants, and to identify targets for drug response (GTEx Consortium et al., 2015).

A key advantage of the TWAS approach is that it does not require expression information on the target cohort used for the association step. Instead, to obtain expression-trait associations, the genetically mediated parts of expression are 'imputed' by utilising a suitable reference panel. Initially branded as 'PrediXcan' (GTEx Consortium et al., 2015), TWAS was first applied to a range of immune related disorders in the Wellcome Trust Case Control Consortium studies. Here, it was found that in addition to confirming many known loci, this approach also identified novel risk genes (for example *DCLRE1B* for IBD). In a later study, Gusev et al. (2016) showed that the TWAS framework may be generalised to work from summary statistics, which has the added benefit of being able take advantage of publicly available data. In the same study, they further demonstrated the utility of TWAS on quantitative traits such as BMI and height, together with highlighting the strengths of the approach in linking association to function via mouse models.

1.4.1 TWAS framework

The TWAS framework consists of two main steps, the imputation of the transcriptome by generating a polygenic score for expression for each gene, and an association step (GTEx Consortium et al., 2015). These steps are summarised by the following two equations

$$\widehat{E_i} = \sum_{j}^{J} G_j^{eQTLi}, \widehat{\beta}_j^{eQTLi} \tag{1.16}$$

$$\widehat{Y} = \widehat{E_i}\widehat{\beta_i^E}, \tag{1.17}$$

where $\widehat{E_i}$ denotes the imputed expression for gene i in a particular a tissue and G^{eQTLi} and $\widehat{\beta}^{eQTLi}$ denote the J eQTL SNPs for this gene and their coefficients, respectively. SNPs may also be pre-filtered by using LASSO or Elastic net regularizers (GTEx Consortium et al.,

2015). Once all gene-level predictors of expressions are built, the phenotype ($\widehat{Y}$) is regressed on each of them separately (eq 1.17). This step is very similar to classical GWAS, the only difference is that here the coefficients estimated ($\widehat{\beta_i^E}$) relate to gene-level predictors rather than SNPs.

1.4.2 The potential benefits of the TWAS framework

A key benefit of TWAS is that it reduces the dimensionality of the dataset, as potentially thousands of SNPs may be summarised into a single, gene-level predictor. This may allow SNPs with smaller, but congruent effects, which are individually too weak to be detected, to contribute to the signal on the level of a gene. This also reduces the multiple testing burden, which even with the conservative Bonferroni correction, would be only ~$5 * 10^{-6}$ for a 5% Type I error rate (GTEx Consortium et al., 2015). Therefore, the TWAS approach may increase power to detect novel associations not accessible to the standard GWAS framework.

1.4.3 Limitations of TWAS

The main limitations of the TWAS method lie in the difficulty of distinguishing the origins of an expression-trait association. In addition to the sought after direct effect of a variant on expression levels, there are two additional alternative explanations. One alternative is that the expression driving variant may simply be in LD with another variant that is the true cause of the association. This problem may be further exacerbated if the true signal actually originated from a coding variant tagged by the regulatory marker in the model, as this would then lead the investigators to falsely infer that the effect involves gene regulation rather than protein alterations. Another alternative explanation for a TWAS association is pleiotropy, where a single variant may affect the trait both directly, as well as through modulating the expression levels. Zhu et al. (2016) proposed the following potential remedies to alleviate these problems. To distinguish pleiotropy from direct effect, they proposed to use Mendelian randomization (which conditions the phenotype on the variant's direct effect). To further distinguish linkage from pleiotropy, they also developed a method known as 'HEIDI'. This method is based on a heterogeneity test, where the null hypothesis is that all SNPs in a locus have the same effect on expression, if the true nature of the association originates from pleiotropy.

1.5 Protein burden score tests

The proteome-wide association study, or 'PWAS', is a recently proposed method (Brandes et al., 2019a) that aims to detect associations between protein-coding genes and phenotypic variance by utilising predicted protein function alterations. While the authors named their method 'PWAS', and there are some similarities to TWAS as they both incorporate external data to perform gene-based tests, in spirit, it is closer to a genome-wide weighted burden association test. Additionally, 'PWAS' does not involve measuring real protein quantities in tissues or cells; instead, it relies on aggregating the predicted functional effects of SNPs that jointly affect a protein coding gene. Therefore, to avoid confusion, from here onward I will be referring to this method as a 'protein burden test' and not as 'PWAS'.

1.5.1 Protein burden test method outline

Similarly to TWAS, the protein burden test consists of two steps, the generation of per-gene protein scores and an association step. One important difference between this and the TWAS framework is that this method does not require a reference panel for a specific tissue or expression profile. Instead, the protein burden association test relies on the predicted molecular consequence of individual variants. On one hand, this makes this approach one step further removed from biology than its TWAS counterpart. On the other hand, as the predicted protein function is common to all tissues, putative associations identified by this method may be more generally applicable.

1.5.1.1 Generating the protein burden scores

The first step in the protein burden test method is to quantify the impact of relevant variants on the function of the affected proteins using FIRM, a related machine-learning model that considers the proteomic context of each SNP (Brandes et al., 2019b) (also developed by the same group). The authors of this method have kindly agreed to share their generated scores for the imputed UKBB panel of variants which I have used for my analyses. The predicted effect score of a SNP is a value between zero and one, which represent complete loss of function and no functional effect, respectively. An important distinction is that FIRM is designed to quantify the damage of variants at the molecular, rather than on the clinical outcome level, which makes these scores more suitable for non-clinical quantitative traits such as height or BMI.

The tool offers two functions, $PWAS_D$ and $PWAS_R$, which aggregate the per-SNP FIRM scores and combine them with the genotyping data to produce per-gene predictors for

each individual for dominant and recessive gene-scores, respectively. These functions are summarised by

$$G_i^D = PWAS_D(\mathbf{X}, S, \mu_D, p_D), \tag{1.18}$$

$$G_i^R = PWAS_R(\mathbf{X}, S, \mu_R, p_R, q), \tag{1.19}$$

where $\mathbf{X}$ and S denote the genotype probabilities and the FIRM effect scores, respectively. The hyper parameters (μ, p and q) control the probability that the FIRM score for a SNP acts independently of other markers in a gene. This tool produces two scores, one for recessive (G_i^R), and another for dominant (G_i^D). In correspondence I exchanged with the tool's authors they recommended that a single score, representing the additive effect (G_i^A), may be obtained by averaging the dominant and recessive scores as: $G_i^A = (G_i^D + G_i^R)/2$.

1.5.1.2 Protein burden association tests

Similarly to TWAS, once the gene-level scores are obtained for each individual, a univariate OLS linear model is fit for each gene where the phenotype is regressed on each protein score as

$$Y = G_i^A \beta_i^A + e, \tag{1.20}$$

where β_i^A is the coefficient that quantifies the additive contribution of the protein score (G_i^A) to the phenotype.

1.5.2 Potential benefits of the protein burden test

The advantages of the protein burden test approach are also similar to the TWAS method. This framework also offers a reduction in dimensionality by aggregating many SNPs into a single gene-level predictor, in addition to giving different weights to potentially relevant predictors. Additionally, as this method is a burden test, variants with a lower MAF but with a congruent effect on the phenotype may still contribute to the aggregate signal.

1.6 Genetic risk prediction

1.6.1 Polygenic scores

GWAS have not only provided us with maps of genotype-to-phenotype associations (Visscher et al., 2017), but have also ushered in an era of availability of large-scale human genetic data. Instead of focusing on individual associations, a popular alternative use of this genetic data

is to estimate for a given individual the aggregate genome-wide propensity for a given trait. These quantities, depending on the field, are variously referred to as genetic profile scores, genetic risk scores, genetic merit, genomic best linear unbiased prediction or molecular breeding values (for animals) (Moser et al., 2009). In the field of human genetics they are most commonly referred to as polygenic risk scores (PRS) or polygenic scores (PGS), for disease and quantitative traits, respectively. Individual-level genetic risk prediction holds the promise to identify individuals at increased risk for monitoring, prevention, stratified treatment or lifestyle changes (Torkamani et al., 2018). From here on, for the sake of consistency, I will be using the term 'PRS' to refer to scores concerning both disease and non-disease phenotypes.

1.6.2 The origin of PRS

The origin of modern PRS in phenotype prediction may be traced back to two converging methodologies in the human and animal quantitative genetics literature. In the area of risk prediction in the field of human genetics, the number of variants considered from GWAS was incrementally expanded until it reached genome-wide coverage. In the field of agricultural science, pedigree derived estimates of kinship were replaced by relatedness based on molecular data for best linear unbiased prediction (BLUP) based breeding value estimation. In humans, one of the earliest successful attempts that demonstrated an improvement in risk prediction by considering multiple markers was a 13 SNP composite score to predict coronary heart disease risk. It was shown, that by jointly considering all markers that achieved genome-wide significance, this early PRS had a predictive power beyond any of the individual associations (Ripatti et al., 2010).

Parallel to the aforementioned developments in human genetics, early theoretical work by Meuwissen et al. (2001) showed that Henderson's equations (Henderson, 1950) for BLUP could be made substantially more accurate by considering dense genome-wide markers instead of expected kinship coefficients. These findings have subsequently led to improved genomic selection in a wide range of applications, such as for wheat yield (Crossa et al., 2010) and dairy production (Moser et al., 2009).

1.6.3 Current methods for building PRS

Most current methods for constructing a PRS fall into two broad (overlapping) categories, univariate regression of the phenotype against each SNP individually, and whole-genome regression based methods that consider the effect of all SNPs jointly on the phenotype.

To predict an individual's expected genetic propensity for a trait given their genetic background, the goal in both cases is to calculate a per-allele dosage effect for each marker, assuming an additive effect. The PRS may then be computed as a weighted sum of all considered risk alleles

$$E(\widehat{Y}|\mathbf{X}) = \mathbf{X}\widehat{\beta}_{SNP}, \tag{1.21}$$

where $\mathbf{X}$ is the genotype matrix for n individuals and p SNPs (0,1, or 2), $\widehat{\beta}_{SNP}$ is a column vector of the p estimated SNP effects, and $\widehat{Y}$ is a column vector of the n predicted PRS values. This formula is agnostic to where the $\widehat{\beta}_{SNP}$ come from, which may originate from either univariate or whole-genome regression based models.

1.6.3.1 Univariate regression based models

In typical GWAS data the number of individuals is less than the number of markers $(n < p)$; therefore, multiple-regression OLS is not possible as the $\mathbf{X}^T\mathbf{X}$ matrix is not invertible. Instead, GWAS rely on univariate, marginal regressions of the phenotype on each SNP individually. The per-SNP effects produced by this model represent the starting point for all methods in this category. After this initial step, the main consideration is the selection of which markers to include in the predictor. In recent years, many studies confirmed that the PRS' accuracy may be substantially improved if the selection criteria is relaxed, and SNPs with a lower than genome-wide significance level are allowed in the prediction model (Shi et al., 2009). Following on from this insight, a range of GWAS p-value thresholds is evaluated in cross-validation, and the threshold which is determined to have the highest predictive accuracy on the validation set is selected. As effect size estimates originate from a marginal regression model in a GWAS, increasing the number SNPs that contribute to the PRS can create the problem of redundant contributions of SNPs in LD that tag overlapping signals. Such redundant contributions may then decrease the predictive accuracy of the PRS. While it has been found that when the number of SNPs considered is large (>10K), this effect is mitigated for highly polygenic architectures (Kim et al., 2017), strategies that deal with this problem explicitly were developed that work by either LD pruning or modelling LD into the predictor. The simplest way to deal with redundant signal contribution by correlated SNPs is to remove markers that exceed a pre-specified pairwise LD threshold of, say, $r^2 > 0.2$, preferentially keeping markers with a lower p-value (Mavaddat et al., 2019). The resulting PRS building strategy is referred to as $P+T$ (pruning and thresholding).

There are several advantages that univariate regression based models have over basic whole-genome regression approaches. Estimating each marker effect separately can more easily accommodate SNPs with large effect. Additionally, genotype-level data is not required;

therefore, PRS may be built from the more widely available GWAS summary statistics, including meta-analyses of multiple GWAS.

Fine-tuning techniques that rely on the initial marginal regression. Recently, several more advanced methods have emerged that improve PRS accuracy by fine-tuning the above framework. Two such frameworks are step-wise regression and LASSO based approaches. A common first step for both of these methods is an initial filtering of SNPs based on GWAS p-values which is performed to reduce the number of markers considered. The outline of these two methods are summarised in the following paragraphs.

The step-wise regression approach starts by considering ~1Mb windows around each locus of associated SNPs in a series of forward regression models. This process sequentially adds SNPs with the lowest-pvalue until no more variants can be added below a pre-specified threshold. At the end, a joint model is built by re-estimating the effects of all SNPs that were selected in the previous step to generate the PRS (Mavaddat et al., 2019).

In contrast, LASSO based methods perform variable selection on a joint model of all SNPs that survived the initial p-value filter. LASSO's shrinkage parameter for this is usually determined by cross-validation (Choi et al., 2018; Mavaddat et al., 2019). Recently, it was also shown that it is possible to adapt the LASSO framework to build PRS from summary statistics (Mak et al., 2017).

As both LASSO and step-wise regression based approaches fit joint models (of the filtered variants from a GWAS source), their benefits and drawbacks are also similar to whole-genome regression models (discussed in detail in 1.6.3.2). On the positive side, estimated marker effects are conditioned on the rest of the predictors in the model; hence, their coefficients may be more accurately estimated. On the negative side, these approaches require larger sample sizes and are also more computationally demanding.

1.6.3.2 Whole-genome regression based models

Originating from the animal breeding literature, methods in this category aim to model the phenotype from the genotype by considering all SNPs simultaneously. These prediction models are fit in two stages that I will describe below.

At the first stage, a realised genetic relatedness, or kinship, matrix ($\mathbf{K}$) is produced, and the additive genetic and noise variances (σ_a^2 and σ_e^2) are obtained as I previously described by equations 1.10 and 1.11. At the second stage, depending on the method, marker effect sizes may be estimated in two different ways. In the case of the linear mixed-effects model (LMM) framework, the genetic component of the training-set phenotypes (the breeding values) are estimated, and then the individual SNP-effects are back-calculated from this. Assuming no covariates, the molecular breeding values g (conceptually analogous to a PRS), are estimated

as (Morota and Gianola, 2014)

$$g = \mathbf{K}(\mathbf{K} + \frac{\sigma_a^2}{\sigma_e^2}\,\mathbf{I})^{-1}y, \tag{1.22}$$

and then the individual marker effect estimates ($\widehat{\beta}_{SNP}$) are back-calculated from g, via the following equation (Morota and Gianola, 2014):

$$\widehat{\beta}_{SNP} = \mathbf{X}^T(\mathbf{X}\mathbf{X}^T)^{-1}g. \tag{1.23}$$

In contrast, ridge-regression(RR) based models estimate marker effects directly via the following equation (de Vlaming and Groenen, 2015):

$$\widehat{\beta}_{SNP} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T y. \tag{1.24}$$

Depending on the shrinkage parameter (λ), RR lies between univariate OLS, which considers each SNP separately (high λ), and multiple-regression OLS that considers all SNPs jointly (low λ). At $\lambda = 0$, equation 1.24 actually reduces to the OLS estimate

$$\widehat{\beta}_{SNP} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T y. \tag{1.25}$$

Therefore, depending on the value of λ, the attractive property of OLS that controls for LD between markers is partially preserved (de Vlaming and Groenen, 2015). Assuming no covariates, when the shrinkage parameter is set to the ratio of the additive genetic and noise variances ($\lambda = \sigma_a^2/\sigma_e^2$), it has been shown that the $\widehat{\beta}_{SNP}$ from a RR are identical to the effect size estimates from a LMM (de Vlaming and Groenen, 2015). In both cases, once the individual $\widehat{\beta}_{SNP}$ are obtained, these may be used in eq (1.21) to produce the final PRS. The first widely used implementation of this framework applied to human complex trait genetics was by the software GCTA which was published in 2010 (Yang et al., 2011). This method essentially translated the theoretical work by Meuwissen and others in the animal breeding literature, and applied it to human PRS and heritability estimation.

Basic whole-genome regression methods assume an infinitesimal model, where all markers are assumed to have an infinitesimally small effect on the phenotype, each from the same distribution:

$$Y \sim N(0, \mathbf{K}\sigma_a^2 + \mathbf{I}\sigma_e^2). \tag{1.26}$$

For traits where a substantial amount of heritability may be due to fewer SNPs with larger effect sizes, these may be incorporated into the model in several ways. Such effects may be modeled as fixed effects in a LMM, or alternatively, more advanced models that allow

an uneven distribution of heritability may be employed. Examples of methods that can accommodate such effects include the extended GCTA-LDMS (The LifeLines Cohort Study et al., 2015) and the LDAK prediction models.

LDAK's Multi-BLUP (Speed and Balding, 2014) takes into account regional heritability by allowing different areas of the genome to contribute disproportionately to the predictor, owing to the different effect-size distributions assumed for each z region of K:

$$Y \sim N(0, \mathbf{K}_1\sigma^2_{a1} + \mathbf{K}_2\sigma^2_{a2} + \cdots + \mathbf{K}_z\sigma^2_{az} + \mathbf{I}\sigma^2_e). \tag{1.27}$$

In conclusion, while they do require genotype-level data to be fit, whole-genome regression based models have two advantages, they control for LD and work well when the number of individuals is far less than the number of SNPs ($n << p$).

1.6.3.3 LDpred

To date LDpred was the method of choice for the studies that most successfully demonstrated the utility of PRS (Khera et al., 2018; Lee et al., 2018). The main reason for LDpred's success is that is a powerful approximation of whole-genome regression PRS methods, but one that only requires the more readily available summary statistics from univariate regression models and an LD reference panel matching the ancestry of the training set (Vilhjálmsson et al., 2015). This method's reliance on summary statistics also offers the additional benefit of reduced computational complexity, as LDpred's resource requirements scale linearly with the number of markers, rather than the quadratic scaling of whole-genome regression methods that make the latter infeasible for larger cohorts. This method offers two models, *LDpred-inf* and *LDpred-p* that assume an infinitesimal and a non-infinitesimal genetic architecture, respectively.

LDpred-inf obtains estimates of SNP effects ($\widehat{\beta_{inf}}$) in ~2Mb tiled windows via the following equation

$$E(\widehat{\beta_{inf}}|\beta_{GWAS}, \mathbf{D}) = \left(\frac{M}{nh^2}I + \mathbf{D}\right)^{-1}\beta_{GWAS}, \tag{1.28}$$

where $\mathbf{D}$ denotes the LD matrix (sourced from a reference panel), β_{GWAS} are the univariate GWAS SNP effect estimates, M is the number of SNPs in the model, n is the number of individuals in the original GWAS, and finally, h^2 is an estimate of SNP heritability. Assuming an infinite sample size ($\frac{M}{nh^2}I \approx 0$), the intuition behind this formula may be understood as simply scaling the GWAS SNP estimates by their LD with other markers. The SNP effect estimates from this model may be analytically computed, and in practice, they are a close

approximation of BLUP estimates that I previously described in section 1.6.3.2 (Vilhjálmsson et al., 2015).

In the *LDpred-inf* model SNP effects are drawn from the same distribution that assumes that all considered markers are causal; however, they are continuously weighted based on local LD. This is an improvement over *P+T* as LDpred explicitly models the effect of LD, where nearby SNPs are allowed to contribute overlapping effects. This strategy offers an advantage over pruning, as unless the overlap in signal is perfect, pruning would result in the discarding of information to the degree of non-overlap. On the other hand, LDpred's continuous weighting scheme allows all associated SNPs to contribute to the final score, but at a weight proportionate to the uniqueness of their signal.

The other main model of this method, *LDpred-p*, caters for the more realistic scenario of non-infinitesimal genetic architectures, where only a subset of variants, a causal fraction p, is expected to contribute. This is accomplished in LDpred via modeling SNP effects as drawn from a Gaussian mixture prior of

$$\beta_p \sim \begin{cases} N\left(0, \frac{h^2}{Mp}\right), & \text{with probability p} \\ 0, & \text{with probability (1 - p),} \end{cases} \tag{1.29}$$

where β_p denotes the true SNP effects and p is the fraction of SNPs believed to be truly associated (a quantity that may be determined via cross-validation). An analytical solution to the posterior mean of SNP effects is not tractable here; therefore, LDpred obtains these via a numerical approximation using a Gibbs sampler. The Gibbs sampler is a Bayesian approach that approximates a posterior multivariate distribution of the variables of interest by iteratively sampling them conditioned on their current values. The outline of this iterative solution is as follows (Privé et al., 2020). At each iteration, each SNP's residualized marginal effect β^j_{resid}, which is the unique contribution of SNP j conditioned on other nearby markers in LD, is obtained by

$$\beta^j_{resid} = \beta^j_{GWAS} - \beta^T_{-j} D_{-j,j}, \tag{1.30}$$

where β_{-j} and $D_{-j,j}$ denote column vectors for all variants without the jth SNP for the current iteration's SNP effect estimates and the pairwise LD between SNPs, respectively. After a pre-specified number of iterations, LDpred obtains the posterior mean for SNP j via

$$E(\widehat{\beta^j_p} | \beta^j_{GWAS}, \mathbf{D}) = \frac{\bar{p}_j \beta^j_{resid}}{1 + \frac{Mp}{nh^2}}, \tag{1.31}$$

where $\widehat{\beta_p^j}$ denotes the final LDpred effect estimate for SNP j and $\bar{p}_j$ is the probability of SNP j being truly associated. Once again, the intuition behind this formulation may be grasped by assuming an infinite sample size, which would then reduce this to $\beta_{resid}^j \bar{p}_j$; that is, the GWAS SNP's unique contribution conditioned on LD, weighted by the estimated probability that SNP j is truly associated. This non-infinitesimal *LDpred-p* model may be thought of as conceptually similar to LDAK's Multi-BLUP (Speed and Balding, 2014), as this model can set the contribution of non-associated SNPs to exactly zero (unlike *LDpred-inf*). As *LDpred-p* still allows for SNPs in LD that represent overlapping signal to contribute appropriately (as opposed to the hard threshold employed by *P+T*), it is often the preferred choice for building PRS in practice.

There are also a few limitations of LDpred. For a mixed ancestry prediction test set a suitable LD reference panel may be unavailable. Additionally, as the input for the method are SNP summary statistics originating from univariate additive models, LDpred cannot be used to incorporate non-linear genetic effects into its PRS.

1.6.4 Recent applications of PRS

Owing largely to the ever increasing cohort sizes, the accuracy of phenotype prediction has been improving substantially over last few years. This development has wide-ranging potential applications from prediction of complex behavioural traits, to disease and disease sub-type classifications.

In the domain of behaviour genetics, in a recent study on educational attainment, with a combined sample size of 1.1 million individuals, a PRS was built that explained ~13% of the phenotypic variation out of a h^2_{SNP} of ~15% (Lee et al., 2018). In the field of medical genetics similar improvements have been observed. For individuals in the extreme tails of the distribution, the predictive utility of these PRS have been recently shown to be comparable to that of monogenic mutations for some common disorders, such as coronary artery disease and type 2 diabetes (Khera et al., 2018). When combined together with conventional clinical predictors, PRS hold the promise to select the subset of individuals at the highest risk at a population level (Torkamani et al., 2018). While at this stage this is still hypothetical, it may be soon possible for these patients to benefit from improved interventions, screening and life-style modifications to alter their disease course outcomes.

PRS may also be used to help to elucidate disease biology by stratifying patients into subgroups based on genetic heterogeneity. For example, PRS demonstrated potential for patient stratification in a study involving breast cancer, by predicting the risk for specific breast cancer subtypes via utilising subtype specific marker effect sizes (Mavaddat et al.,

2019). In another study, PRS demonstrated that the genetic aetiology of inflammatory bowel disease substructure forms a continuum that ranges from ulcerative colitis through colonic Crohn's disease to ileal Crohn's disease (Cleynen et al., 2016).

Finally, the aggregation of the effects of many variants into a single score, the basic motivation behind PRS, has also been instrumental in the development of the TWAS framework which I have covered in detail previously in section 1.4.

1.6.4.1 Limitations

The most severe limitation of current PRS is a reduction in the expected prediction performance due to differences between the panel that the PRS was trained on, and the target cohort for which the PRS is intended to be evaluated on. So far, differences between training and test sets that impact PRS performance have been identified in two areas, genetic ancestry and population characteristics.

Populations with different genetic ancestries exhibit divergent MAFs and LD patterns. This is a challenge for PRS, as an important factor in statistical power to detect GWAS signal is MAF; thus, relevant loci may not be detected between ancestries if the MAF spectra differ beyond a certain degree. Also, PRS aggregate signal across many variants, assuming that associated loci are suitable proxies for the latent causal variants due to LD. However, LD patterns may be specific to the GWAS training set, and may not tag the same causal signals should LD differ substantially between populations. Therefore, the less well matched the training and test sets are on MAF and LD, the lower the transferable true association signal would be, and ultimately, the lower the expected performance of PRS would become. Recent examples for this are PRS trained on European ancestry reference panels for educational attainment and height that explained ~11% and ~10% population variance in a held out test set of the same ancestry, respectively. However, these very same PRS only explained ~3% and ~1.6% variance, respectively, for populations with a non-European ancestry (Lee et al., 2018; Martin et al., 2019).

PRS are also sensitive to even more subtle variations in mismatches of population characteristics between training and test sets. Recently, it was shown that within the same genetic ancestry, stratification by age, sex and socio-economic status also had a substantial negative impact on PRS performance. For example, the accuracy, evaluated by r^2 (squared correlation) between the phenotypes observed and those predicted by the PRS, for diastolic blood pressure was ~1.3x greater in females than in males, and an educational attainment PRS had less than half the predictive accuracy for individuals in the lowest socio-economic status than those in the highest socio-economic status (Mostafavi et al., 2020).

Finally, a common limitation of all PRS generation methods surveyed so far is that they rely on a linear model that considers additive effects only; thus, seek to model the phenotype as a linear combination of genotypes and their estimated effects. Therefore, any non-additive genetic variation would not be accounted for; hence, the PRS predictive accuracy may fall short of their maximal potential.

1.6.5 Genetic prediction incorporating non-additive effects

All of the PRS building methods examined so far were limited to consider additive-effects only. However, methods already exist that may extend these predictors by incorporating non-additive genetic components into the PRS without explicitly assessing individual epistatic effects.

(Ridge-)BLUP based models operate by essentially regressing phenotypic similarity on genotypic similarity. In the case of classical (Ridge-)BLUP, this genotypic similarity is represented by the pairwise relatedness values in an additive kinship matrix. Therefore, in models that extend this by considering non-additive effects, the additive kinship matrix is replaced by a non-additive kinship matrix, which may be readily obtained from additive kinship matrices. For example, to generate non-additive pairwise relationship values for the dth order of interactions, the following formula would be applied (Jiang and Reif, 2015)

$$\mathbf{K}^{\#d} = \mathbf{K_1}\#\mathbf{K_2}\#\ldots\#\mathbf{K_d}, \tag{1.32}$$

where # is the Hadamard product operator. In other words, the pairwise additive relationship values are element-wise raised to the dth power. It is also possible to incorporate all conceivable interactions between the p SNPs. The elements of such an infinitesimal epistatic kinship matrix may be generated by the Gaussian kernel function as (de Vlaming and Groenen, 2015):

$$k(x_i, x_j) = exp\left[\frac{-||x_i - x_j||^2}{h}\right], \tag{1.33}$$

where x_i and x_j are individuals i and j, $||.||$ denotes the norm in the Euclidean space and h is a bandwidth parameter that controls the rate at which the weight of interactions decays, with smaller values corresponding greater importance given to higher-order interactions (Endelman, 2011). Methods that implement this framework include extended E-GLUP and reproducing kernel Hilbert space (RKHS) regression (also known as kernel-ridge regression (KRR) (de Vlaming and Groenen, 2015)). By utilising such kinship matrices it is possible to obtain a BLUP from an infinite number of predictors (interactions); however, it is no longer possible to obtain individual marker effects, and therefore equations (1.21) and (1.24) are

no longer applicable (de Vlaming and Groenen, 2015). Instead, to obtain predictions for out-of-sample individuals ($\widehat{Y}_2$), the model fit is slightly altered, and phenotype predictions are obtained by an equation very similar to (1.22) (de Vlaming and Groenen, 2015)

$$\widehat{Y}_2 = \mathbf{K}_{21}(\mathbf{K}_1 + \frac{\sigma_a^2}{\sigma_e^2} I)^{-1} Y, \tag{1.34}$$

where $\mathbf{K}1$ is the same kinship matrix as before (genetic similarity between the training set individuals) and $\mathbf{K}_{21}$ represents the genetic similarity between the out-of-sample and the training set individuals.

In human genetics such models have been scarcely utilised (Weissbrod et al., 2016). However, in agricultural applications studies have already shown that such methods may outperform additive models for daily weight gain in pigs (Su et al., 2012) and yield in maize breeding (Crossa et al., 2013).

1.7 Neural-network based methods

1.7.1 The origins of neural-networks

Neural networks (NN) are a machine-learning prediction framework loosely inspired by how biological neurons function (LeCun et al., 2015). Their main use is to build prediction models from large datasets where complex, non-linear relationships between the input features contribute substantially to the outcome. Subsequent sections describe more details on their technical aspects, but first, a brief history of NNs is provided here.

Initially proposed by McCulloch and Pitts (1943), over the next 40 years the theory of NNs were developed by scientists working in vastly different fields, frequently unaware of the contributions of their peers. The first NN capable of learning was created by the American psychologist Rosenblatt (1958) and the first multi-layer networks originated in 70s. Ivakhnenko, a Soviet mathematician, created networks that went as deep as eight layers in 1971 (Ivakhnenko, 1971). The original version of the backpropagation algorithm, not specifically intended for NNs, was derived around at the same time by Werbos and John (1974). Finally, inspired by the work of neuroscientists Hubel and Wiesel (1962), the earliest version of the convolutional NN (CNN) was invented by Japanese computer scientists Fukushima and Miyake (1982). Thus, by the early 80s most of the main algorithmic ingredients existed; however, NNs remained in relative obscurity until several decades later, the early 2010s. Their recent resurgence was brought on by the alignment of all the separate

components previously described, together with the orders of magnitude of increase in training data and computational processing power available (LeCun et al., 2015).

Today, algorithms based on the NN framework are an intrinsic part of many aspects of our lives, including state of the art image recognition tasks (Zhao et al., 2019), self-driving cars (Bojarski et al., 2016), movie recommendation systems (Zhang et al., 2019) and numerous applications in the field of biomedical sciences (Ching et al., 2018).

1.7.2 What are neural-networks exactly?

NNs derive their ancestry primarily from regression-like methods (Schmidhuber, 2015). In fact, a single neuron NN is equivalent to logistic regression (Schumacher et al., 1996). Therefore, I will explain their mechanism in relation to logistic regression, starting from a single neuron and extending it step-by-step, to arrive at a more complete NN, up to the complexity that I will be using later in this work.

Consider the following equation that describes logistic regression

$$Y = \sigma(\mathbf{X}\beta_X), \tag{1.35}$$

where $Y \in \mathbb{R}^n$ is a column vector of binary outcomes, $\mathbf{X} \in \mathbb{R}^{n \times p}$ is a design matrix of n observations and p input features and $\beta_X \in \mathbb{R}^p$ is its corresponding coefficient. To improve the clarity of the notation, I omit a separate intercept term which is assumed to be included as a column of ones in $\mathbf{X}$ with a corresponding element in β_X. Finally, σ is the logistic function defined as

$$\sigma(x) = \frac{1}{1 + exp(-x)}. \tag{1.36}$$

In NN literature many standard statistical terms are known by alternative names which I will introduce as they arise. For example, the logistic function is referred to as the sigmoid activation function, the intercept is called the bias and the model coefficients are known as weights. While the sigmoid function is non-linear in a sense that the linear predictor is transformed into the range of zero to one, the only non-linearity it may represent is the logistic function itself. To turn (1.35) into a NN two changes are needed. First, another neuron (another logistic regression) is added that obtains its input from the first neuron's output as

$$Y = \sigma(\sigma(\mathbf{X}\beta_X)\beta_h). \tag{1.37}$$

Here, the first logistic regression (neuron) becomes the hidden layer ($\sigma(\mathbf{X}\beta_X)$), whose output, termed feature/activation maps or representations, is transformed by the other neuron to

produce the final result with the learned scalar coefficient β_h. A 'network' like the above with two neurons would have very limited capacity to learn; thus, to turn this into a real NN, one final change is required:

$$Y = \sigma(\sigma(\mathbf{XW_1})W_{out}). \tag{1.38}$$

Here, I made two substitutions. I replaced β_X with a matrix of weights $\mathbf{W_1} \in \mathbb{R}^{p\times a}$, and replaced β_h with $W_{out} \in \mathbb{R}^a$, which turned the latter into a column vector. This change expanded the hidden layer's width by the addition of a number of neurons that occupy the columns of $\mathbf{W_1}$. Thus, the changes so far may be thought of as adding multiple logistic regressions that learn from the original input $\mathbf{X}$, which is first transformed into a space of $\mathbb{R}^{n\times a}$; and finally, this is passed forward to the output logistic regression for another round of non-linear transformation to generate the final prediction.

The representational capacity of a NN, the complexity of the function that the model may learn, grows exponentially with the number of neurons (LeCun et al., 2015). It was shown that even a simple, single hidden layer NN like the above with a sufficiently large a, may already approximate any function (Cybenko, 1989). In practice however, NN models are extended in depth and not in their width. The reason for this is that empirically, deeper architectures are faster to train, generalise better and deeper layers may learn higher-level abstractions that makes them easier to interpret (Bengio et al., 2007). Thus, adding k hidden layers completes the formula for a basic, fully-connected NN (FNN) as

$$Y = \sigma_k(\ldots\sigma_2(\sigma_1(\mathbf{XW_1})\mathbf{W_2})\ldots W_k). \tag{1.39}$$

Depending on the activation function of the final output neuron, this model may be used for different types of prediction tasks. If σ_k is the sigmoid function, it performs binary classification. If σ_k is the identity, then the NN may be used for regression tasks. Finally, for multi-class classification problems the *'softmax'* function is used which is described by

$$softmax(x) = \frac{exp(x)}{\sum_i^J + exp(x_i)}, \tag{1.40}$$

which produces probability estimates for J classes for a given observation x.

Once the number of layers get sufficiently deep, NN models may be termed 'deep learning'. There is no consensus as of when exactly a NN model constitutes 'deep learning' and when it remains 'shallow learning'; however, there is an ongoing trend towards increasing depth, with some NNs having reached a depth of over a 1000 layers (He et al., 2016). Whether such depth is truly necessarily is still subject to debate (Zagoruyko and Komodakis, 2016); nevertheless, currently used NN models are almost always deeper than a single layer.

1.7.3 How are neural-networks fit?

To generate the output of the network as I described so far is straightforward: it is a series of matrix multiplications interspersed by application of element-wise non-linearity. This sequential transformation of the input features towards an output is termed *forward propagation*.

To allow the model to obtain the appropriate parameters to predict the output accurately, the network weights are learned from the data by training via the application of the *back-propagation algorithm* (Werbos and John, 1974). As a NN is essentially a series of nested functions of functions, the backpropagation algorithm obtains the errors via the application of the chain rule of differentiation. Thus, the partial derivative of the first hidden layer for the NN described in 1.39 is obtained by

$$\frac{\partial}{\partial \mathbf{W_{in}}} = \frac{\partial}{\partial W_{out}} \left(W_{out} \frac{\partial}{\partial \mathbf{W_k}} \right) \left(\mathbf{W_k} \frac{\partial}{\partial \mathbf{W_{k-1}}} \right) \dots \left(\mathbf{W_2} \frac{\partial}{\partial \mathbf{W_1}} \right) \mathbf{X}. \tag{1.41}$$

These partial derivatives, usually referred to as gradients, are calculated by automatic differentiation algorithms by most implementations, such as those found in Tensorflow or Pytorch (Paszke et al., 2017). Each layer's contribution to the total error (same quantity as the residuals in statistical terminology) is a product of its gradient and the weights of a layer deeper to itself, if there was one; thus, the equation 1.41 may be rewritten as

$$\frac{\partial}{\partial \mathbf{W_{in}}} = \delta_{out} \delta_k \delta_{k-1} \dots \delta_1 \ \mathbf{X}. \tag{1.42}$$

Finally, the total weight delta (Δ) for layer h_i is calculated by

$$\Delta_i = \left(\delta_i^T \, h_{i-1} \right)^T, \tag{1.43}$$

where h_{i-1} is the output (activation map) of the previous layer, or in the case of the first hidden layer, the input $\mathbf{X}$ itself. The backpropagation algorithm allows all updates for a NN to be obtained highly efficiently, as the expensive calculation of the derivatives only need to be performed just once per iteration.

1.7.3.1 Stochastic gradient descent

Similarly to logistic regression, because of the non-linearity, NNs are fit iteratively. However, because of the computational requirements for most practical applications, the model does not fit the entire training set at once. Instead, small random subsets, termed 'mini batches', are passed through the network to obtain incremental changes. These are then scaled by η,

the learning rate parameter, to obtain the final per-iteration changes to the weights as

$$\Delta \mathbf{W_i} = -\eta \Delta_i', \tag{1.44}$$

where Δ_i' is the mini-batch sized version of delta obtained in 1.43, and $\Delta \mathbf{W_i}$ is a matrix of updates for the iteration. This latter is then element wise added to the corresponding weight matrix of layer i to complete the iteration. This entire process is termed stochastic gradient descent or SGD (Robbins and Monro, 1951). An epoch is defined at the time point when the network has processed all mini batches once; hence, the training time for NNs is measured in epochs.

1.7.3.2 Weight initialisation

As all neurons are defined identically and trained via the same algorithm on the same data, one may ask, how come they do not end up learning the same parameters? The answer to this lies in weight initialisation, as each neuron is initialised differently via random weights.

To cover all possible ways to initialise neurons is not my intention here; however, there are two themes common to most of them. One commonality is that the starting values are drawn from truncated Gaussian distributions, and the other is that the variance of these distributions are inversely proportionate to the connections of the given neuron (which is expected to keep neuron variances similar between layers). The truncation operation is employed as when a large number of parameters are randomly initialised, a small fraction of them may be assigned very low values ($> 2 * SD$), which would then cause those neurons to respond very slowly to training.

One of the most popular weight initialisation methods is HE initialisation scheme (He et al., 2015) which scales the *sd* of the weight distributions as

$$sd(\mathbf{W_i}) = \sqrt{\frac{2}{\#in_i}}, \tag{1.45}$$

where $\#in_i$ is the number of input connections for layer i, which is in turn equivalent to the number of neurons in layer $i-1$.

1.7.4 Advanced neural-network concepts

A common extension to the basic NN framework I described so far, which I will be referring to when describing the work of others but not use in my own analyses, are convolutional neural-networks (CNN). I cover CNNs in detail in Appendix 2 in section B. However, in the

next sections I will describe a few more advanced concepts that are relevant to my own work which include different optimizers, activation functions and special layer types that facilitate regularization.

1.7.4.1 ADAM optimizer

The original SGD optimizer I described in 1.44 is frequently substituted by more elaborate algorithms that yield superior results at a lower number of epochs under most circumstances. The main innovation of these optimizers is that they apply per-parameter adaptive learning rates that consider past updates via momentum. The most popular of these, ADAM (Kingma and Ba, 2014), modifies the standard SGD equation (1.44) to

$$\Delta \mathbf{W_i} = \eta * \frac{\mathbf{M}'}{\sqrt{\mathbf{V}'}} + \varepsilon, \tag{1.46}$$

where each term above is defined as:

$$\begin{aligned}
\mathbf{M} &= f * \mathbf{M} - f * \Delta_i, \\
\mathbf{V} &= \gamma * \mathbf{V} - \gamma * \Delta_i^2, \\
\mathbf{M}' &= \frac{\mathbf{M}}{1 - f^{t+1}}, \\
\mathbf{V}' &= \frac{\mathbf{V}}{1 - \gamma^{t+1}}.
\end{aligned}$$

$\mathbf{M}$ and $\mathbf{M}'$ are the momentum and its bias corrected estimates, respectively, $\mathbf{V}$ and $\mathbf{V}'$ are the second moment of the weight derivatives and its bias corrected estimates, respectively. f and γ are friction hyperparameters which apply decay to the aforementioned two variables. t is the current epoch, which is used to scale up the learning rates in the first few epochs (as f and γ are both < 0; thus, in later epochs $\mathbf{M}'$ and $\mathbf{V}'$ tend to $\mathbf{M}/1$ and $\mathbf{V}/1$). Finally, ε is a small value added for numerical stability.

1.7.4.2 Layers that address the vanishing gradient problem

While the sigmoid activation function theoretically already enjoys the universal approximation property (Cybenko, 1989), in practice, it suffers from the vanishing gradient problem. The vanishing gradient problem arises, as the consecutive transformations of the input cause gradients to become smaller and smaller towards the input layer, which then produce correspondingly diminishing updates to the model. To address this issue, several strategies were invented which are reviewed in next two sections.

1.7.4.3 ReLU and batch normalization

The currently preferred way to apply non-linearity to a NN is via the Rectified Linear Unit, or '*ReLU*' function (Glorot et al., 2011), which is described by

$$ReLU(x) = max(0,x). \tag{1.47}$$

This function sets all negative input values of x to 0, otherwise it returns the original input. The reason why the *ReLU* eliminates the vanishing gradient problem is that as the function is nearly linear, it does not saturate; thus, it yields a derivative of either zero or one.

A frequently deployed complementary strategy is the usage of a layer type known as the *batch normalization layer* (Ioffe and Szegedy, 2015). This layer scales the output (x) of each layer to have a zero mean and a unit variance by

$$x_N = \frac{x - \bar{x}}{sd(x)}, \tag{1.48}$$

where $\bar{x}$ and $sd(x)$ are the mean and standard deviation of the mini batch output of the preceding layer, respectively. Then, before passing it forward, the batch normalization layer also shifts its output via a linear regression as,

$$BN(x) = \beta x_N + \gamma, \tag{1.49}$$

where β and γ are the regression's coefficient and intercept terms, respectively. These latter are hyperparameters that are estimated via the usual application of the backpropagation algorithm.

The batch normalization layer also improves training by addressing the internal covariate shift problem. To clarify, without batch normalization, the input distribution for each hidden layer would change with each iteration which would force each hidden layer to continuously adapt to its changing inputs, making training less effective.

1.7.4.4 SELU

Showing particular promise to FNNs, an alternative to the ReLU followed by batch normalization paradigm, is the application of a '*SELU*' activation layer (Klambauer et al., 2017). This is an activation type that accomplishes both of the aforementioned layers' goals in a single step by

$$SELU = \lambda \begin{cases} x, & \text{if } x > 0 \\ \alpha e^x - \alpha, & \text{if } x \leq 0 \end{cases}, \tag{1.50}$$

where α and λ are constants set to ~1.673 and ~1.051, respectively.

1.7.4.5 Regularization of neural-networks

Given the non-linearity inherent in their nature, NNs are particularly susceptible to overfitting. To reduce the potential for this problem, several different strategies were developed. Some of the most popular these strategies are the application of L1/2 norms, early stopping and dropout layers. A brief review is provided of each in the next three sections.

1.7.4.6 L1 and L2 norms

A basic way to reduce the scope for overfitting is by the addition of an L1 or L2 norm to the loss function of the NN model, usually referred to as 'weight decay' in NN literature. The application and behaviour of these norms are identical to how one would employ them in standard linear models. Their effects are also similar, the layers onto which the L1 or L2 norms are applied to acquire properties similar to LASSO or Ridge regression, respectively.

1.7.4.7 Early stopping

Early stopping is a simple yet effective technique to reduce the potential for overfitting (Prechelt, 1998). The mechanism of this technique is as follows. NNs are routinely trained by utilising both a training set and a validation set. The model is trained first until a pre-specified epoch, during which its predictive performance is recorded on both the training and the validation set. After the training completes, the NN performance is evaluated on the validation set retrospectively, and the epoch after which the model stopped improving on the validation set is selected as the ideal number of epochs to train.

1.7.4.8 Dropout layer

Dropout is a NN-specific technique that emerged recently that achieves effective regularization with many attractive properties (Srivastava et al., 2014). Considering the output z_i of neuron i of a hidden layer, the dropout layer is applied by

$$z_i' = \begin{cases} 0, & \text{with probability } p \\ \frac{z_i}{1-p}, & \text{otherwise} \end{cases}, \tag{1.51}$$

where z_i is the output of hidden layer i and p is the probability specified for dropout. In the case the neuron remains active for the given training iteration, its output z_i is scaled by $1/(1-p)$ to ensure the same expected value for the overall layer output. The reason why

dropout achieves regularization is that it reduces the co-adaptation between neurons between different layers, which is a condition where a neuron relies on a specific pattern in the output from the previous layer.

1.8 Thesis objectives

The overarching objective of this work is to identify non-linear genetic effects that influence phenotypic variance. Therefore, the hypothesis pursued is that there is a substantial non-linear polygenic component to complex traits, which I hope to infer either directly or indirectly using traditional statistical approaches and the neural-network framework. The chapters are conceptually organised along an axis of increasing complexity of the effects that I seek to infer, which grow from additive effects in Chapter 2, through to two-way epistasis in Chapter 3, up to higher-order interactions in Chapter 4.

The GWAS quality control and statistical methodology framework that I reviewed in section 1.3.2 is applied in Chapter 2, where I employ these strategies to prepare datasets for further analyses and also to increase the confidence in my subsequent results. Polygenic scores and the LDpred tool that I rely on in Chapters 2 and 3 to build prediction model baselines and also to build gene-level predictors, were covered in section 1.6. The two gene-level approaches, TWAS and protein scores, that I described in sections 1.4 and 1.5, respectively, are deployed in Chapter 3. The regression model with a term for interaction that I introduced in section 1.1.7 is applied in Chapter 3, where I use it to evaluate the evidence for statistical epistasis in both SNP and gene-level data, as well as across these domains. Finally, the neural-network framework, which I reviewed in section 1.7, is applied in Chapter 4 in an attempt to infer epistasis on the same datasets that I prepared in Chapter 3.

Chapter 2

Additive models and common quality-control steps

2.1 Chapter 2 outline

The work presented throughout this thesis makes use of some the largest datasets in the field of human genetics. In this chapter I perform quality control on these key datasets, which is of crucial importance as I will rely on the same data in all subsequent chapters as well.

The technical details of the UK Biobank and IBD datasets are described in section 2.2. The quality control and filtering protocol I used to process my datasets are detailed in section 2.3. Section 2.4 describes the strategy for organising my datasets into training, validation and test sets, and section 2.5 details the additive models I built that were used in comparisons against publicly available results. I found that my data QC efforts were successful in recovering the main association signals as compared to relevant studies from the literature; thus, I determined that my data was of a sufficiently high standard, and my cohorts were well powered to address the research questions in subsequent chapters. Finally, section 2.6 describes a novel method that improves genetic risk prediction for traits with shared genetic aetiology by leveraging sub-phenotype information to fine tune PRS.

2.2 Datasets

2.2.1 Overview of the phenotypes considered

Throughout this thesis I will be working with five phenotypes: height, body mass index (BMI), fluid intelligence, asthma and inflammatory bowel disease (IBD). The following two

sections provide a brief overview of each trait, emphasising aspects relevant to my work, and also explain my rationale for selecting them.

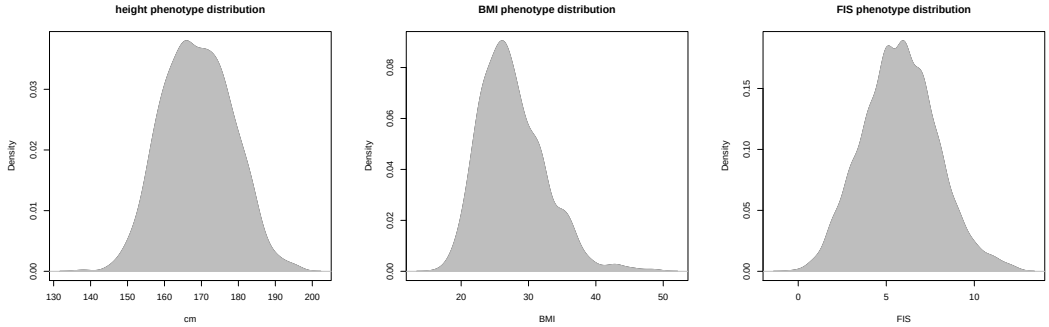
2.2.1.1 UK Biobank traits: height, BMI, fluid intelligence and asthma

Height and BMI (weight divided by height squared) are canonical quantitative traits with high heritabilities of ~80% and ~50%, respectively (Elks et al., 2012; Visscher et al., 2012). These two traits also offer some of the largest sample size available today (~700K (Yengo et al., 2018)); therefore, they represent an attractive go-to option to show the utility of novel methods as a proof of concept in a situation where sample size is less of a limiting factor. Current state of the art PRS models can now explain ~25% and 6% of phenotypic variance for height and BMI, respectively (Yengo et al., 2018).

Average population values for both height and BMI have been increasing in the developed world during the last century. There are many factors underpinning this increase, including increased access to nutrition, changes to culture and sexual selection favouring taller males (for height) (Stulp et al., 2015). In the UK Biobank (UKBB) cohort, the mean height is 168cm (SD: 9.3cm) and the mean BMI is 27 (SD: 4.8). The distribution of both traits is approximately normal, which I confirmed via a Kolmogorov–Smirnov test of a 1,000 randomly sampled individuals (Fig 2.1). BMI in the UKBB is moderately positively skewed (1.096), which is consistent with the well documented effect of BMI increasing across successive generations (Peeters et al., 2015). As the cohort's age range covered just over a generation, with a minimum and maximum age of 37 and 73, respectively, this effect may have contributed to the aforementioned skewness. During the next decade this increase in BMI is expected to result in up to 20% of the world population to become obese. This development may create a substantial public health burden due to obesity's connections to health risks, such as type 2 diabetes, cardiovascular diseases and certain cancers (Hruby and Hu, 2015).

Cognitive ability, or 'intelligence', may be defined as an abstract problem solving skill that does not rely on direct recall from memory (Plomin and von Stumm, 2018). This phenotype is also a highly polygenic trait, with adult heritability estimates ranging from 50-80% (Hill et al., 2018; Polderman et al., 2015). I selected this trait due to its perceived complexity, and the fact that it is not a disease trait, but rather an example of what may be considered 'positive genetics' (when genetic variation contributes to traits that may be considered beneficial (Plomin and Deary, 2015)).

The first principal component of test scores across many cognitive tests is known as the 'intelligence quotient' or IQ. Professional cognitive tests, such as Raven's progressive matrices (Raven, 1936), are administered under strict supervision over a time period of up



(a) Height (b) BMI (c) FIS

Fig. 2.1 **Distributions of the three quantitative phenotypes in the UKBB**. Height, body mass index (BMI) and fluid intelligence score (FIS).

to 40 minutes (Raven et al., 1988). Due to being part of a larger battery of measurements, the relevant field in the UKBB, fluid intelligence score (FIS), was generated from a much simpler test, the unweighted sum of 13 questions to be answered in two minutes. To find out if this difference between the FIS metric and more standard tests had any impact on my analyses, I performed several checks that are detailed under section 2.2.2. The discrete 14 possible outcomes of the FIS data made the Kolmogorov–Smirnov test inapplicable; however, visual inspection suggested that the distribution of this trait also follows a normal distribution. During the last century IQ scores have risen across the globe, ascribed to improved nutrition and access to education (Baker et al., 2015), a phenomenon known as the Flynn effect. On the other hand, recent reports indicate a slow decline of the genetic component of cognitive ability over the same period, as measured by PRS stratified by age (Kong et al., 2017).

The aforementioned three quantitative phenotypes are considered as classic polygenic traits that, aside from a few notable monogenic forms (Chiurazzi and Pirozzi, 2016; Durand and Rappold, 2013; Fawcett and Barroso, 2010), arise due the joint action of many variants with small effect, a property which makes them an ideal choice for methods that aim to model the phenotype from a large number of markers. An additional consideration in favour of these particular traits was that they cover a spectrum that ranges from the simple, additive physiological traits, such as height, to the more complex cognitive traits, such as FIS. On one extreme, recent studies indicate that all of height's heritability can be explained by additive genetic effects (Wainschtein et al., 2019). At the other extreme, twin studies suggest that non-additive genetic variation may contribute to the phenotypic variance of higher-level cognitive functions (Polderman et al., 2015).

The last UKBB trait, asthma, is also a complex polygenic trait that is characterized by respiratory inflammation and obstruction of the airways, which affects over 339 million

people world-wide (Vos et al., 2017). Recently, it was reported that asthma has a low to moderate genetic overlap with psychiatric disorders such as hyperactivity, anxiety and major depressive disorder (Zhu et al., 2019). Asthma is also a substantial source of public health loss and economic burden. In the next 20 years this condition is expected to cost over $960 billion in the USA alone (Yaghoubi et al., 2019) . Asthma's high population prevalence, ~20% in the developed world (Thomsen, 2015), together with a high estimated heritability of 55-90% (Hernandez-Pacheco et al., 2019), make it an ideal test subject for disease phenotypes. Another reason for the inclusion of the asthma phenotype was that it is also a representative immune related disorder, an attribute that allowed me to draw on my group's area of expertise and auxiliary data available, such as expression data from relevant tissues.

The UKBB includes 59,313 individuals (~12%) marked as positive for self-reported asthma, some of which were included in the UK BiLEVE study (Wain et al., 2015). The aims of the UK BiLEVE study were to examine the genetic bases of smoking behaviour and chronic obstructive pulmonary disease, a condition which has a moderate genetic correlation (0.38) with asthma (COPDGene Investigators et al., 2017). The strategy of this study included an over-sampling of individuals from the extremes of lung function distribution from the main UKBB cohort, and genotyping them on a different platform (the UK BiLEVE Axiom™ Array). The details of how I handled this differential sampling are described in section 2.4.0.1.

2.2.1.2 IBD and its subphenotypes

Inflammatory bowel diseases (IBD) are chronic inflammatory conditions of the gastrointestinal tract that encompass many subphenotypes. It is believed that these complex, relapsing disorders involve an inappropriate immune response to the enteric microbiota that interact with environmental risk factors in genetically susceptible individuals. Its two main clinical entities are Crohn's disease (CD) and ulcerative colitis (UC).

The genetic overlap between UC and CD may be described as substantial but imperfect. The majority of the ≥ 240 genome-wide significant associations are shared (de Lange et al., 2017; The International IBD Genetics Consortium (IIBDGC) et al., 2012), and their genome-wide genetic correlation was quantified at 0.56 (The UK-PSC Consortium et al., 2017). However, there is also considerable genetic heterogeneity, many shared variants exhibit a heterogeneity of odds, and some loci affect only one of the subphenotypes. Two notable examples for incongruent effects are *NOD2* and *PTPN22* which are risk variants for CD, but have a protective effect against UC (Furey et al., 2019).

Given its lower incidence and smaller sample sizes (~17,5K, for details see Table 2.3), I chose IBD to be included in this work to serve as a more realistic model for evaluating any novel methods.

2.2.2 UK Biobank genotype and phenotype data diagnostics

The UKBB project is currently the largest biobank resource in the United Kingdom that includes both genetic and phenotypic data on 487,409 individuals (Sudlow et al., 2015). In addition to the directly genotyped data of ~805,000 markers, it also contains ~97 million imputed variants (Bycroft et al., 2017). Participants between the ages of 40-69 were recruited during the years 2006-2010. The UKBB is a population based cohort which is expected to serve as a prospective epidemiological resource for diseases that may manifest in its target age range during the next decades. Some evidence suggests a "healthy volunteer" bias in the UKBB recruitment, as its participants were found to be slightly above average in health, education and socio-economic status, relative to the general UK population (Fry et al., 2017). However, as none of my analyses relied on comparisons with other cohorts, I did not expect the validity of my conclusions to be affected by this.

The field identifiers and estimated SNP heritabilities of the four UKBB phenotypes, standing height, BMI, FIS and self-reported asthma are summarised in Table 2.2. For brevity, I will be referring to standing height as height and self-reported asthma as asthma from this point onward.

For FIS, there were two relevant fields, 20016 and 20191. 20016 was recorded in person (at three different time points) and 20191 was recorded via an online follow-up. The tests were short (two minute long) touch screen based questionnaires that assessed the participant on cognitive reasoning tasks. To investigate how the fact that this phenotype was measured at several different time points under different circumstances may have impacted the recorded values, I calculated the correlations for the 1,217 individuals for whom I had a value for all four occasions which are presented in table 2.1.

	time1	time2	time3	online
time1	1	0.628	0.621	0.562
time2		1	0.653	0.601
time3			1	0.590
online				1

Table 2.1 **Correlations between the four occasions the FIS UKBB phenotype was recorded.** 'time1', 'time2' and 'time3' are the three different time points where the participants were assessed via in-person tests. 'online' represents the online follow-up test.

phenotype	type	field	SNP h^2	Neff
height	quantitative	50	0.485	360,388
BMI	quantitative	21001	0.248	359,983
FIS	quantitative	20016,20191	0.22	117,131
asthma	binary	20002_1111	0.171	148,259

Table 2.2 **UK Biobank summary of phenotypes.** 'SNP' h^2 is the LDSC estimated SNP heritability, 'Neff' is the effective sample size. Data was obtained from the Neale lab's 'SNP-Heritability Browser' online service from https://nealelab.github.io/UKBB_ldsc/index.html, accessed on 01/03/2020.

I observed a slightly lower correlation between the averaged in-person and online tests, ~0.63 and ~0.58, respectively. I performed a paired t-test and I found that the average scores were significantly lower (p-value $< 2.2 * 10^{-16}$) for the in-person recording versus the online follow-up, 6.155 and 6.405, respectively. A recent study by Fawns-Ritchie and Deary (2020) evaluated the validity of the UKBB cognitive tests, and found that, despite their non-standard format, these correlated well with more standard intelligence tests (r = 0.83); thus, I deemed that the FIS phenotypic data was of a sufficiently high standard to proceed.

2.2.3 IBD datasets

IBD is a well studied immune related disorder, and my own group has published a number large scale GWAS on IBD in recent years (de Lange et al., 2017; Luo et al., 2017). The datasets on which these studies were based on were made available for my analyses during my PhD. These datasets included the Wellcome Trust Case Control Consortium (WTCCC) 1 and 2, together with another dataset, internally identified as GWAS3. In subsequent chapters, I will be referring to these datasets as GWAS1, GWAS2 and GWAS3. These datasets were imputed via the internal Sanger imputation service (utilising the merged UK10K + 1000 Genomes Phase 3 reference panel) by a fellow team member, Loukas Moutsianas, and then filtered to exclude variants with a MAF < 0.001 and an INFO < 0.4. Further details of sample collection, imputation and initial quality control protocols are described in the original publications of each study (Barrett et al., 2009; de Lange et al., 2017; WTCCC et al., 2007). Table 2.3 summarises the specifications of these studies, including sample size counts and the genotyping platforms.

study	original platform	phenotype	cases	controls	SNPs
GWAS1	Affymetrix GeneChip	CD	1,196	2,919	7,582,624
GWAS2	Affymetrix 6.0	UC	1,918	2,776	8,476,301
GWAS3	Human Core Exome v12.1/0	IBD	8,062	9,492	8,017,981
		CD	3,810	9,492	8,020,419
		UC	3,765	9,492	8,020,586

Table 2.3 **Platform and study size details for the three IBD datasets.** 'GWAS1', 'GWAS2' and 'GWAS3' refer to the WTCCC1, WTCCC2 and the internal GWAS dataset, respectively.

2.3 Quality Control

2.3.1 Common quality control steps

To facilitate meaningful comparisons between the more experimental NN approaches and the classical statistical methods I will use in Chapter 3 and Chapter 4, I employed a common quality-control strategy implemented for each trait separately. I employed this strategy to ensure that all methods were evaluated on the same version of the datasets, starting from the same conditions.

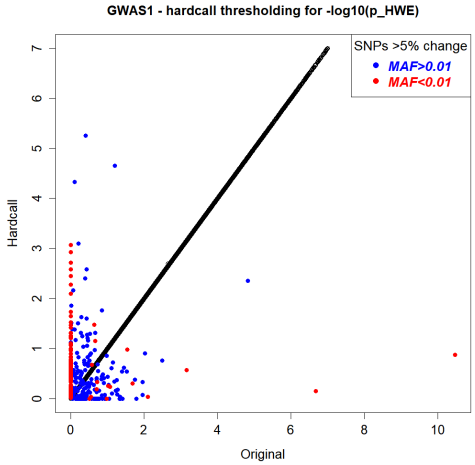
2.3.1.1 Converting genotype probabilities to hard calls

The raw data files that I started my analyses from were the imputed genotypes for the UKBB and IBD datasets in BGEN 1.2 and VCF formats, respectively. Both of these formats store genotypes as probabilities represented by real values. However, as many of the tools used in this thesis, such as LDpred and my own NN framework, only support PLINK1 genotype files (.bed/.bim), which are hard calls (0, 1 or 2 alternative alleles), I had to convert the data to this format. Using PLINK2 with a hard threshold rate of 0.1, I converted allele dosages that were greater than 0.1 away from a nearest hard call to be recorded as missing, and the rest thresholded to the nearest integer. This meant that unless the dosage for the alternative allele fell between $0.0 < dosage < 0.1$, $0.9 < dosage < 1.1$ or $1.9 < dosage < 2.0$, it was recorded as missing.

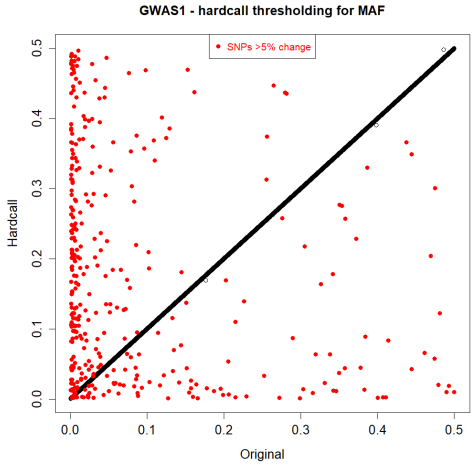
Converting genotype probabilities to hard calls is a lossy process that may result in substantial changes in allele frequencies for variants where allele dosages are uncertain. One possible option would have been to randomise the hard calling process to preserve the same allele frequencies that were recorded in the original files. I decided against this, as this would have permuted the inter-variant relationships. This would have been a problem, as the arrangement of alleles with respect to each other is a crucial element for detecting non-linear

genetic effects, as tests for statistical epistasis compare the effects of different haplotypes on phenotypic variance. Therefore, hard-calling variants was still the best option, despite the potential problems arising from changes to allele frequencies.

To identify variants where hard-calling variants may have created a problem, I performed Hardy-Weinberg tests and computed MAFs in the datasets before and after their conversion. Upon a visual inspection of the plots (Figs 2.2a and 2.2b), I deemed that removing variants that differed by more than 5% in either the $-log10$ of the Hardy-Weinberg test p-value or MAF between the original and converted datasets would eliminate the change in allele frequencies caused by hard-calling issue. This filtering removed 3,826,495 and ~13,430 variants in the UKBB and IBD datasets, respectively.



(a) **The effect of hard calling on HWE p-values.** Variants retained after filtering are displayed in black and SNPs removed are coloured by their MAF.



(b) **The effect of hard calling on MAF.** Variants retained after filtering are displayed in black and SNPs with a greater than 5% difference after conversion are highlighted in red.

2.3.1.2 Post-imputation quality-control for the UKBB genotypes

I excluded individuals who were sex-discordant, which I determined by comparing the '*Submitted Gender*' and the '*Inferred Gender*' fields in the UK Biobank *Sample-QC* file. I also removed individuals who were not defined as '*white British*' or had third degree relatives in the cohort, as described in the UK Biobank documentation. The aforementioned filtering left 376,007 individuals for further analyses.

To ensure only high quality markers remained for my analyses, and to reduce the multiple testing burden, I excluded all variants that had a MAF < 0.1% or an imputation INFO score < 0.8. I relied on the INFO score metric that came with the UKBB data release; however,

I recomputed MAFs from the subset of individuals that actually remained in my analyses. Finally, I only kept SNPs with unique positions that passed filters for a missing genotype filter of < 2% and a Hardy-Weinberg test of $P_{HWE} < 10^{-7}$. These steps left a total of 12,211,706 SNPs for further analysis.

The HLA region is an extremely polymorphic area of high LD that has many confirmed associations for immune related diseases (International Inflammatory Bowel Disease Genetics Consortium et al., 2015). However, because the HLA region is unlike other areas of the genome, any potential insights from this locus could be considered unrepresentative with respect to the rest of the genome. Therefore, considering both the additional computational burden that it would have took to maintain the HLA region in my analyses, and that I was interested in drawing general conclusions on method performance over the genome, I decided to exclude this area. I removed markers in the HLA region by excluding SNPs from the range 6:28477797-33448354, in B37 coordinates.

2.3.1.3 Post-imputation quality-control for the IBD genotypes

The IBD studies were all previously quality-controlled and imputed using the Sanger imputation service by other members of my lab. To facilitate my own analyses, I performed the following additional QC steps for each dataset. I only kept SNPs with unique positions, with an imputation INFO > 0.8, a MAF > 0.1% and a missing genotype rate < 2%. Next, I excluded all SNPs that significantly deviated from the Hardy–Weinberg equilibrium with a $P_{HWE} < 10^{-5}$ in controls or $P_{HWE} < 10^{-7}$ in all individuals. Finally, I removed markers in the HLA region by the exclusion of SNPs from the range 6:28477797-33448354, in B37 coordinates. These steps left between 7,582,624 - 8,476,301 markers for further analysis across the different studies. The full details of each dataset and each subphenotype are presented in Table 2.3.

To control for cryptic population structure or any residual batch effects within my datasets, I performed PCA within each dataset (which were previously filtered to only include individuals of European ancestry). To perform the PCA, I used the subset of SNPs available in the IBD datasets (~83,585) which were identified in the UKBB documentation as suitable for this purpose based on QC passed status, MAF and lack of LD. I carried out PCA to estimate the top 20 principal components with the software FlashPCA 2.0 (Abraham et al., 2017).

2.3.1.4 Phenotype quality control

Complex trait phenotypes are affected by factors other than genetic variation and these could potentially confound the analysis if they are causally associated with both the outcome of interest as well as the genotype (Anderson et al., 2010). In a traditional GWAS of a quantitative trait, covariates are usually added into a linear regression model where their individual effects may be isolated via

$$Y = G\beta_G + Z\beta_Z + e, \tag{2.1}$$

where Y, G, Z and e denote the phenotype column vector, the SNP, the covariate and a random noise term, respectively. β_G and β_Z are the coefficients for the SNP and the covariate, respectively. In this model, β_Z, and its p-value, would allow one to evaluate the importance of the Z covariate while the variable of interest, G, is held constant.

However, the non-linear nature of neural-network models does not allow investigators to obtain similarly reliable estimates of the effect of individual predictors the same way as it is possible for linear models (covered in detail in Chapter 4 4.2.5). As my intention was to use the same version of the data for all methods, I decided to control for the covariates' effect by regressing them out of the phenotype ahead of the main analyses. This process also transformed binary phenotypes into continuous ones, which also made all analyses into linear regression-like problems. All subsequent work in this chapter, as well as all analyses in later chapters was performed on these phenotype residuals.

I will now describe the protocol to obtain these phenotype residuals. First, I fit a regression with all considered covariates in the model. This was logistic regression for the binary traits and linear regression for the quantitative traits. Then, I performed backward selection by removing the term with the highest p-value one-by-one, until there were no terms left with a p-value threshold of > 0.05 (Bonferroni corrected based on the number of covariates). Finally, I fit the reduced model with only the surviving terms, and the phenotype residuals from this model were then taken forward as the outcome against which all subsequent analyses were performed.

To identify potential covariates, I cross-referenced the covariates that my lab had access to against covariates that similar UKBB studies have used for the same phenotypes (Johansson et al., 2019; Savage et al., 2018; Yengo et al., 2018). The full list of covariates I considered were *age*, age^2, *sex*, *PC*1-20, $Townsend\_deprivation\_index$, $centre$ and $batch$. For the IBD analyses these were *sex* and *PC*1-*PC*20. Table 2.4 summarises the results from this step.

I note that the *sex* covariate for the IBD datasets was not always identified as significant by my variable selection process. The incidence of both UC and CD are known to vary by

phenotype	**significant covariates**
BMI	sex, age, age^2, $Townsend\_deprivation\_index$, $centre$, $batch$, $PC4-5$, $PC7$, $PC9-11$, $PC14$, $PC16$, $PC20$
Height	sex, age, age^2, $Townsend\_deprivation\_index$, $centre$, $batch$, $PC1$, $PC4-5$, $PC7-9$, $PC11-16$
FIS	sex, age, age^2, $Townsend\_deprivation\_index$, $centre$, $PC4-5$, $PC7$, $PC11-12$, $PC14$, $PC16$, $PC18-20$
Asthma	sex, age, age^2, $Townsend\_deprivation\_index$, $centre$, $PC5$, $PC9$
GWAS1 CD	sex, $PC1$, $PC3$
GWAS2 UC	$PC1$, $PC3$
GWAS3 IBD	sex, $PC1$, $PC2$, $PC4$, $PC5$
GWAS3 CD	$PC1$, $PC2$, $PC4$
GWAS3 UC	sex, $PC2$, $PC4$

Table 2.4 **List of significant covariates for both the UKBB and IBD datasets.** Covariates were selected by a two stage backward selection process to be considered for each dataset and phenotype combination.

sex depending on the patients' age group. However, this effect may only be consistently shown in large scale meta-analyses (Shah et al., 2018); therefore, the relatively small sample size of my studies may explain why it was not always identified as significant in my own datasets.

2.3.1.5 Further filtering of genotypes for the TWAS and protein burden score tests

As both the TWAS and protein burden analyses use the same genotype data that I processed through the previously described QC steps, the genotype data itself did not require additional QC.

For the protein burden tests, to simplify my analyses, I intersected the post-QC genotype panels of the four UKBB phenotypes to yield a single set of SNPs, which resulted in a loss of less than 10,000 markers. Additionally, I intersected the resulting panel with the list of FIRM scores that had a numeric entry, which left a total of 61,081 exonic SNPs that had protein affecting scores.

For the TWAS, as my analyses relied on LDpred to build the per-gene level predictors (described in detail in Chapter 3 in section 3.2.2.1), I applied the following filtering steps. I subset my QC-passed GWAS data to the HapMap3 SNP panel before proceeding (a recommendation for practical performance gains by the authors of the LDpred tool: https://github.com/bvilhjal/ldpred/wiki/Q-and-A, accessed on 01/11/2019). Then, I intersected this

subset of markers with the SNPs for which I had expression data available in the BLUEPRINT summary datasets which left 692,298 markers.

2.4 Experimental setup for later analyses

2.4.0.1 Cohort organisation in the UK Biobank

Due to the non-linearity, neural-network based methods are especially prone to overfitting (a phenomenon when a model learns the noise patterns in a data to achieve a better fit on the training set but fails to generalise to new data). Therefore, to prepare my datasets for my work in Chapter 4, I divided my datasets in the following manner. I divided the full cohort into two partitions, one for training and validation ('*Main Set*'), and another for testing ('*Test Set*'). For all but the asthma phenotype, I split the datasets based on the two chips used, the UK Biobank Axiom™ and UK BiLEVE Axiom™ arrays which contained ~90% and ~10% of the individuals, respectively. I decided to use the individuals on the UK BiLEVE chip as the Test Set to eliminate a potential batch effect arising from the different platforms.

For the asthma experiments I chose not to include individuals on the UK BiLEVE Axiom™ Array to avoid any potential bias that could arise from the fact that this chip was specifically designed to facilitate the UK BiLEVE study. The aim of this study was to examine lung function, and the chip included a special subset of markers that had shown previous association to asthma. Therefore, I decided to only include the individuals on the Biobank Axiom™ array and generated all data partitions from within that. Finally, I generated 20 bootstrap samples to be able obtain variance estimates for PRS that I built subsequently. This process entailed sampling with replacement from the Main Set the same number of individuals to be included in a bootstrap sample as the total number of individuals. The resulting set of individuals served as a training set for the bootstrap sample. Sampling with replacement results in some individuals being sampled more than once, while others may not be included at all. I kept track of this latter category of unique individuals that were not sampled into the training bootstrap sample, which I then used as the corresponding validation set. This process yielded three non-overlapping subsets of my original data that I subsequently used for training, validation and testing. Table 2.5 summarises the size and partitions of all datasets used in subsequent chapters that relied on the UKBB data.

2.4.0.2 Dataset organisation for the IBD datasets

Individuals in the GWAS3 study were separated into three subsets based on their phenotypes (CD, UC and IBD). Within each of these datasets, bootstrap training and validation samples

phenotype	Main set	bootstrap training	bootstrap validation	Test set
BMI	332,059	332,059	~122,000	43,948
Height	332,059	332,059	~122,000	43,948
FIS	137,088	137,088	~34,000	21,775
Asthma	298,853	298,853	~107,000	33,206

Table 2.5 **The number of individuals in the various data splits for each experiment for the UKBB phenotypes.** The validation set sizes are shown as approximate, as the number of unique individuals not sampled into the training set varied slightly in each bootstrap sample due to the random nature of the resampling process.

were generated in a manner identical to the one I described above in section 2.4.0.1. The GWAS1 and GWAS2 studies were selected to be used as the Test Sets for CD and UC, respectively.

2.5 Additive association tests

This section describes the technical details of the additive association tests that I performed on all phenotypes and datasets. The results from this initial association step form the basis for my later interaction analyses in Chapter 3 and Chapter 4.

2.5.1 GWAS

I performed a standard GWAS on the 'Main Set' of individuals (Table 2.5) on each dataset and cohort by applying PLINK's '*–assoc*' function, which fits an OLS linear regression model that regresses the phenotype on each individual SNP.

2.5.1.1 Post-association QC

GWAS signal can be recognized by a particular LD signature that provided the inspiration for the naming of the Manhattan plots. The basic principle is that, provided there is adequate coverage, associated SNPs are supported by other nearby markers with signal ($-log10(p)$) linearly proportionate to their LD with the index variant (Farh et al., 2015). Many false positive associations may be visually identified as being either isolated or in a group with no coherent LD structure structure underpinning them (for an illustration, see Fig 2.3) . Such false positives may be generated at the various steps of the sample and marker processing stages (Anderson et al., 2010), or even by the imputation algorithm (Lin et al., 2010).

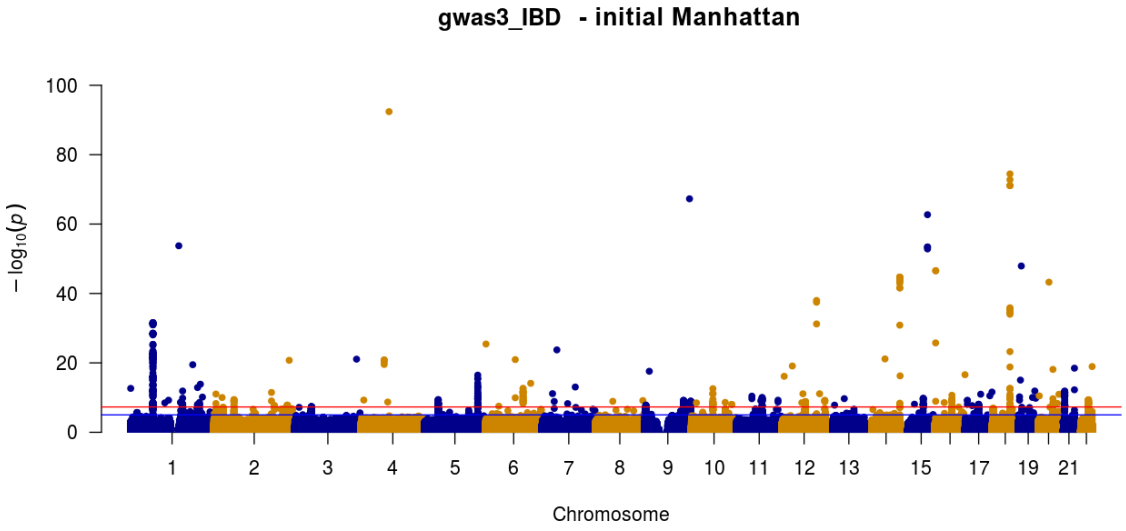


Fig. 2.3 **Manhattan plot visualising the GWAS1 study without applying post-association QC to consider LD patterns.** There are many associations above the genome-wide significance level with no LD structure to support them, a property that marks them out as potential false positives.

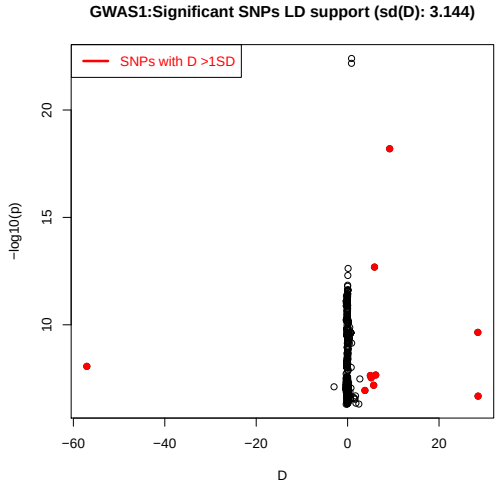
Traditionally, the quantitative allelic signals (intensity plots) of SNP associations suspected of being false positives are individually inspected for unexpected clustering patterns. In case of an imputed variant, several directly genotyped markers may be examined in the region. To make filtering for potential false positives practicable for the number of analyses in my project, I decided to take the expected relationship between LD and genuine signal, and derive rules that may be automatically applied. Working with all datasets, I based this test on an OLS regression model that relates association signal to LD. I extracted the LD-friends (defined as SNPs having an $r^2 > 0.2$ with the target variant) for all the SNPs with an association p-value $< 5 * 10^{-8}$. Then, I fit an OLS linear regression model on these SNPs

$$-log10(p) = \beta_{r^2} r^2 + e, \tag{2.2}$$

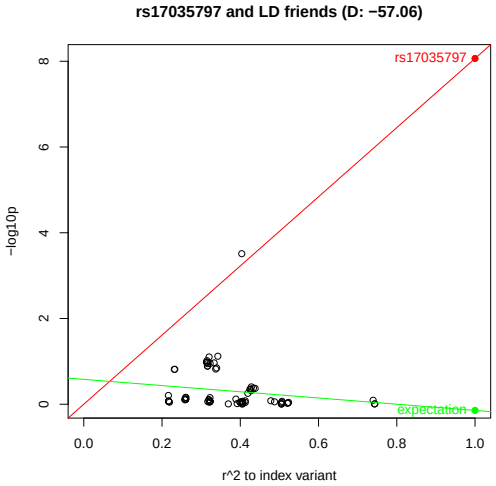
where r^2, β_{r^2} and e denote the LD to the target variant, its coefficient and the noise term, respectively. Next, using this model, I predicted the $-log10(p)$ of the target association using an r^2 of one (the target association's correlation squared with itself). Finally, I defined the value D, to quantify the difference between observed and expected $-log10(p)$ as

$$D = \frac{-log10(p) - (-log10(p)_{expected})}{-log10(p)_{expected}}. \tag{2.3}$$

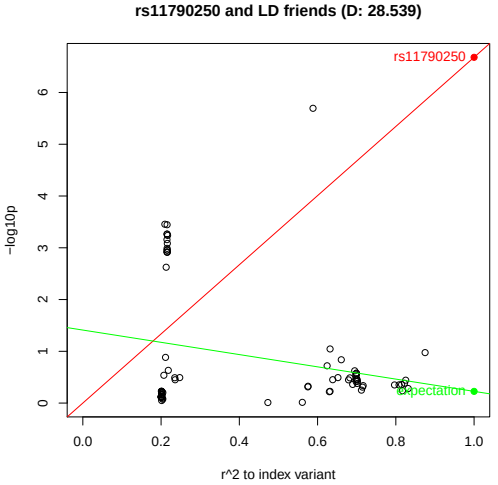
I note that D may be either negative or positive, depending on if the target SNP has higher or lower significance level than what would be expected by considering nearby variants. Upon examining the distribution of D and how it related to significance (Fig 2.4a), I set the exclusion criteria for markers as $abs(D) > 1 * SD(D)$, or if a SNP had less than four LD-friends. I reasoned that SNPs that fail this latter criterion may come from an area that was insufficiently covered, poorly imputed or that the variant is very rare. This step eliminated 748 SNPs across the IBD datasets. For the UKBB, this process removed 1,791, 1,679, 1,583 and 572 SNPs for FIS, height, BMI and Asthma, respectively. To see illustrative examples of how this algorithm was used to eliminate potential false positives, refer to Fig 2.4b and 2.4c.



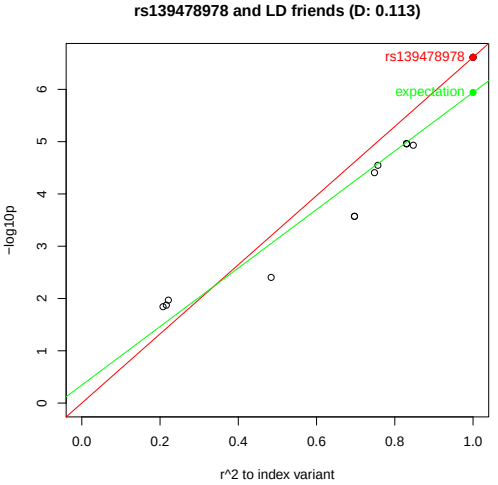
(a) **Demonstration of how the GWAS signal of SNPs in a study depends on D.** Outliers with an $SD(D) > 1$, highlighted in red, are the variants that were filtered out.



(b) **Illustrative example of how a potential false positive is identified by the algorithm.** The target variant is highlighted in red. The green line is the regression's line of best fit from the tagging variants. The green dot represents the prediction for the target variant's predicted significance.



(c) **Example with a D value of the positive extreme where the LD structure does not support the association.** Here, the algorithm filtered the SNP out as a potential false positive.

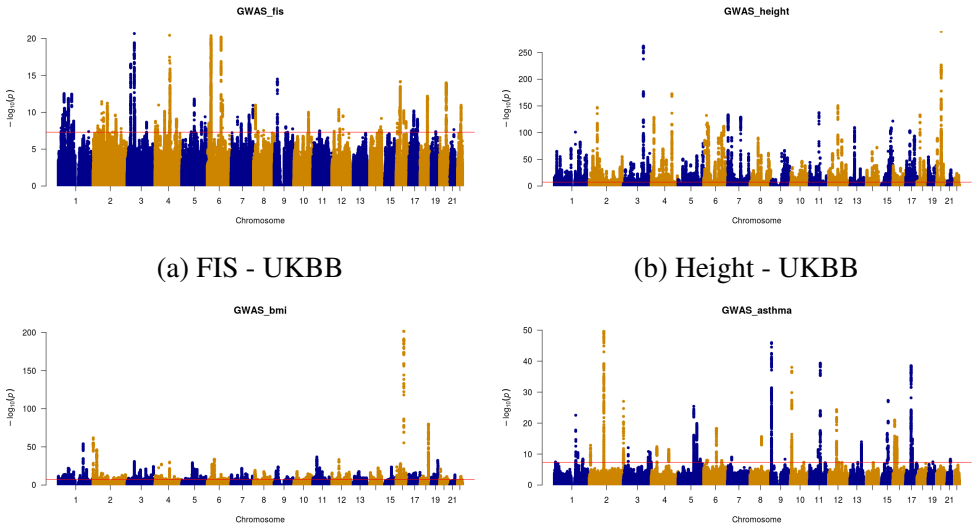


(d) **Example where the LD structure supports the variant as a genuine association.** The D value here is small, as the variant's predicted significance level is very close to the actual $-log10(p)$.

Fig. 2.4 **Four examples that illustrate common cases where the application of the automated filtering either eliminated potential false positive associations, or alternatively, retained those consistent with the nearby signal.**

2.5.1.2 UKBB association test results

I performed a GWAS on all variants via PLINK1.9's '–assoc' functionality for each of the UKBB phenotypes (height, BMI, FIS and asthma). Then, I subjected these initial results to the post-association QC steps described in section 2.5.1.1. The final results after this step are presented in Fig 2.5.



(a) FIS - UKBB

(b) Height - UKBB

(c) BMI - UKBB

(d) Asthma - UKBB

Fig. 2.5 **Manhattan plots visualising the UKBB GWAS.** y-axis shows the $-log10$ of the additive association p-values and the x-axis displays the genomic coordinates. The red line represents the genome-wide significance level of $5 * 10^{-8}$.

2.5.1.3 IBD association test results

I performed a GWAS via PLINK1.9's '–assoc' functionality on each individual study and on both subphenotypes within GWAS3. Then, I processed these initial results through the post-association QC steps described in section 2.5.1.1. The final results from this step are presented in Figs 2.6 and 2.7.

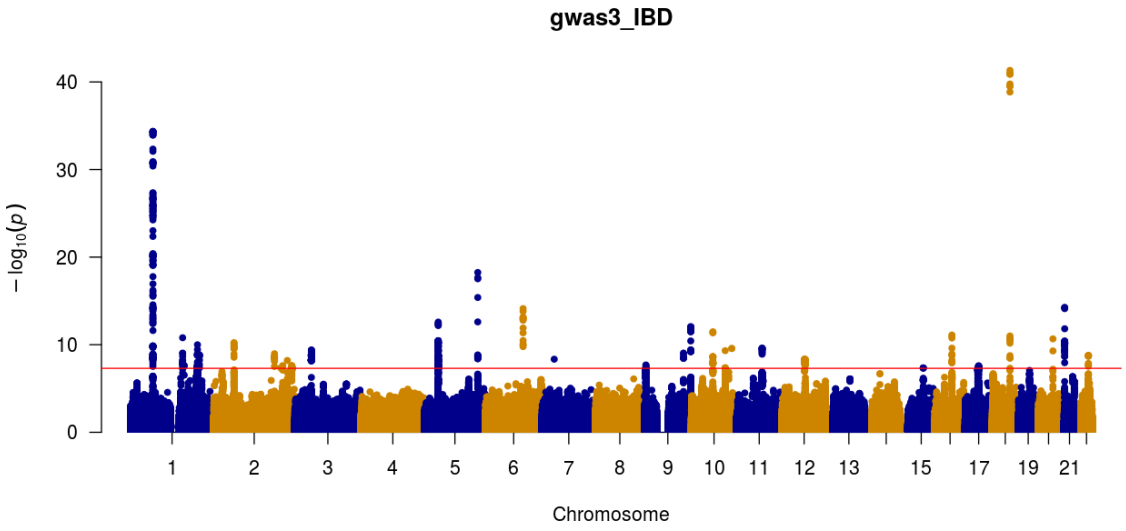
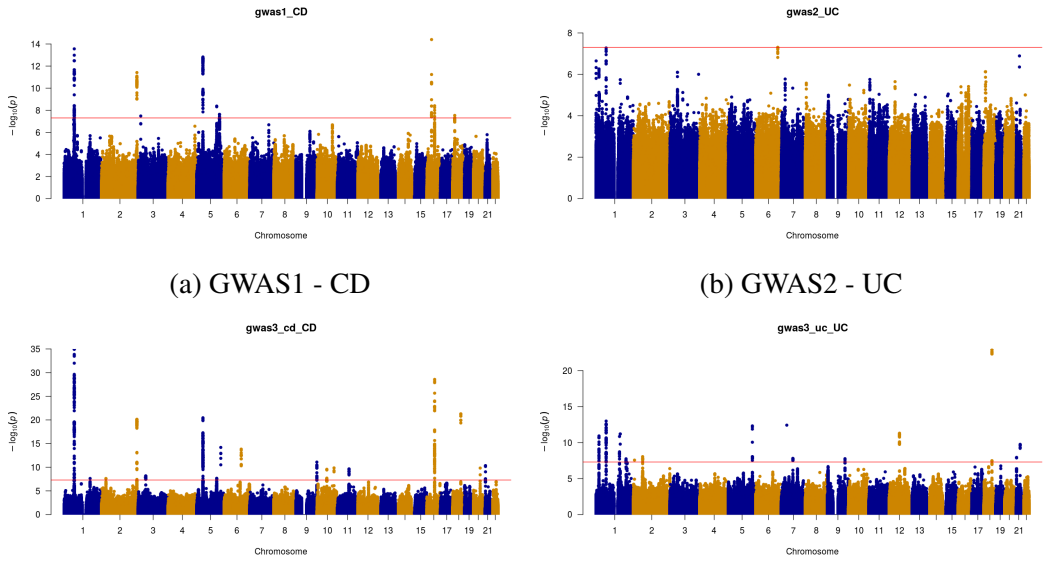


Fig. 2.6 **Manhattan plot visualising the GWAS3 dataset IBD association result.** y-axis represents the $-log10$ of the additive association p-values and the x-axis displays the genomic coordinates. The red line represents the genome-wide significance level of $5 * 10^{-8}$.



(a) GWAS1 - CD

(b) GWAS2 - UC

(c) GWAS3 - CD

(d) GWAS3 UC

Fig. 2.7 **Manhattan plots visualising the IBD, CD and UC GWAS.** y-axis represents the $-log10$ of the additive association p-values and the x-axis displays the genomic coordinates. The red line represents the genome-wide significance level of $5 * 10^{-8}$.

2.5.2 Summary of the additive association experiments

My main objective with the additive association tests described so far was to ensure that my data meets quality standards adequate for my subsequent analyses in later chapters. Most of my cohorts and analyses were not novel in a sense that the same datasets, or a subset of them, were already used for previously published analyses. Therefore, I only make a few general observations, and highlight specific landmarks in my results and how they relate to findings in the relevant literature. I do this only to convince my readers of the validity of my experimental procedure so far, not to claim any novel insights, which I hope to derive from later analyses in Chapter 3 and Chapter 4.

To evaluate the validity of the results of my UKBB analyses, I searched for comparable studies in the literature. For height and BMI I chose Yengo et al. (2018), for asthma I selected Johansson et al. (2019), and for FIS I used the study by Savage et al. (2018). I note that even though our datasets were not identical, since those studies were meta-analyses that involved other cohorts besides the UKBB, visual inspection of our Manhattan plots suggested a strong qualitative similarity between my results and the published records. To quantify the similarity in our results, I compared z-scores for two of my UKBB traits (height and BMI) that had comparable publicly available summary statistics. The results from these analyses are presented in Table 2.7 and Fig 2.8.

My IBD datasets were different versions of the same studies that were used for a meta-analysis by de Lange et al. (2017); therefore, that study presented itself as a natural basis for comparison. Once again, I observed qualitative similarities between our corresponding Manhattan plots. I also cross-checked a few key landmark associations for each trait from my analyses against those found in the supplementary table S3 of the de Lange et al. (2017) study. The results from this are shown in Table 2.6.

trait	top SNP	gene	p-value	de Lange p	chrom	position
IBD	rs11581607	*IL23R*	$1.114 * 10^{-34}$	$4.59 * 10^{-111}$	1	67707690
CD / IBD	rs2076756	*NOD2*	$2.716 * 10^{-29}$	$1.42 * 10^{-38}$	16	50756881
UC	rs10263242	N/A	$4.400 * 10^{-7}$	$9.07 * 10^{-21}$	7	107489762

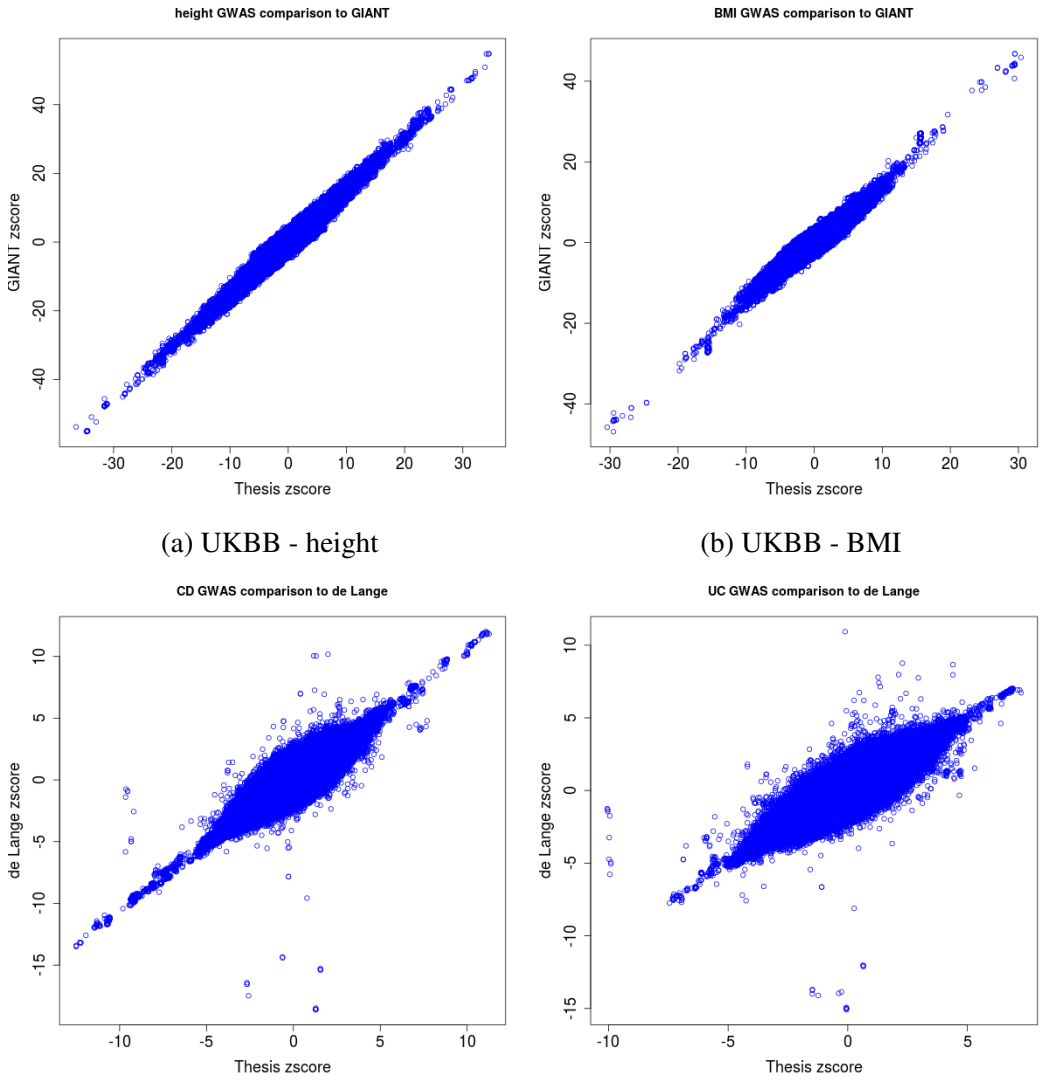
Table 2.6 **Landmark associations for my IBD analyses.** Comparisons of associations between the GWAS3 dataset and the study by de Lange et al. (2017). 'de Lange p' is the p-value from the de Lange et al. study, and 'chrom' indicates the chromosome.

For IBD, I identified a variant (rs11581607) in the locus of *IL23R* with a p-value of $1.114 * 10^{-34}$. For CD, I recovered *NOD2* via a variant (rs2076756) with a p-value of $2.716 * 10^{-29}$. Finally, possibly owing to the lower heritability of UC, I was unable to locate

a suitable proxy for the most strongly associated locus (tagged by rs6017342 in de Lange et al.) within its LD bracket that achieved genome-wide significance in my analysis; however, I managed to identify a variant in the second most significant locus (tagged by rs10263242) with a p-value of $4.4 * 10^{-7}$. To obtain a broader sense of congruency between our results, similarly to the UKBB analyses, I selected two traits (UC and CD) from the GWAS3 dataset and compared their association z-scores to the summary statistics by de Lange et al. (2017). The results from this comparison are presented in Table 2.7 and Fig 2.8.

phenotype	**correlation**	**correlation** ($p < 5 * 10^{-8}$)
CD	0.926	0.994
UC	0.932	0.975
height	0.919	0.994
BMI	0.890	0.989

Table 2.7 **The four traits I selected for a quantitative comparison against reference studies from the literature.** The values in the correlation column are Pearson correlation coefficients between the z-scores from my association results and those of the literature. The values in the column 'correlation' ($p < 5 * 10^{-8}$), are Pearson correlation coefficients computed between z-scores that were restricted to have an additive association $p < 5 * 10^{-8}$.



(a) UKBB - height

(b) UKBB - BMI

(c) GWAS3 - CD

(d) GWAS3 UC

Fig. 2.8 **Plots comparing the GWAS z-scores of my results against relevant studies in the literature.** x-axis ('Thesis zscore') represents the z-scores from my analyses, and the y-axis represents z-scores for the same variants I obtained from reference studies in the literature.

.

I found that the overall correlation between my results and the reference studies was strong, ranging from ~0.89 (BMI) to ~0.93 (UC). I also restricted the calculation to those variants with an association $p < 5 * 10^{-8}$; here, I observed even stronger correlations that ranged from ~0.98 (UC) to ~0.99 (height). This latter increase of correlations may be explained by the reduction of random discrepancies of the less significant associations due to different sample sizes and variations in data processing steps. In summary, my results were

highly congruent with the literature; thus, I felt confident that my analyses so far would form a sound basis for my later work.

2.6 Leveraging shared genetic effects to improve genetic risk prediction for IBD

As I described in section 2.2.1.2, IBD is a collective term for conditions with overlapping genetic aetiologies (de Lange et al., 2017). Its two main clinical entities, CD and UC, share a substantially but imperfectly overlapping genetic aetiology with a genetic correlation of 0.56 (The UK-PSC Consortium et al., 2017). A recent review of UC and CD (Furey et al., 2019) summarised that, while the majority of confirmed SNPs have effects of the same direction and similar magnitude, there were also incongruent associations that differentiated the two subphenotypes. I was interested in if such an imperfectly shared aetiology may be used to improve the performance of PRS by developing an approach that could exploit heterogeneity of effects between the two subphenotypes.

2.6.1 Establishing baselines

To evaluate the potential benefits of more advanced approaches, I first needed to establish a baseline prediction performance for the two subphenotypes. I trained two sets of PRS, one on cases that only consisted of the target subphenotype (UC or CD alone), and another one from all IBD cases. This baseline PRS would also answer the question of the bias-variance trade-off inherent in predicting a phenotype from the smaller but more precise study, or from the larger but mixed study. On one hand, SNP effect estimates from the smaller subphenotype dataset would be expected to have lower bias but a higher variance. On the other hand, SNP effect estimates from the combined dataset would be expected to yield a higher bias but lower variance estimates.

I used the LDpred tool (described in the Introduction in section 1.6.3.3) to construct the baseline IBD PRS. I began my analysis by subsetting my post-association QC datasets to the HapMap3 panel. Then, I extracted an LD reference panel of 5,000 individuals from the GWAS3 dataset to be used in LDpred. I then performed a GWAS for all bootstrap samples to produce association summary statistics. Next, I generated the full default range of PRS, one for each causal fraction hyper parameter ($p : \{1, 0.3, 0.1, 0.03, 0.01, 0.003, 0.001, 3 * 10^{-4}, 10^{-4}\}$) for the first bootstrap sample. Then, the best performing p was selected, based on the performance of the generated PRS against the first bootstrap sample's validation set. The same p (0.3) was selected by this process for all three phenotypes. Next, I ran LDpred

to adjust the summary statistics for the rest of the 20 bootstrap samples (using the same p). Finally, I built 20 PRS for the two Test Sets, GWAS1 and GWAS2 for CD and UC, respectively. The performance of these PRS are presented in Fig 2.11.

The performance of the PRS were evaluated by r^2 (squared correlation) between predicted and observed phenotypes, which were 0.026 vs 0.027 and 0.012 vs 0.014 between the subphenotye and mixed datasets for CD and UC, respectively. I also performed paired t-tests on each pair, and I found that they were not significantly different. From the point of view of the variance-bias trade-off my results made intuitive sense. I approximately doubled the number of cases (Table 2.3) for phenotypes that share approximately half of their genetic aetiology ($r = 0.56$); thus, the values of the trade between sample size (variance) and a more precise phenotype (bias) approximately cancelled each other out. In conclusion, I interpret my findings to support the established results of a substantial but imperfect genetic overlap between CD and UC (Furey et al., 2019).

My initial results established a baseline reference for PRS performance for the prediction of both disease subphenotypes. The next question I was interested in was if it was possible to improve on the baselines by finding the best balance between the SNP estimates from each PRS. That is, to choose the larger sample size and lower variance where the SNP effects were congruent between subphenotypes, but to favour the more precise phenotype and lower bias where SNP effects were found to be heterogeneous.

2.6.2 Estimating SNP heterogeneity of effect in the IBD studies

To find the best balance for SNP effects between UC and CD I used Cochran's Q-test to estimate a per-SNP heterogeneity of effect via

$$Q = \frac{(\beta_{CD} - \beta_{UC})^2}{SE_{CD}^2 + SE_{UC}^2}, \tag{2.4}$$

where β_{CD}/β_{UC} are the SNP coefficient estimates for CD and UC, respectively, and SE_{CD}/SE_{UC} are the standard errors of the estimates for CD and UC, respectively. The Q test statistic is distributed according to χ^2 with one degree of freedom.

However, as the UC and CD studies used the same individuals as controls, not accounting for this effect would have resulted in the inflation of Type I errors. Therefore, I estimated the Q test statistic via a procedure described by (Lin and Sullivan, 2009) as

$$Q_{adjusted} = \frac{(\beta_{CD} - \beta_{UC})^2}{SE_{CD}^2 + SE_{UC}^2 - 2\rho SE_{CD} SE_{UC}}. \tag{2.5}$$

This is very similar to the original formula (eq 2.4), the only difference is an extra term in the denominator that adjusts for the overlap between the studies. Here, ρ is the quantity that measures the extent of the overlap. To determine ρ I evaluated the following two possibilities. An approximation formula for ρ was described by Lin and Sullivan (2009)

$$\rho_{approx} = (n_{cu0}\sqrt{\frac{n_{c1}n_{u1}}{n_{c0}n_{u0}}})/\sqrt{n_u n_c}, \tag{2.6}$$

where n_c and n_u are the total number of individuals in the CD and UC studies, respectively, n_{cu0} is the number of overlapping controls, n_{c1} and n_{u1} are the number of cases in CD and UC, respectively, and n_{c0} and n_{u0} are the number of controls in CD and UC, respectively. I also considered an alternative strategy to estimate ρ via the calculation of an empirical correlation of SNP estimates between the two studies. I selected a subset of SNPs in the GWAS3 IBD dataset that had an IBD association $p > 0.01$, and I computed ρ from these summary statistics as

$$\rho = cor(\beta_{CD}/SE_{CD}, \beta_{UC}/SE_{UC}). \tag{2.7}$$

I found that the ρ and ρ_{approx} values were similar, 0.269 and 0.286, respectively, so I chose to proceed with ρ. To get a sense of how the Q-values are distributed across the genome, I produced a Manhattan plot from these values (Fig 2.9). To reassure myself of the validity of my progress so far, I examined the largest peak on this plot on chromosome 16, and I identified it to be within the *NOD*2 locus, which is a confirmed site of high heterogeneity between CD and UC (The International IBD Genetics Consortium (IIBDGC) et al., 2012).

2.6.3 Finding the balance between the subphenotypes and IBD

To improve on the baseline PRS I described in section 2.6.1, I considered two methods to balance SNP effect estimates. Both of these methods were based on the same idea, to favour the SNP estimate for the phenotype with the greater evidence of being appropriate, but differed in the way this was implemented. One approach I evaluated was to build a composite PRS based on a hard threshold, and the other approach was to continuously weight each SNP via a blending factor. The end goal in both approaches was to create a new set of summary statistics by modifying each SNP's coefficient before generating a new PRS via LDpred.

I decided to use local FDR (lFDR) as a metric for strength of association for both approaches. In contrast with Bonferroni correction or Benjamini & Hochberg's FDR, lFDR performs not only multiple testing correction, but it also provides a per-predictor statement

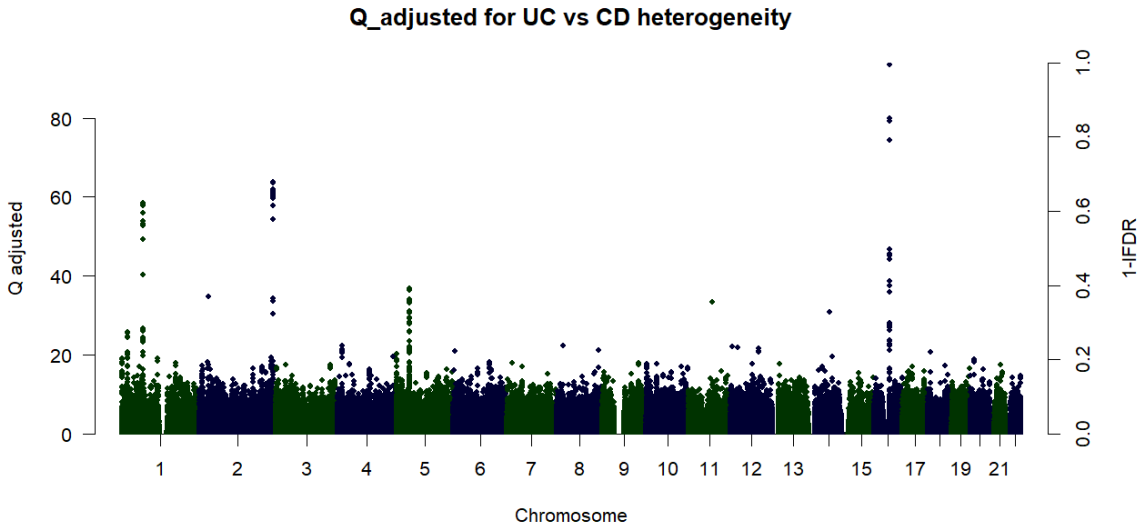


Fig. 2.9 **Manhattan visualising the adjusted Q values that measured SNP heterogeneity of effect between CD and UC.** Left y-axis shows the adjusted Q-values and right y-axis shows 1-lFDR. x-axis represents genomic coordinates.

about the probability that a particular SNP is consistent with the null hypothesis:

$$lFDR_i = Pr(H_i = 0 | P_i = p_i), \tag{2.8}$$

where H_i is the null hypothesis for predictor i, p_i is the SNP's association p-value and P_i is the evaluated probability.

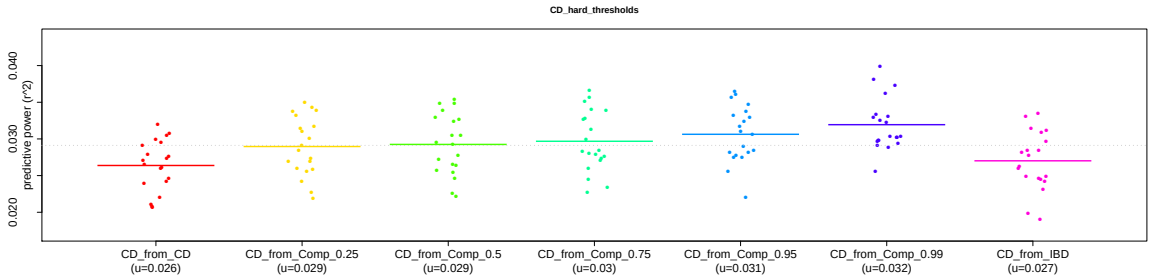
I implemented the composite PRS method by swapping the SNP summary statistics between the subphenotype and IBD as

$$SS^i_{threshold} = (1-I)SS^i_{subpheno} + ISS^i_{IBD}, \tag{2.9}$$

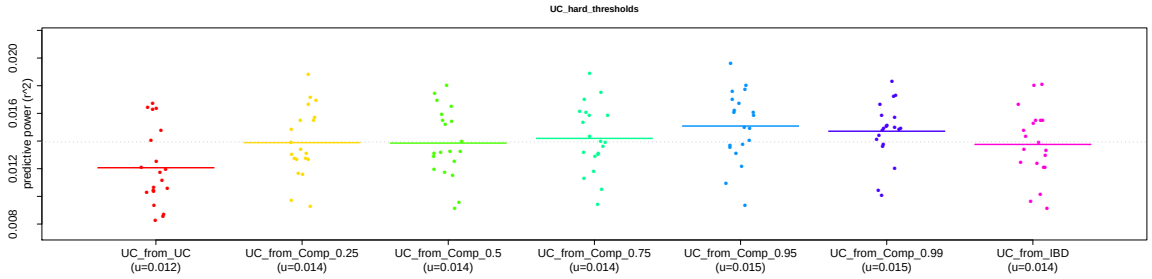
where $SS^i_{threshold}$ is a summary statistic associated with SNP_i for the current threshold that included β, SE, p and N (the number of individuals used to perform the association). I is an indicator function defined as

$$I = \begin{cases} 1, & \text{if lFDR >t} \\ 0, & \text{otherwise,} \end{cases} \tag{2.10}$$

which chose the IBD summary statistic if the lFDR indicated no heterogeneity of effect, and the subphenotype if it did indicate heterogeneity of effect. This selection was evaluated based on a range of five thresholds $t = \{0.25, 0.5, 0.75, 0.95, 0.99\}$. The results from these analyses are presented in Fig 2.10.



(a) CD composite PRS - hard thresholds



(b) UC composite PRS - hard thresholds

Fig. 2.10 **Dot-plots for the IBD subphenotype compose PRS hard threshold experiments.** y-axis represents the r^2 between the predicted and observed phenotypes. The dots represent bootstrap samples and the coloured bar is the mean across all bootstrap samples. The grey dotted line represents the mean across all experiments. The suffix after each plot's name indicates the lFDR threshold used to swap between subphenotype and IBD SNP summary statistics.

To build the continuously weighted PRS, I blended the summary statistics appropriate for linear interpolation (β and N) between the subphenotype and IBD via

$$SS^i_{blend} = (1 - lFDR)SS^i_{subpheno} + (lFDR)SS^i_{IBD}. \tag{2.11}$$

As the analogous relationship is not linear for standard errors, I interpolated those via

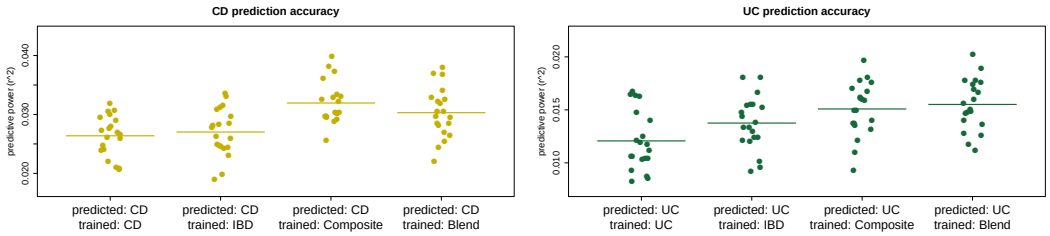
$$\begin{aligned} O &= (1 - lFDR)^2 SE^2_{subpheno} + SE^2_{IBD} lFDR^2 \\ SE_{blend} &= \sqrt{O + 2lFDR(1 - lFDR)SE_{subpheno}SE_{IBD} * cor(\beta_{subpheno}, \beta_{IBD})}. \end{aligned} \tag{2.12}$$

The p-value for the blended SNP effect was then derived from the new blended SNP coefficient and its standard error. This process yielded a new set of summary statistics, which I then used to generate new PRS scores via LDpred by almost the same procedure that I previously described in section 2.6.1. The only difference in the construction of these PRS

was that I did not need to re-estimate p (the causal fraction), as these were identical across all three phenotypes; thus, I was able to reuse the same hyperparameter.

2.6.4 Results for predicting IBD subphenotypes

The final results of the most performant PRS are presented in Fig 2.11. I observe that both the blended and best hard threshold composite PRS outperformed their baseline PRS counterparts.



(a) CD final PRS plots

(b) UC final PRS plots

Fig. 2.11 **Dot-plots for the IBD subphenotype composite and blended PRS experiments.** y-axis represents the r^2 between the predicted and observed phenotypes. The dots represent bootstrap samples and the coloured bar is the mean across all bootstrap samples. The naming convention is as follows. The first line of each PRS represents the target phenotype on which the PRS was evaluated on and the second line represents the source on which the PRS was trained on. For example, "*predicted: CD trained: Blend*" is the PRS that was evaluated on the CD phenotype and was trained using the blended PRS approach.

2.6.5 Discussion of the improved IBD subphenotype PRS

I took advantage of the substantial but imperfect overlap in the genetic aetiologies of CD and UC to develop an approach that improves the performance of PRS by exploiting the genetic correlation and heterogeneity between the two subphenotypes. The performance of the subphenotype-from-IBD PRS was better, although not significantly, than the single-trait PRS with an $r^2 = 0.012$ vs $r^2 = 0.014$ and $r^2 = 0.026$ vs $r^2 = 0.027$ for UC and CD, respectively. The PRS generated from my novel approaches further improved on the single-trait baselines with an $r^2 = 0.015$ (p-value: $6.824 * 10^{-4}$) and an $r^2 = 0.031$ (p-value: $1.109 * 10^{-4}$), which represent an overall improvement of ~25% and ~19% for UC and CD, respectively.

IBD is a good model trait for disorders where larger GWAS datasets to estimate SNP effect sizes that yield more accurate PRS are unavailable due to the relatively low population prevalence of the disease (for IBD this is ~0.3% (Ng et al., 2017)). Therefore, my work may be used to derive a general principle to improve PRS performance in situations analogous to

my IBD subphenotype datasets. That is, where a larger pooled study may be available which consists of genetically overlapping subphenotypes that present clinically distinct entities. Disease domains where this may apply include psychiatric, metabolic and immune related disorders. In summary, my approach may be of particular relevance to uncommon disorders, where individual studies for (sub) phenotypes may be too small to build a serviceable PRS on their own.

The performance of the blended and composite PRS was not significantly different for UC (p-value: 0.145); however, the latter significantly outperformed the blended approach for CD (p-value: 0.003). The blended approach offers several advantages over the composite method however, as it is faster to compute (as it does not require the evaluation of a range of thresholds), and more importantly, it does not require genotype level data.

The method described so far is suited for situations where there is a substantial gap between heritability and the accuracy of the PRS due to low power. In a scenario where sample sizes are very large, the estimates of the SNP coefficients may already be accurate; thus, this approach may offer limited benefits. This method also relies on the existence of a substantial, but imperfect genetic correlation. Therefore, for traits where rG is either zero or one, this method may also not be appropriate, as in those circumstances the subphenotype or the combined phenotype SNP estimates would be expected to perform better, respectively. Additionally, rG on its own may not completely describe the shared genetic aetiology between two diseases, and I expect that the variance of the distribution of genetic heterogeneity may also play an important role. For example, an rG of 0.5 between two diseases may be possible without any loci of high heterogeneity (such as the ones shown in Fig 2.9, like *NOD2*) in which case I would expect my approach not to offer an advantage over a combined phenotype PRS. To quantify the ranges of genetic aetiologies under which my method would be expected to offer an advantage, a range of rGs and distributions of heterogeneous sites would need to be explored via simulation studies.

In the domain of immune mediated disorders, recent related work (Burren et al., 2020) showed that a wide range of clinically-related diseases have substantial overlap in their genetic architectures, which may be potentially exploited to better characterise their aetiologies in modest sample size cohorts. By adopting a similar approach, I expect that the method I described here could be generalised for the multi-trait scenario, where the accuracy of PRS may be further enhanced by borrowing information between more than two diseases.

Chapter 3

Regression based models of statistical epistasis

3.1 Chapter 3 outline

This chapter covers my search for two-way interactions via classical statistical methods that belong to the regression framework. Section 3.2 details my approach for dimensionality reduction that produced the transcriptome and protein score views of my UKBB cohorts. Section 3.2 describes my search for epistasis in the GWAS data and in the derived gene-level domains in the UKBB. Cross-domain experiments where the different genomic views (SNPs, TWAS and protein scores) were integrated to search for interactions across the different domains are described in section 3.4.

In the analyses described in section 3.5, I pursued a hypothesis-driven approach to search for statistical epistasis in the IBD datasets, where the search-space was reduced to only consider the evidence for haplotype-specific interactions between specific coding and regulatory variants.

3.2 Dimensionality reduction in the UKBB

As I described in the Introduction in section 1.1.7, managing the dimensionality of the search-space, by the reduction of the total number of tests to increase power, is of key importance to increase the chance to successfully detect statistical epistasis. Therefore, I employed the following dimensionality reduction strategy. I generated derived gene-level predictor datasets that summarise information on the gene-level based on genetically predicted expression levels and protein burden scores. Additionally, I applied the established best practices of

filtering predictors on both additive effects and LD (Cordell, 2009; Marchini et al., 2005; Van Steen, 2012a; Wood et al., 2014).

In the subsequent sections where I describe the various processing steps, I will be referring to the '*Main Set*' and '*Test Set*' edits of the UKBB cohort. These were created in the previous chapter and a detailed explanation of their parameters can be found under Chapter 2 section 2.4, where Table 2.5 provides the specifics on the exact number of individuals in each set.

3.2.1 Transcriptome and protein score data-sets

Transcriptome-wide association studies (TWAS) and the protein burden score tests allow one to search for signal on the gene-level rather than on the SNP-level, and these frameworks offer several important advantages. Aggregating many SNPs into a single predictor reduces the dimensionality, which in turn reduces the multiple testing burden. Additionally, such gene-scores may capture signal in scenarios where multiple SNPs with small but genuine congruent effects do not meet the genome-wide significance threshold individually; however, when aggregated into a single predictor they may collectively reach significance.

The next two sections describe how I generated the TWAS and protein score datasets that I will use to perform my analyses subsequently in this chapter and in Chapter 4 as well.

3.2.1.1 FIRM protein scores

FIRM is a machine-learning model that considers the proteomic context of missense SNPs. This model evaluates each variant based on its location within the protein sequence, the nature of the amino acid substitution and finally, annotations from the UniProt, Pfam and ClinVar databases. Thus, FIRM scores quantify each SNP's predicted effect at the biochemical functional level, rather than on the clinical outcome at the organism level. This makes FIRM unique compared to other variant effect prediction tools which assess mutation pathogenicity (Brandes et al., 2019b).

The predicted effect score of each SNP is a value between zero and one, which represents complete loss of function and no harmful effect on the protein, respectively. The authors of this method have kindly agreed to share their database of generated scores for 97,013,422 UKBB markers.

3.2.1.2 BLUEPRINT transcriptome data

One of the aims of the BLUEPRINT epigenome project is to provide high-resolution transcriptomic profiling of cis-genetic factors in three major human immune cell types, CD14+ monocytes, CD16+ neutrophils and naive CD4+ T-cells (Chen et al., 2016). For brevity, I

will refer to these cell types as monocytes, neutrophils and T-cells from here on. This project includes a reference panel that has expression data on 194, 192 and 171 individuals and summary statistics for 84,982,294, 76,901,636 and 87,575,990 marker-expression quantitative trait locus mapping association tests for monocytes, neutrophils and T-cells, respectively.

3.2.2 TWAS for asthma in the UKBB

As I described in the Introduction in section 1.4, the TWAS framework may be used to derive biological insight on a gene-level basis; however, for my purposes I was primarily interested in using it as a dimensionality reduction tool. The TWAS framework consists of two main stages, the generation of PRS that capture the genetic component of the expression of each gene, and an association step that relates the phenotype to these PRS.

3.2.2.1 Imputing the transcriptome

To date most successful TWAS were aimed to identify individual gene-phenotype associations. These studies relied on filtering on MAF and/or on eQTL p-value, followed by the application of either LASSO or elastic net to identify markers suitable to predict the transcriptome (GTEx Consortium et al., 2015; Gusev et al., 2016; Zhu et al., 2016). However, I believe that continuous weighting is preferable to discarding information when possible. Therefore, I opted for using the LDpred method instead, as it has been shown to outperform PRS generating methods that rely on hard thresholds (Khera et al., 2018; Lee et al., 2018). The reason behind LDpred's success is that, in contrast to hard thresholding and filtering approaches that eliminate SNPs completely (such as those relying on L1 norms), it applies a continuous weighting scheme that leverages all of the data from all variants. This considers both the confidence in SNP association signal as well as local LD structure (Vilhjálmsson et al., 2015). Therefore, I chose LDpred to impute gene expression based on the three reference panels I described in section 3.2.1.2.

I generated per-gene expression PRS that relied on the summary statistics extracted from the BLUEPRINT data for each gene for my cohort. There were 16,516, 14,621 and 16,945 genes available for monocytes, neutrophils and T-cells, respectively. I then combined the eQTL summary data with the individual GWAS genotypes to aggregate SNPs into expression-level predictors for each individual. The step-by-step procedure to generate these scores was as follows. First, I exported out the SNPs in my cohort that had a matching eQTL summary result in the BLUEPRINT data into a separate PLINK file. Next, I generated the LD-adjusted eQTL SNP coefficients using the LDpred '*gibbs*' function. It is important to emphasise that at this stage LDpred did not consider the GWAS phenotype. All SNP

coefficients refer to the SNPs' relationship to gene expression in a tissue, rather than to disease status. Thus, the LDpred LD-shrinkage was based purely on the eQTL summary data and an LD reference panel generated from the GWAS genotypes. The GWAS phenotype itself was only considered at the last stage, where I used it to select the highest performing causal fraction parameter (p) for each gene, based on the gene expression PRS performance at predicting the GWAS trait. This is in contrast with standard TWAS approaches, such as PrediXcan (GTEx Consortium et al., 2015), where the target phenotype is not considered when building the expression-scores. However, I wanted to determine the causal fraction of SNPs based on the performance on an independent subset of the cohort of the target trait. I reasoned that this would emphasise eQTLs most relevant to the GWAS phenotype, as that was the final association target, not the gene expression (as in the PrediXcan study). Finally, I built a per-gene PRS using the LDpred '*score*' function for all genes and all individuals in each of the three tissues via

$$\widehat{E_i} = G_{gene}\beta_{eLDpred}, \tag{3.1}$$

where $\widehat{E_i}$ denotes the imputed expression for gene i in a particular tissue, G_{gene} denotes the SNPs in the gene and $\beta_{eLDpred}$ denotes the adjusted eQTL coefficients for these SNPs which were determined in the previous step. I repeated this procedure for all bootstrap samples, for the Main Set and for Test Set datasets as well.

3.2.2.2 Expression association to the phenotype

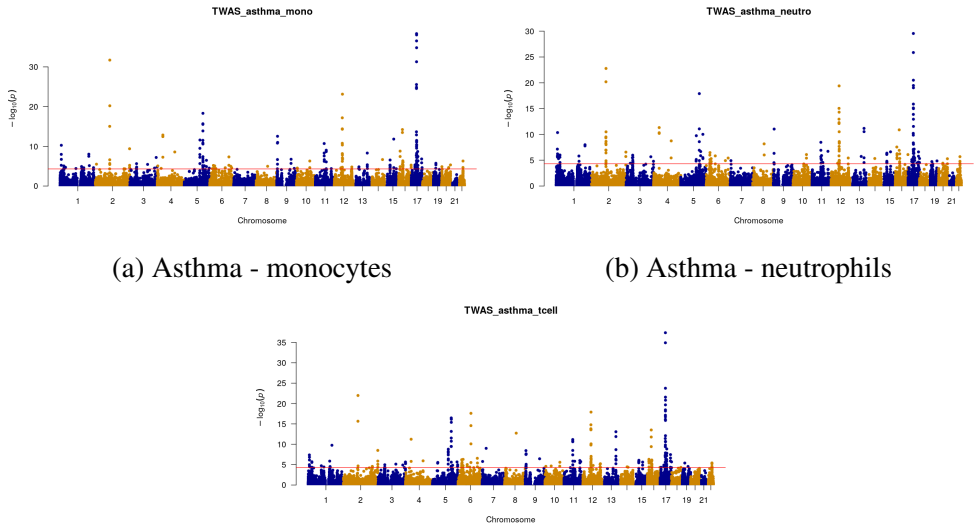
To perform the standard TWAS additive association test on the Main Set, I fit a simple univariate OLS linear model of the phenotype against each gene's predicted expression level as

$$Y = \widehat{E_i}\beta_i^{GeneExpr} + e, \tag{3.2}$$

where $\widehat{E_i}$ denotes the expression for gene i in a particular tissue, $\beta_i^{GeneExpr}$ is its associated coefficient and e is a noise term.

3.2.2.3 UKBB asthma TWAS dimensionality reduction results

The results for the three tissues investigated for the UKBB asthma phenotype are presented in Fig 3.1. I observe that these results appear to closely mirror their GWAS counterpart from Chapter 2 (Fig 2.5), and upon visual inspection it may be said that they resemble lower resolution versions of the latter. The three asthma TWAS among themselves also look very similar to each other, which is not surprising, since the only difference between them is the differential weighting of the gene-level predictors derived from the three tissues.



(a) Asthma - monocytes

(b) Asthma - neutrophils

(c) Asthma - T-cells

Fig. 3.1 **Manhattan plots visualising all three tissues in the UKBB asthma TWAS.** y-axis represents the $-log10$ of the additive association p-values. x-axis shows the genomic coordinates. Red line represents the (Bonferroni corrected) genome-wide significance level of $5 * 10^{-6}$.

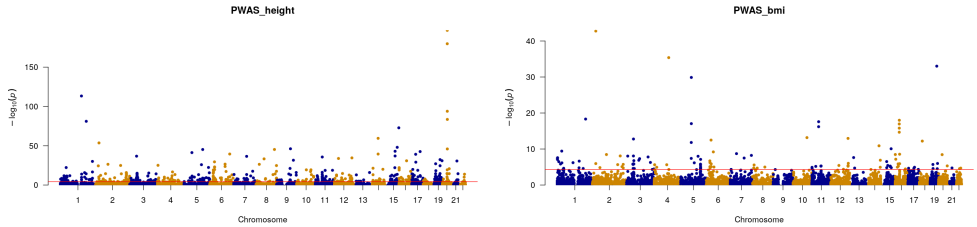
3.2.3 Protein burden score tests in the UKBB

Similarly to TWAS, protein burden tests may also be deployed to identify individual genes with relevance to the phenotype. However, just like with TWAS, I was mainly interested in this framework's dimensionality reduction capability. My workflow for conducting the protein burden score analyses followed closely the one described by the authors of this method (Brandes et al., 2019a), the details of which I described in the Introduction in section 1.5.1. I performed gene-score generation step on the Main and Test Sets, as well as all bootstrap samples using the 'PWAS' tool's "*calc_gene_effect_scores*" function. I also filtered out all genes which had less than two constituent SNPs, as in that case applying a FIRM score as a weight to a single predictor would not have provided an advantage over the original GWAS. This process generated a total of 7,283 gene-scores that I then used for the association step.

3.2.3.1 Protein burden score dimensionality reduction results

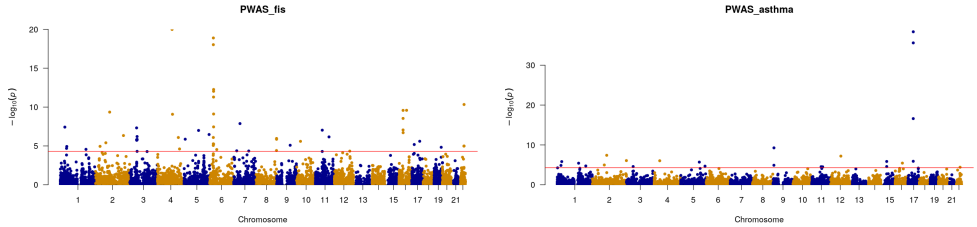
The protein burden test results for the four UKBB phenotypes are presented in Fig 3.2. The UKBB protein burden score test results also appear to be broadly congruent with their GWAS Manhattan counterparts, which reflect the fact that they were both derived from the same

underlying genotype datasets. Visual inspection suggested that the protein score results were slightly more noisy than their GWAS counterparts. However, this apparent noise may be explained by the fact that these were gene-level associations, generated from a much sparser panel of only 61,081 underlying SNPs across only 7,283 genes; thus, the same level of LD support would not be expected to be present as for their GWAS counterparts.



(a) Height - PWAS

(b) BMI - PWAS



(c) FIS - PWAS

(d) Asthma - PWAS

Fig. 3.2 **Manhattan plots visualising the PWAS test results for the four UKBB traits.** y-axis represents the $-log10$ of the additive association p-values. x-axis shows the genomic coordinates. Red line represents the $-log10$ p-value threshold of $5 * 10^{-6}$.

3.2.4 Filtering the protein burden and gene expression scores

For the gene level predictors that I produced in sections 3.2.2.1 and 3.2.3, I employed a similar filtering strategy that I used on SNPs (described in section 3.2.5). I performed FDR correction on the full unfiltered list of scores. As the gene-level predictors are real numbers in a format that is not compatible with PLINK, I was unable to use standard LD clumping. Instead, I implemented my own LD filtering strategy that also considered evidence of association. Briefly, this consisted of eliminating all except one of the predictors that were within 2000kb windows and had a pairwise $r^2 > 0.1$, preferentially keeping gene-scores with lower additive association p-values. Finally, I intersected these index gene-scores with those that had an FDR < 0.05 to select the top most likely independent associations among these. The summary of this filtering process is shown in Tables 3.1 and 3.2.

3.2.5 GWAS data

To reduce the potential for haplotype effects to induce statistical epistasis, and also to keep the dimensionality of my datasets low enough to be suitable for my later neural-network analyses, I applied following filtering steps to reduce the number of SNPs to the low thousands. I performed FDR correction on the full unfiltered list of SNPs. Then, I used PLINK's LD clumping feature on the genotype data and GWAS summary statistics to filter out SNPs within 2000kb windows that had an $r^2 > 0.1$. Finally, I intersected these LD-clumped index SNPs with those that had an FDR < 0.05 to select the top most likely independent associations among these. This process resulted in 1,277, 7,547, 3,247 and 656 SNPs for FIS, height, BMI and asthma, respectively.

3.3 Interaction tests

Using the Main Set of my UKBB cohort, I fit the following regression model with an interaction term to test for statistical epistasis

$$Y = \beta_1 P_1 + \beta_2 P_2 + \beta_{1,2} P_1 * P_2 + e, \tag{3.3}$$

where Y denotes a phenotype column vector and e is a random noise term. The P are the predictors, which may refer to either SNPs or gene-level predictors, such as protein burden scores or TWAS expression scores, and the βs are their corresponding coefficients. The total number of tests I performed for each experiment are summarised in Tables 3.1 and 3.2.

phenotype	**SNP**		**Protein scores**	
	pre/post filtering	number of tests	pre/post filtering	number of tests
FIS	94,918 / 1,277	814,727	129 / 97	4,656
Height	689,573 / 7,547	28,474,832	1,234 / 991	490,545
BMI	345,034 / 3,247	5,269,882	416 / 334	55,611
Asthma	19,361 / 656	214,841	44 / 37	666

Table 3.1 **Summary of the number of predictors and interaction tests performed in the UKBB cohort.** The columns 'pre/post filtering' display the number of SNPs or PWAS scores pre and post LD filtering out of the total number of < 0.05 FDR corrected predictors. The 'number of tests' columns show the total number of interaction tests performed post-filtering using either the SNPs or the protein burden scores.

Tissue	TWAS	
	pre/post filtering	number of tests
monocytes	715 / 358	57,292
neutrophils	628 / 297	39,061
T-cells	743 / 344	52,651

Table 3.2 **Summary of the number of TWAS scores and interaction tests performed for the asthma phenotype.** The column 'pre/post filtering' displays the number of TWAS scores pre and post LD filtering out of the total number of $FDR < 0.05$ corrected predictors. The 'number of tests' column shows the total number of interaction tests performed post-filtering.

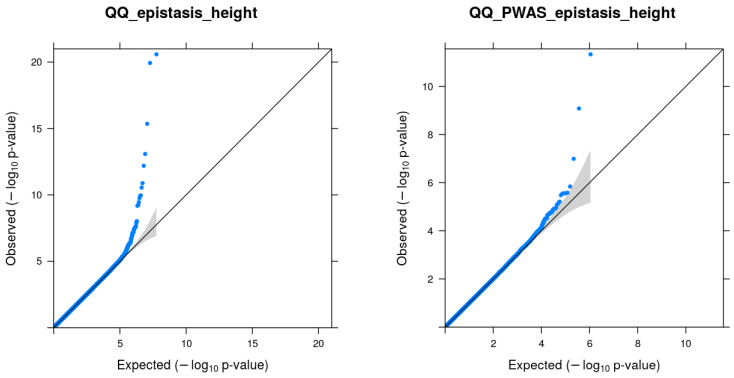
3.3.1 Post-association QC

As I described in the Introduction in section 1.1.6.1, attempts at detecting statistical epistasis require additional QC considerations unique to interaction test analyses. These considerations include haplotype effect induced statistical epistasis (Wood et al., 2014), and thresholding artefacts that affect traits where the measurements do not cover the true underlying range of the phenotype (Fish et al., 2016; Wei et al., 2014b).

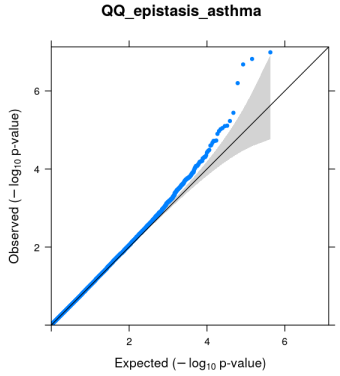
I examined the QQ-plots of my initial interaction tests (Fig 3.3) and I observed that only the height SNP/protein burden score tests and the asthma SNP tests appeared to deviate from the null. I decided to examine these two phenotypes in more detail to assess the potential for the aforementioned two factors to have induced false positives into the results.

The height analysis relied on a much denser set of markers (7,547 SNPs and 991 protein scores) than any of the other phenotypes; thus, it may have been particularly vulnerable to haplotype effects. Therefore, I further restricted my tests to reduce the potential for false positives by eliminating one of any two predictors that were either within the same recombination block (within the boundaries of one cM) or closer than 500kb. I determined the 500kb limit empirically, as after the application of the one cM filter there were still a few interactions in close proximity with p-values outside of the 95% CI. Closer inspection revealed that these variant/gene pairs were near the cM borders. I measured the furthest distance between them to be ~260Kb in the height GWAS SNP analysis. As the boundaries of the recombination blocks that I used were approximate (Burren et al., 2014), also considering the poor track record of replication of epistatic associations (Wood et al., 2014), I chose to be conservative and excluded one of each pair of variants that were less than 500Kb apart. I also applied the same filtering strategy to all of the remaining UKBB datasets.

The described LD filtering strategy reduced the number of SNPs to 955, 1,732, 1,671, 451, for FIS, height, BMI and asthma, respectively. For the protein score analyses this left 99, 781, 317 and 38 predictors for FIS, height, BMI and asthma, respectively. Finally, for



(a) Height - SNP (b) Height - Protein burden scores



(c) Asthma - SNP

Fig. 3.3 **QQ-plots visualising the p-values of the two-way interaction term for the height SNP and protein burden score domain and asthma SNP domain.** Grey area represents 95% confidence intervals.

the asthma phenotype, the same filtering process left 215, 187 and 204 TWAS gene-level predictors for monocytes, neutrophils and T-cells, respectively.

3.3.2 Interaction test results

Tables 3.3 and 3.4 summarise the final post-QC results for the two-way interaction test analyses. The QQ-plots for all experiments are presented in Figs 3.4, 3.5 and 3.6.

Visual inspection indicated that the interaction p-values do not show a trend that systematically deviates from the null in any of the QQ-plots, which is consistent with the notion that the deviations I observed for height and asthma before the post-association QC were caused by the aforementioned haplotype effects. Considering individual pairs of interactions, aside from asthma, none of the analyses generated an interaction test result that had an FDR < 0.05.

There was a single pair of SNPs (rs117290331 and rs115122203) for the asthma phenotype that had an FDR < 0.05 (FDR=0.015). The details of this association are provided in Table 3.5. Given that this association involved relatively rare variants, a MAF of 0.016 and 0.007 for rs117290331 and rs115122203, respectively, there was also a potential concern that this association may have been a false positive induced by an imputation error.

phenotype	**SNP**		**Protein scores**	
	minimum FDR	number of tests	minimum FDR	number of tests
FIS	0.411	455,535	0.989	4,852
Height	0.099	749,501	0.632	304,591
BMI	0.896	697,501	0.748	50,087
Asthma	0.015	101,475	0.178	703

Table 3.3 **Summary of post-QC results for the two-way interaction tests for all four UKBB phenotypes for both SNP and protein scores.** The 'minimum FDR' column represents the lowest FDR observed in a given experiment, and the 'number of tests' column displays the total number of tests performed.

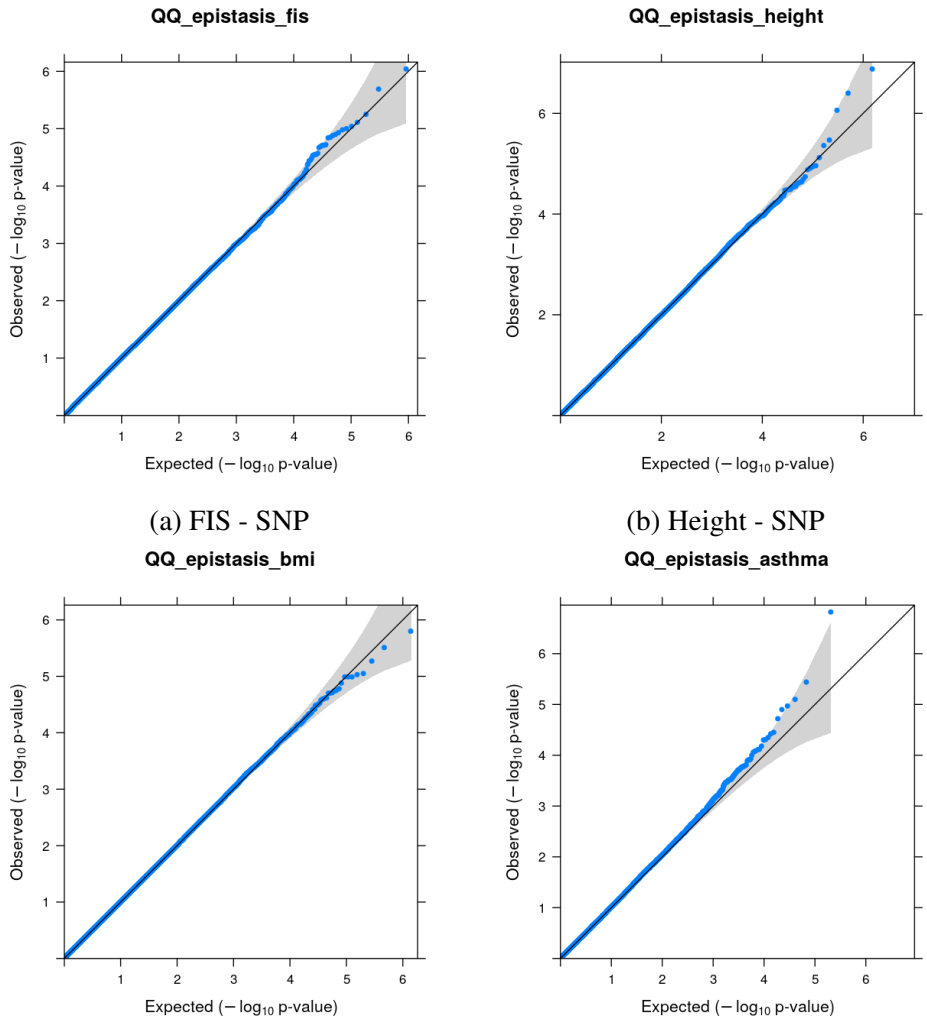
Tissue	**TWAS**	
	minimum FDR	number of tests
monocytes	0.734	34,716
neutrophils	0.422	23,653
T-cells	0.764	31,878

Table 3.4 **Summary of post-QC results for the three TWAS tissues for the asthma phenotype** The 'minimum FDR' column represents the lowest FDR observed in a given experiment and the 'number of tests' column displays the total number of tests performed.

.

term	**p-value**	**beta**	**MAF**
rs117290331	$5.92 * 10^{-4}$	0.011	0.016
rs115122203	$3.58 * 10^{-3}$	0.014	0.007
interaction	$1.53 * 10^{-7}$	0.136	*N*/A

Table 3.5 **Summary of the model terms of the linear regression between SNPs rs117290331 and rs115122203 for the asthma phenotype.** Values in the 'beta' column represent the regression coefficient.



(a) FIS - SNP

(b) Height - SNP

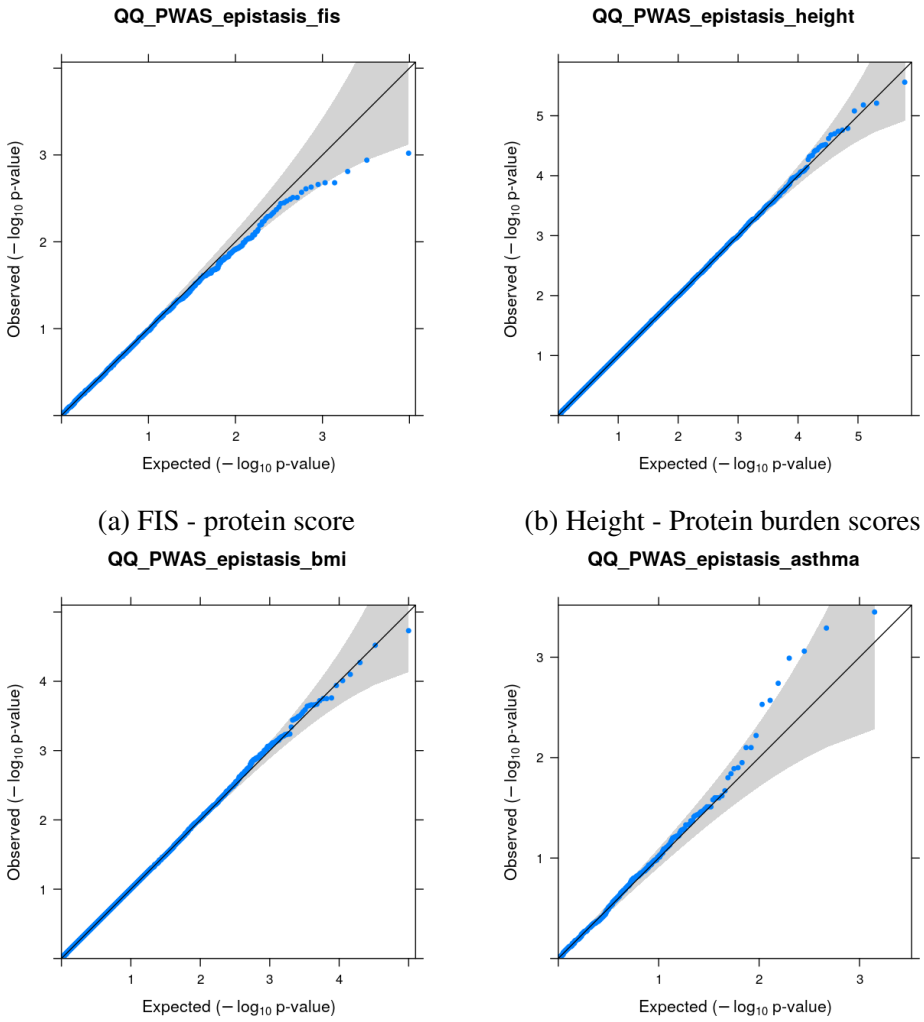
(c) BMI - SNP

(d) Asthma - SNP

Fig. 3.4 **QQ-plots visualising the p-values of the two-way interaction term for the post-QC analyses for the four UKBB traits in the SNP domain.** Grey area represents 95% confidence intervals.

There is an additional interaction detection method that tests if significant deviations exist from the expected allele frequencies in a contingency table conditioned on case status (Vittinghoff and Bauer, 2006). If the epistatic effect is real, then cases carrying the interacting alleles at both loci should be over-represented, relative to what would be expected from the alleles' additive effects. I applied this method to this putative interaction via Fisher's exact test for count data. The SNP pair remained significant with a p-value of $1.23 * 10^{-4}$.

As nearby markers' interaction association signal is expected to decay in proportion to their r^4 with the index pair (Wei et al., 2014b), I performed the same interaction test with proxies for the aforementioned index variants. As rs115122203 was imputed, to evaluate



(a) FIS - protein score

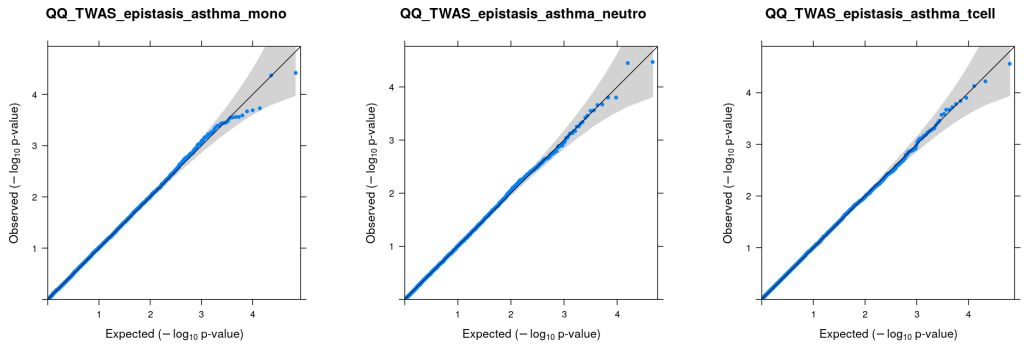
(b) Height - Protein burden scores

(c) BMI - protein score

(d) Asthma - protein score

Fig. 3.5 **QQ-plots visualising the p-values of the two-way interaction term for the post-QC analyses for the four UKBB traits in the protein score domain.** Grey area represents 95% confidence intervals

if the imputation process had affected the signal, I searched for a proxy for that SNP that was on the original genotype panel. I identified the best available proxies for both index variants, rs117893879 and rs61364965, which had an r^2 of 0.95 and 0.66 with rs117290331 and rs115122203, respectively. I repeated the interaction association test for this pair and obtained a p-value of $2.19 * 10^{-4}$. Then, I also performed the same interaction association test in the Test Set for both the index and the proxy pairs. I found that that neither of the Test Set tests were significant with p-values of 0.737 and 0.664 for the index and proxy tests, respectively. While the proxy pair's signal decay remained plausible, given that the best tagging proxy for rs115122203 had an r^2 of only 0.66 with the index, neither of the index



(a) Asthma - monocytes (b) Asthma - neutrophils (c) Asthma - t-cells

Fig. 3.6 **QQ-plots visualising the p-values of the two-way interaction term for the post-QC analyses for the asthma phenotype in the TWAS domain.** Grey area represents 95% confidence intervals

Cases

rs117290331	**rs115122203**		
	0	1	2
0	32,097 (0.948)	523 (0.015)	5 (0.0)
1	1,166 (0.034)	43 (0.001)	0 (0.0)
2	8 (0.0)	0 (0.0)	0 (0.0)

Controls

rs117290331	**rs115122203**		
	0	1	2
0	252,634 (0.954)	3,639 (0.0137)	16 (0.0)
1	8,242 (0.031)	106 (0.0)	1 (0.0)
2	63 (0.0)	1 (0.0)	0 (0.0)

Table 3.6 **Genotype count tables for the asthma phenotype for cases and controls.** The values in parentheses are proportions.

nor the proxy pairs replicated in the Test Set; thus, I concluded that this association is a false positive.

3.4 Cross-domain interaction tests

As I previously described in section 3.2.1, one of the benefits of aggregating SNPs on the gene-level is this may increase power to find novel signal that was not detectable in the source SNP data. The same phenomenon could also occur for interactions between the derived gene-level predictors and SNPs, which would conceptually represent statistical epistasis between individual variants and genes. To investigate if these types of interactions were present in my datasets, I performed interaction tests between SNPs and gene-level predictors.

3.4.1 Cross-domain filtering

As the signal for the gene-level predictors is a product of external data and the original SNP association signal, potential interactions between these domains could only offer unique insight if the gene-level predictors represent non-overlapping associations with their source GWAS signal. Therefore, I performed cross-domain filtering to eliminate all predictors that represented overlapping signal between the GWAS data and the derived gene-level predictors.

3.4.1.1 Gene filter for asthma TWAS and protein burden scores

I used the LD filtered subset of genes that also had an additive association FDR < 0.05 for the asthma phenotype to perform cross-filtering between the three TWAS tissue types to only keep the gene with the lower p-value. I applied the same filtering steps between the surviving TWAS predictors and the protein burden scores. This filtering process left 304, 236 and 283 TWAS gene-level predictors for monocytes, neutrophils and t-cells, respectively, together with 32 protein burden scores.

3.4.1.2 SNP-Gene cross-filtering

To ensure that only those SNP-gene interaction pairs are evaluated where the gene-score association signal was not driven by an underlying GWAS SNP that was also in the model, I employed the following filtering strategy. For each gene, I noted its additive association p-value (p_{gene}). Then, I located the gene's constituent SNPs, which were the variants that were weighted and aggregated into the gene-score. Among these, I identified the SNP with the lowest GWAS p-value ($p_{GWAS\_indexSNP}$). This SNP was either one of the constituent SNPs that was used to produce the gene-score, or the index SNP of an LD-clump, if it happened to belong to an LD-clump. Finally, I compared the strength of the signals between the GWAS and the gene-score to determine which one to keep by the following logic. If

$p_{gene} > p_{GWAS\_indexSNP}$, then I excluded the gene, otherwise I excluded all the SNPs that were used to build the gene-score instead.

As the new set of predictors were only filtered on recombination blocks individually before I integrated them, merging the datasets may have created new opportunities for the haplotype effect problem to arise again. Therefore, I once again applied a filter to remove variants or genes that were less than one cM apart in the integrated datasets. Table 3.7 summarises the end result of this filtering process.

phenotype	**number of SNP**	**number of protein scores**	**number of TWAS scores**
FIS	946	20	not used for TWAS
Height	1,613	192	not used for TWAS
BMI	1,622	73	not used for TWAS
Asthma	418	9	152

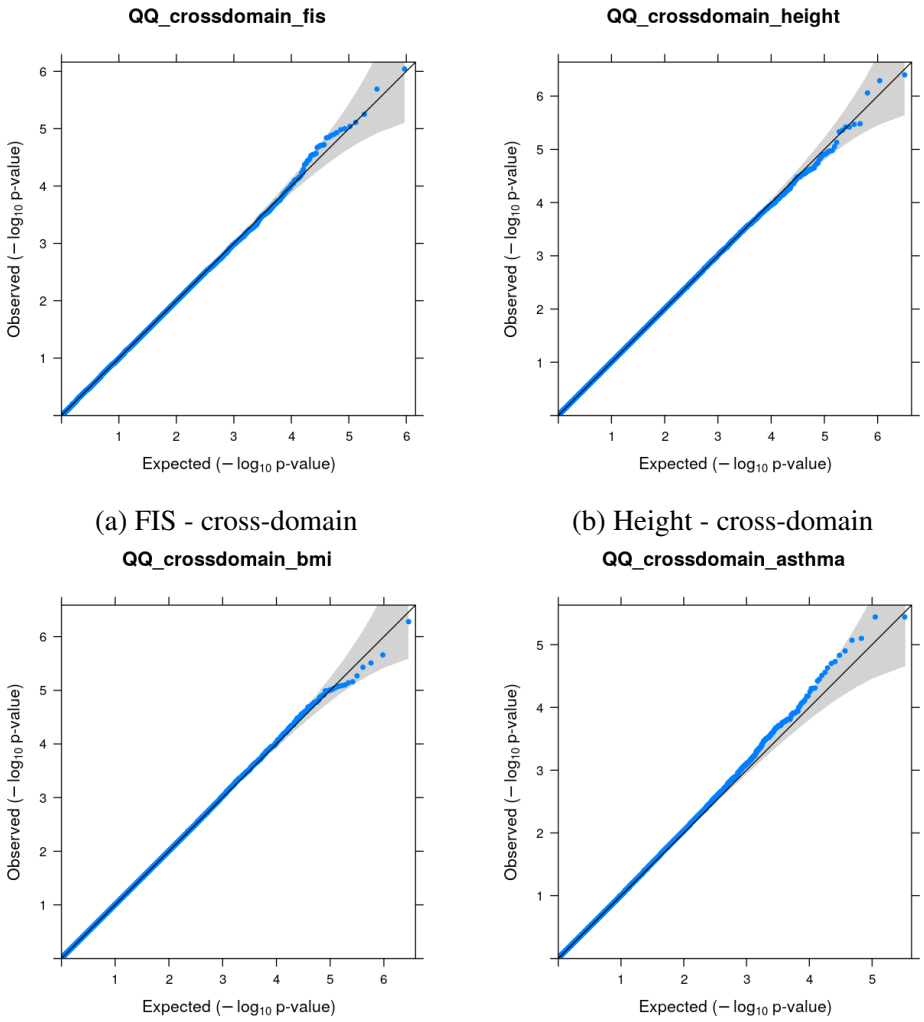
Table 3.7 **Summary of the cross-domain filtering process.**

3.4.2 Cross-domain interaction results

To search for interactions across domains, I performed the same test as described in section 3.3, along with the same post-association QC steps I detailed in section 3.3.1. My results are presented in Table 3.8 and Fig 3.7. I note that all the top associations occurred between SNPs, and aside from BMI, these were all identical to the SNP-only interaction tests I shown in 3.3. For the BMI experiment this differed only because the top SNPs were removed in the cross-filtering process.

Phenotype	**Cross-domain tests**	
	minimum FDR	number of tests
FIS	0.423	466,095
Height	0.419	1,628,110
BMI	0.762	1,435,665
Asthma	0.305	167,332

Table 3.8 **The results of the cross-domain two-way interaction tests for all four UKBB phenotypes.** The 'minimum FDR' column shows the lowest FDR observed in a given experiment, and the 'number of tests' column displays the total number of tests performed.



(a) FIS - cross-domain

(b) Height - cross-domain

(c) BMI - cross-domain

(d) Asthma - cross-domain

Fig. 3.7 **QQ-plots visualising the p-values of the two-way interaction term for the four UKBB trait cross-domain analyses.** Grey area represents 95% confidence intervals.

3.4.3 Summary of the UKBB interaction test experiments

I performed experiments to test for the presence of statistical epistasis using two different strategies. I evaluated the evidence in each of the genomic domains individually, and I also integrated these different views to perform cross-domain interaction tests between non-overlapping additive signals. After the application of filters to reduce the potential for false positives, all of my experimental results were consistent with the null hypothesis of no evidence for statistical interactions modulating phenotypic variance in any of the UKBB traits.

I realise that my QC filtering approach was highly conservative. Local (within recombination block) interactions may be biologically more plausible than those at least a recombination block apart (Wei et al., 2014a). Thus, I discarded information that may have contained genuine signal together with false positives. However, as there is no reliable way to distinguish loci that are only in physical linkage from those that also involved in biological function, my preference was to obtain fewer or no results of what may be considered as genuine epistasis. I define epistasis as 'genuine' that arises from the way information is stored in the genome, rather than what is generated by the physical properties of the DNA molecule. An alternative strategy would have been to instead of removing one variant in each pair that were within the same block to only remove interaction tests of pairs that were within the same block, and to allow variants to interact with others outside of their recombination blocks. However, given that the overall objective of my work was to compare standard methods against neural-network based models on the same datasets, I could not do this as such a per-interaction filtering is not feasible within the neural-network framework. Finally, neural-networks perform better with fewer predictors and a larger number of samples; thus, keeping a larger number of predictors would not have been feasible for this reason either.

There are several other possible explanations for the lack of positive results. Despite the large sample size of the UKBB, I may still not have had adequate statistical power to detect epistasis. It is also possible that my power would have been sufficient; however, the SNPs involved in the interactions were either not imputed or were filtered out by my initial QC steps. Finally, it is also possible that statistical epistasis does not contribute to phenotypic variance in any of the four UKBB traits.

3.5 Interaction tests in the IBD datasets

As the IBD datasets were an order of magnitude smaller than the UKBB, I believed that an exhaustive search, even after pre-filtering on additive effects, was not a feasible approach. Therefore, I decided to pursue a hypothesis-driven approach that utilised a biological prior to reduce the search-space for epistasis. As this prior assumed haplotype-specific interactions, before describing my analysis, I will also provide the necessary background on haplotype phasing in the following sections.

My overall analysis involves fitting regression based models on phased SNP data, to infer the existence of haplotype-specific interactions between variants. I will describe in detail each stage of my analyses for interaction detection in the subsequent sections; however, I will first outline my overall strategy here, so that each individual component's role may be better understood in the overall scheme. My analysis consists of the following three steps:

1. Collate association summary statistics to identify plausible missense and eQTL signals (section 3.5.3.1).
2. Phase haplotypes to obtain information on the missense and eQTL variants' chromosomal arrangement (section 3.5.3.2).
3. Evaluate two statistical models that have the ability to detect haplotype-specific statistical epistasis (section 3.5.4).

3.5.1 Biological insight to reduce search-space

A recent study by Castel et al. (2018) indicated that interactions may be more easily detected where a cis-eQTL allele modulates the expression of a gene which has a nearby missense allele on the same chromosome. Fig 3.8 illustrates this hypothesis graphically. They successfully deployed this strategy to infer epistasis both indirectly in the population, by observing that deleterious haplotypes were removed by purifying selection, and also in cancer and autism patients where they found an enrichment of deleterious haplotypes. Inspired by their results, I thought that a similar approach may be a viable strategy to identify statistical interactions that increase susceptibility to IBD.

3.5.2 Statistical haplotype phasing

3.5.2.1 The definition and the utility of haplotype phase

Obtaining the chromosomal arrangement of alleles by separating the nucleotide content of an individual's maternally and paternally derived chromosomes is known as phasing. The information obtained by phasing, termed the haplotype, has utilities for imputation, calling of genotypes, detecting genotyping errors, inferring demography, studying recombination events and the detection of signatures of selection (Browning and Browning, 2011).

3.5.2.2 Overview of phasing methods

Currently used methods to obtain phase may be broadly organised into two categories. The first group consists of specialised experimental methods that assemble haplotype contigs (series of overlapping DNA sequences) from sequence reads. The second category contains computational approaches that aim to infer the underlying haplotypes that generated the observed genotypes by using a phased reference population. Given that all of my experiments relied solely on in-silico analyses, I will only cover approaches that belong to this latter category.

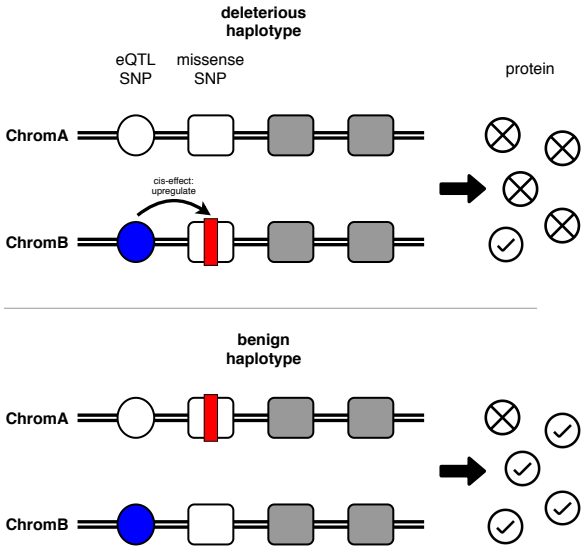


Fig. 3.8 **Missense-eQTL schematic diagram** Top: Illustration of the haplotype-specific interaction effect between a missense variant and a cis-regulatory SNP. In the deleterious haplotype configuration, the missense and the eQTL upregulatory alleles are on the same chromosome which results in an increase of the faulty gene product. Bottom: a benign haplotype configuration, where the hypothetical individual carries the same alleles, but not on the same chromosome, which would result in a greater abundance of the normal gene product.

3.5.2.3 Statistical Methods

Due to their relative speed and low cost, most large-scale phasing efforts currently rely on computational methods. As most current techniques produce allele dosage estimates, these statistical methods work by estimating the true underlying haplotype configurations that generated the observed genotypes. These methods will be described in the next two sections.

3.5.2.4 Trio and pedigree based phasing

In the simplest case, where parental genotype information is available, and the only interest is to obtain phase information for the child, then short range haplotype information may be derived by performing genetic analysis (Marchini et al., 2006). Genetic analysis involves tracing all alleles' origins, relying on Mendel's law of segregation that states that each gamete receives only one allele. This analysis assumes no recombinations, and that at least one individual is homozygous for the target markers. To obtain phase information on parents and for whole chromosomes, more complex methods and larger families (with at least four children or multiple generations) are needed (Roach et al., 2011). The practicalities of recruiting such individuals into studies limits the utility of pedigree based phasing;

therefore, most studies rely on population-based phasing of unrelated individuals that utilise the framework of hidden markov models.

3.5.2.5 Hidden markov model based phasing

The realisation that haplotype distributions are more realistically represented by basing them on approximate coalescent models (Li and Stephens, 2003), gave rise to Hidden Markov Model (HMM) based phasing methods. These models capture the fact that new haplotypes are derived from old haplotypes by the processes of mutation and recombination. As such events are rare, over short distances a given individual's haplotype may be estimated from genetically similar individuals' haplotypes (Stephens et al., 2001).

HMMs assume that a markov process generates a sequence of underlying hidden states that emit observations. A key property of this model is that it is memoryless, only the current state and current observation affect transition probabilities between states. In the context of haplotype phase inference, these hidden states represent the underlying true haplotypes, and the observations represent the genotypes of an individual. Therefore, HMMs seek to find the most likely haplotype configuration that generated the observed genotype as

$$G = h_1 + h_2, \tag{3.4}$$

where G denotes the observed genotype, and h_1 and h_2 denote the first and second haplotypes, respectively. Due to recombination events, observed genotypes are modelled as an imperfect mosaic of 'template haplotypes', which are a subset of sampled haplotypes from a reference dataset. Therefore, the probability for the phase of S set of markers is given by (Delaneau et al., 2012):

$$S = p(D|G',H). \tag{3.5}$$

In words, phase is the probability of the haplotype pair, the diplotype (D), conditioned on a pool of haplotypes (H), which are also consistent with the observed genotypes of the to-be phased population (G').

SHAPEIT, the currently most widely used phasing method for large scale data (Bycroft et al., 2017), achieves further performance gains by several algorithmic tweaks. Like its predecessor, PHASE (Stephens et al., 2001), it breaks the genotypes into disjoint segments of 5-8 SNPs. The most probable haplotype for each of these segments is then determined, and then ligated together to produce a complete haplotype. The key innovation of SHAPEIT lies in how these compatible haplotypes are considered. Instead of maintaining a full list of all possible complete haplotypes, the same information is represented in a haplotype binary tree. Here, each node is a haplotype segment that consists of a heterozygous SNP and all the

homozygous markers before the next heterozygous SNP. These nodes have two children that represent the two possible switch orientations with the next segment. In this representation, complete haplotypes are captured by valid paths from the tree's root to a leaf node. Such a tree would still grow exponentially with the number of heterozygous SNPs; therefore, to further reduce the complexity of the algorithm, SHAPEIT applies a pre-specified threshold to prune highly unlikely branches to build an incomplete haplotype tree instead (Delaneau et al., 2008). As this graphical model still represents most possible haplotypes, the HMM only needs to estimate the hidden states for the segments, not individuals markers (Delaneau et al., 2012).

3.5.2.6 Phasing summary

Phasing methods are used to identify alleles that are co-located on the same chromosome. Currently, the preferred way to obtain phase at scale, is through the application of statistical methods that utilise large-scale haplotype reference panels such as the HRC (Zheng et al., 2016). A key limitation of current population-based computational approaches is that they are not able to phase rare variants that were not present in the reference panel.

3.5.3 Genotype and summary data

I obtained the summary statistics of the fine mapped IBD associations that my experiments relied on from the Huang et al. (2017) and de Lange et al. (2017) studies. The eQTL summary data that I used to find relevant SNPs that had an eQTL result with the IBD genes (defined in section 3.5.3.1) were sourced from the same BLUEPRINT data that I described in section 3.2.1.2 (Chen et al., 2016), together with two other sources, which were the CEDAR database (Momozawa et al., 2018) and the eQTLGen database (Võsa et al., 2018). The cell-count QTL summary data that I used to cross-check my eQTL variants against known cell-count QTLs was sourced from the database by Astle et al. (2016).

I performed these analyses earlier during my PhD than the data QC work I described in Chapter 2; therefore, I relied on a different version of the same genotype datasets that I described in section 2.2.3. Specifically, I was given access to the same data that was used to publish the study by de Lange et al. (2017). As I wanted to stay close to the workflow that led to the published results, I adopted the same model fit strategy as the authors of that study. An important difference in our workflows was that they treated the disease status as binary phenotypes in a logistic regression model, as opposed to regressing out covariates ahead of the main analysis, like I did in Chapter 2.

3.5.3.1 Collating summary statistics for IBD

I began by identifying all IBD-associated missense variants with a posterior probability of causality greater than 0.5. The criterion that the variants must be fine mapped was important, as the hypothesis that I was interested in relied on the assumption that a missense variant yielded a faulty-protein product that increased risk of IBD; hence, I needed to be reasonably certain that these SNPs were indeed increasing IBD risk by affecting protein coding genes. I identified 13 such missense variants. Then, I selected eQTL SNPs with the lowest association p-value for the 13 genes matched to these 13 missense variants via the eQTL databases I described in section 3.5.3. There were 37 such SNPs, which meant that there were more eQTL variants than missense SNPs. Their median and maximum eQTL p-values were $4.07 * 10^{-17}$ and $2.93 * 10^{-5}$, respectively, and the average number of eQTLs per missense SNP was 2.84 with a standard deviation of 1.57. The most common tissue types were T-cell (14) and whole blood (13), and the least common tissue type was monocyte (3).

One important consideration for an analysis where the hypothesis pursued relies on the effect of cis-eQTL SNPs, is a potential confounding mechanism where the alternative allele would modulate expression levels not by regulating transcription levels in individual cells, but rather indirectly, by regulating the total number of cells. To reduce the possibility for this confounder, I cross-checked each of the 37 eQTL SNPs in the summary statistics provided by Astle et al. (2016) against confirmed cell-QTL associations. I found that none of the 37 eQTLs had evidence of also being cell-count QTLs.

3.5.3.2 Obtaining haplotype configurations

To infer if deleterious haplotype configurations increased risk for IBD, I needed to phase the variants involved. To begin, I first had to exclude missense and eQTL SNP pairs that had a $D' > 0.95$, as variants failing this criterion would have made haplotype-specific regression models problematic due to (near) collinearity (a $D' = 1$ would have indicated that only three of the four possible haplotype configurations exist (Slatkin, 2008)). There were 21 SNP pairs that passed this criterion. Next, to increase the number of variants that the phasing algorithm may use to infer the correct configuration of my targets, I added an extra 500 SNP support window on each side around the missense and eQTL variants. Thus, the final segment included an additional 500 SNPs on each side, plus all variants between the missense and the eQTL pair. Finally, I phased these extracts using SHAPEIT3 (Delaneau et al., 2012) to obtain phase information on my target pairs.

3.5.4 Two statistical models to evaluate haplotype-specific interaction effects

In the next two sections, I will describe the two regression based methods that I used to test the hypothesis that IBD risk is increased by the presence of a deleterious haplotype that consisted of an eQTL upregulating allele and a missense allele.

3.5.4.1 '#Bad haplo' model

This model extends the same interaction model that I described in section 3.3 with an extra term that captures the haplotype-specific interaction effect:

$$logit(Y) = \beta_m G_{missense} + \beta_e G_{eQTL} + \beta_{me} G_{missense} * G_{eQTL} + \beta_B B + e, \tag{3.6}$$

where Y, $G_{missense}$, G_{eQTL} and e denote the phenotype column vector, the missense SNP, the eQTL SNP and a random noise term, respectively, and the βs are their corresponding coefficients. The new B term captures the number of deleterious haplotypes. I determined this value for each pair of SNPs for each individual, from the phase I obtained in section 3.5.3.2 by counting the number of times an individual had a missense allele and an eQTL-increasing allele on the same chromosome. Thus, the number of possible values for the B term were $\{0,1,2\}$.

Fig 3.9 shows a hypothetical example of the phenotype column vector and the design matrix for the '*#bad haplo*' model, which may be useful to illustrate a few additional properties of this model. Only the double heterozygotes contribute new information relative to the reduced model that would only include the standard genotype interaction term (eq 3.3), as the B term can be obtained from the combination of the $G_{missense}$ and G_{eQTL} terms for individuals homozygous at either locus.

$$\begin{array}{c} \\ Indi_1 \\ Indi_2 \\ \vdots \\ Indi_n \end{array} \begin{array}{c} \begin{array}{c} Y \end{array} \\ \left[\begin{array}{c} 0 \\ 1 \\ \vdots \\ 0 \end{array}\right] \end{array} \left[\begin{array}{cccc} G_{missense} & G_{eQTL} & G_{missense} * G_{eQTL} & B \\ 1 & 2 & 2 & 1 \\ 2 & 2 & 4 & 2 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 1 & 0 \end{array}\right]$$

Fig. 3.9 **Hypothetical example for a phenotype column vector and design matrix for the *'#bad haplo'* model for n individuals.** Intercept omitted for clarity but was present in the model fit.

3.5.4.2 'haplo regression' model

An alternative regression model where individuals were split into two observations (one for each of their homologous chromosomes) was also evaluated. Here, I fit the same model described by eq 3.3, with the only difference that the predictors were now haplotypes instead of genotypes:

$$logit(Y) = \beta_m h_{missense} + \beta_e h_{eQTL} + \beta_{me} h_{missense} * h_{eQTL} + e. \tag{3.7}$$

where Y, $h_{missense}$, h_{eQTL} and e denote a phenotype column vector, the missense haplotype, the eQTL haplotype and a random noise term, respectively, and the βs are their corresponding coefficients. The advantage of this model is that it only requires three terms, as the third term captures the haplotype-specific interaction effect directly. To illustrate the details of this model further, consider the hypothetical example of a phenotype column vector and a design matrix shown in Fig 3.10.

$$\begin{array}{c} \\ Indi_1 \\ Indi_1 \\ \vdots \\ Indi_n \\ Indi_n \end{array} \begin{array}{c} \begin{array}{c} Y \end{array} \\ \left[\begin{array}{c} 0 \\ 0 \\ \vdots \\ 1 \\ 1 \end{array}\right] \end{array} \left[\begin{array}{ccc} h_{missense} & h_{eQTL} & h_{missense} * h_{eQTL} \\ 1 & 1 & 1 \\ 0 & 1 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{array}\right]$$

Fig. 3.10 **Hypothetical example for a phenotype column vector and design matrix for the haplotype regression model for n individuals.** Intercept omitted for clarity but was present in the model fit.

A disadvantage of this model is that as humans are diploids, there is only one unique phenotype for both chromosomes. Therefore, both haplotypes have to share the same outcome, which creates a situation where all the individuals form two-observation clusters from their own two chromosomes. Such an artefact could cause an artificial deflation of variance estimates in a regression based model. To account for this artefact, I applied a Huber-White cluster variance correction procedure (Williams, 2000) as a post-processing step via the 'rms' R package (Harrell Jr, 2019), which considered each individual as a cluster of two.

3.5.4.3 Results for the haplotype-specific interaction tests

I applied both the '*#bad haplo*' and the '*haplo regression*' models to each of my IBD datasets. I also considered the same covariates as the de Lange et al. (2017) study did, which were *sex*

and the first ten PCs. Then, I used the R package 'meta' (Balduzzi et al., 2019) to perform generic inverse variance fixed-effects meta-analysis to aggregate evidence from all of my IBD datasets. The results from this analysis are presented in Table 3.9.

SNP pair			**'haplo regression'**		**'#bad haplo'**	
gene	missense	eQTL	p	coef	p	coef
SLAMF8	rs34687326	rs75087057	0.299	0.212	0.144	0.391
SLAMF8	rs34687326	rs2501342	0.726	0.027	0.899	0.015
PLCG2	rs11548656	rs8059316	0.128	0.186	0.185	0.234
PLCG2	rs11548656	rs145841372	0.950	-0.016	0.610	-0.147
IL23R	**rs41313262**	**rs2064689**	$\mathbf{2.69 * 10^{-6}}$	**-1.355**	0.048	0.779
PTPN22	rs2476601	rs17464525	0.776	-0.036	0.680	0.062
SMAD3	rs35874463	rs10152593	0.955	0.006	0.258	-0.171
SMAD3	rs35874463	rs8023420	0.230	0.113	0.432	0.113
SMAD3	rs35874463	rs6494626	0.672	-0.036	0.514	-0.079
SMAD3	rs35874463	rs10163040	0.166	-0.163	0.570	-0.092
NOD2	rs2066844	rs1420685	0.701	0.060	0.449	-0.132
NOD2	rs2066844	rs1981760	0.315	-0.192	0.885	-0.029
NOD2	rs2066844	rs4785448	0.412	0.143	0.817	-0.044
NOD2	rs2066845	rs1420685	0.514	0.150	0.823	-0.066
NOD2	rs2066845	rs1981760	0.548	-0.143	0.905	0.036
NOD2	rs2066845	rs4785448	0.689	-0.094	0.192	-0.394
NOD2	rs5743271	rs1981760	0.135	-0.632	0.420	-0.402
NOD2	**rs5743271**	**rs4785448**	$\mathbf{1.24 * 10^{-3}}$	**1.345**	0.641	0.242
PLCG2	rs11548656	rs56704282	0.158	0.173	0.664	0.074
SNAPC4	rs3812565	rs531538571	0.038	-0.223	0.464	-0.109
SNAPC4	rs3812565	rs76179734	0.430	-0.033	0.559	-0.033

Table 3.9 **Results for the two-way interaction tests between the missense and eQTL SNPs for both the '*haplo regression*' and '*#bad haplo*' models.** Values in the 'p' column show association p-values for the haplotype-specific interaction term and values in the 'coef' column show their corresponding coefficient estimates.

3.5.4.4 Post-association QC and discussion of haplotype-specific interaction tests

There were two significant associations in the '*haplo regression*' model, and none in the '*#bad haplo*' model. As the former model requires one less parameter to estimate, in theory, it is possible that it captured associations that the other model could not. However, as it also required a post-processing step to adjust its variance estimates, it may also have been susceptible to artefacts that arose from this procedure. Therefore, I decided to examine the

two pairs of associations (rs5743271, rs4785448) and (rs41313262, rs2064689) in greater detail. I recovered the original, unadjusted p-values of these two interactions, and I found that they were far from significant at 0.8137 and 0.8134, respectively. This already suggested an artefact, as the p-values usually only change towards the other direction, increase slightly due to larger estimated error variances, as a result of the Huber-White adjustment. Next, I examined the haplotype counts for both pairs, and I found that the interaction effect ($h_{missense} * h_{eQTL}$) had very low counts for both associations. The case/control haplotype counts were 0/1 and 0/6 for (rs5743271, rs4785448) and (rs41313262, rs2064689), respectively. As the Huber-White method relies on asymptotic assumptions to adjust variance estimates, such a low number of observations were consistent with an artefact that could induce false positives. After eliminating these two associations, I concluded that no interaction tests were found to be significant after post-association QC. I also have to note that all of the pairs were within the same recombination block; therefore, even if these pairs were found to be not due to technical errors, without fine mapping the regulatory variant the effect may still have been caused by the haplotype-effect artefact described by Wood et al. (2014).

There are several possible explanations for the null results of my analyses. It is possible that I did not have enough power in my datasets to detect statistical interactions between missense and eQTL variants. It is also possible that the power would have been sufficient to detect such interactions, but the combination of SNPs were not available in the panel of SNPs I had access to. Additionally, I may not have considered eQTLs from the relevant cell-types or tissues. Finally, I also have to acknowledge the possibility that haplotype-specific interactions between coding and regulatory variants may not contribute to susceptibility to IBD.

3.6 Concluding remarks

In this chapter I searched for two-way interactions using standard statistical methods involving both hypothesis-free approaches, and analyses that employed a biological prior. I was unable to find evidence in any of my experiments of credible statistical interactions that also survived my QC steps that eliminated variants where the interaction could also have been induced by haplotype effects. As I already covered in their respective sections, the reasons for this could have been a lack of power, or that interacting markers were not in the model, and finally, that statistical interactions do not contribute to phenotypic variance in any of the traits that I examined.

Detecting epistasis in a robust, consistent manner remains an enduring challenge in the field of human genetics. This is in contrast with GWAS, where after the initial protocol was

established over ten years ago (Anderson et al., 2010), the number of confirmed associations has been growing exponentially during the last decade (Visscher et al., 2017). On the other hand, progress in epistasis detection during the same period has been very limited. Confirmed findings of statistical epistasis have been few and far in between, and results have been marred by false positives (Wood et al., 2014) and retractions (Rhinn et al., 2015). Genuine findings appear to be more the exception rather than the rule in the endeavour of epistasis detection. Thus, now I see the Castel et al. (2018) study as one of the isolated successes, rather than the identification of a general principle to could help epistasis detection more broadly. Indeed, I have not seen any other studies that applied their strategy successfully to other traits. As evidence for a substantial contribution of non-linear genetic effects to phenotypic variance has been scarce at best, I see my own negative findings in this chapter as congruent with the broader field.

I did achieve my main objectives however, which was to prepare datasets with an appropriately low dimensionality, and also to perform standard statistical tests that may serve as a frame of reference for my neural-network based approaches in the next chapter. A sufficiently low dimensionality was an important objective, as neural-networks do not cope well with a high number of input features, nor do they provide the same level of control over individual predictors for QC (for example, it is not feasible to selectively exclude tests for variant pairs in the neural-network framework).

As for the future of epistasis detection using standard methods I make the following remarks. As one of the most important factors of statistical epistasis detection is power (Wei et al., 2014b), one potential future trend that may offer hope is the expected increase in sample size offered by upcoming large population cohorts. These cohorts include the *5 million genomes project* (GEL, 2020) in the UK and the '*All of Us*' biobank project in the USA (The All of Us Research Program Investigators, 2019). With the order of magnitude of increase in sample sizes that these cohorts will bring, it is possible that we may see a similar increase in positive findings that accompanied the increase of GWAS sample sizes from individuals in the low thousands to the ~100K scale.

Chapter 4

Prediction and inference on non-linear genetic effects using neural-networks

4.1 Chapter 4 outline

Parallel to my work described in the preceding chapter where I used regression based methods to search for statistical epistasis, I also explored the potential of neural-network (NN) based methods to infer evidence of epistasis indirectly. NNs perform a non-exhaustive random search for interactions between the input features, and their performance is evaluated by predictions in held out data, which is a different standard of evidence than the null hypothesis testing approach that regression based methods rely on.

Section 4.2 reviews relevant previous work, and provides the necessary background on the NN architectures and algorithms that I used in the rest of the chapter. Section 4.3 describes how these methods were applied to synthetic data, where I confirmed via a large-scale simulation study the potential of NNs to be able to infer interactions at a higher accuracy than standard regression based methods. Section 4.4 covers the application of the same NN approaches to the cohorts I prepared and analysed in Chapter 3.

4.2 Neural-networks in genomics

4.2.1 Relevant previous work

The key advantage of NNs is that they can approximate complex non-linear functions and model higher-order interactions between input features, without performing an exhaustive search. For NNs to perform well, there must be a substantial non-linearity in the underlying

problem, and the training data needs to be sufficiently large for the NN to learn this non-linearity (Hestness et al., 2017).

As I described earlier in the Introduction (section 1.1.2), non-linearity (epistasis) may occur between individuals in the form of statistical epistasis, or within a genome in the form of functional epistasis. To date in genomics NN have been most successful in inferring functional epistasis. Here, NN classifiers are trained to learn the relationships between labels (such as TF binding) and DNA sequence context. A successful example in this area was the 'DeepSEA' network (Zhou and Troyanskaya, 2015), which was a shallow convolutional neural-network (CNN) classifier that could accurately predict the presence or absence of regulatory elements in a given nucleotide sequence. A similar model was more recently applied to quantify the consequence of non-coding mutations to autism spectrum disorders by Zhou et al. (2019).

Applications of NNs to infer the existence of statistical epistasis have been less convincing. Here, the disease phenotype is modeled directly from genetic differences between individuals; thus, these models are the NN analogues of PRS generation methods. A few recent applications in the field of agricultural science showed early promise. A NN model by Ma et al. (2017) outperformed baseline linear methods in bread wheat yield prediction by up to 65%. Another NN effort found a ~24% improvement over linear methods by deploying a locally connected CNN model to predict plant traits (Pook et al., 2020). However, similar reliable positive results in humans have been lacking. When I started my PhD, most relevant studies relied on small GWAS cohorts and pre 'deep learning' era models (Motsinger-Reif et al., 2008). Parallel to my own work, in the last few years there have been a number of attempts that relied on larger datasets and more modern NN models.

One of the earliest attempts from the more modern efforts was a study by Montañez et al. (2018), which aimed to predict the obesity phenotype using a NN based PRS. They claimed that from a sample size of only ~2000 individuals their NN model was able to predict obesity with a near 100% accuracy, a trait which is not a 100% heritable. Additional issues of this study included a lack of comparison to baseline linear methods, and also that they did not consider haplotype effects. Their choice of 2,465 SNPs was selected based on a simple additive association p-value filtering step with no consideration given to LD. There were another two more recent recent studies that utilised the UKBB cohort, which were also more comparable in scope to my own work.

The study by Bellot et al. (2018) used the smaller interim UKBB release (~150K individuals) on quantitative traits (including height and BMI) to investigate the relative performance of NN and linear methods for building PRS. A notable difference between their approach and mine was their treatment of LD when selecting SNPs to be included in the PRS. Instead

of trying to eliminate haplotype effects, which could create non-biologically meaningful statistical epistasis via filtering (Wood et al., 2014), they took the position that incorporating them into the model may improve prediction. As a consequence, their NN models were smaller (maximum 128 neurons), but the number of SNPs considered was larger (~10-50K) than what I had in my analyses. However, their approach of incorporating LD, correlations between SNPs (a linear effect), into prediction was a very different goal than my own objective to find evidence for non-linear genetic effects that contribute to phenotypic variance. They found that their NN models did not outperform the linear baselines; however, even if they would have, their approach could not have been used to find evidence for epistasis, as any potential benefits over the linear baselines could also have been due to haplotype effects. Indeed, if the main benefit of NNs would be to improve PRS by modelling LD, I would argue that it would be more effective to use LDpred instead, as the latter method can accommodate ~1 million markers (Vilhjálmsson et al., 2015), rather than just ~50K, the maximum number that their study considered.

The most recent study that made use of the full UKBB cohort, investigated the potential of NNs to build PRS for blood cell traits (Xu et al., 2020). This study was more similar to my own project with respect to variable selection. However, instead of filtering on a recombination block basis, they used conditional analysis to only keep SNPs that represented unique signal. This filtering process left their study with fewer SNPs (160 - 762) than what I used for my own analyses (450 - 1,732). It is important to note, that their variable selection approach may not have fully controlled for haplotype effects, as they found that when they removed SNPs in their models close together on chromosomes 3, 6 (HLA) or 16, this resulted in a deterioration in prediction accuracy of PRS built by their multivariate linear models (that did not include the interaction terms). Subsequently, when explicitly tested, the interaction terms of these removed variants were found to be significant. Taken together, the behaviour of these SNPs is consistent with haplotype effects that could generate spurious statistical interactions as described by Wood et al. (2014). The conclusion of their study was also similar to that of Bellot et al. (2018), as they found that NNs did not perform better than linear PRS building methods.

A common limitation of all previously described studies that aimed to build PRS with NNs was the attitude they took on using NNs to infer the nature of potential non-linear effects. Specifically, they asserted that a NN could outperform conventional approaches because it would learn non-linear genetic effects. However, just because NNs are capable of learning non-linear functions, there is no guarantee that they will detect non-linear effects if they are too weak, or especially, if there are none. Also, even if NNs did learn non-linear effects, unless addressed on explicit terms, there is no guarantee that they were not due to haplotype

effects, or some of the other known artefacts that could generate statistical epistasis (Fish et al., 2016; Wood et al., 2014). Finally, none of the previous studies so far attempted to look for non-linear effects across genomic domains nor have they tried to perform inference using NNs to identify individual interactions.

4.2.2 Opportunities and challenges for NNs in genomics

The reasons why fully-connected NNs (FNNs) and CNNs have faced difficulties in genetic prediction lie in the differences between the domains in which NNs have been applied to successfully, such as images, and genomics. Image labels are predicted from pixel data, where convolutional layers may learn reusable features (I discuss convolutions in more detail in Appendix B in section B). However, convolutions do not fit GWAS data well, and reusable filters are unlikely to be learned from SNPs. The subtle reasons for this have to do with how GWAS SNP data and DNA sequence data (or pixel data) differ. In the domain of sequence classifiers, convolution filters may represent regulatory motifs. However, as SNP data only captures deltas from a reference genome, a series of SNPs do not carry any intrinsic meaning as their sequence context is not preserved. To clarify, a $3x3$ filter in an image classification task may represent an edge. Similarly, for a sequence classifier a motif pattern (example: '*GCA*') may represent a (partial) transcription factor binding site, both of which are reusable features that are likely to carry the same meaning in other areas of the input. In contrast, a pattern of three SNPs (example: '021') with the same values is unlikely to have the same meaning elsewhere. Two SNPs next to each other in the genotype matrix may refer to different types of nucleotides with an arbitrary distance between them. Indeed, both studies (Bellot et al. (2018); Xu et al. (2020)) that empirically evaluated the applicability of convolutions to SNP data found that CNNs perform at a level below FNNs.

An additional important difference between image classification and genetic prediction is that an image's label depends only on the pixels in the image itself; however, complex diseases are never a 100% heritable, and what is not heritable is also not predictable from SNP data. Therefore, the predictive ceiling is correspondingly lower (broad sense heritability).

Finally, images are typically made up of only a few thousand pixels (after downsampling), whereas SNP data ranges from the hundreds of thousands to millions of predictors. In addition, while image data is abundant and cheap, genomic data is still relatively scarce and expensive. These last two factors mean that typical image classification problems have very high samples to predictors ratios (an essentially infinite sample size), whereas in most GWAS there are far more SNPs than individuals. This poses a continued challenge for NNs, as state of the art PRS typically consists of ~500K SNPs (Khera et al., 2018), a dimensionality

that would seem difficult for FNNs of the current generation. Table 4.1 summarises the differences between the typical domains where NNs succeed, such as images, and genomics.

	Image classification	**Genomics**
n >> p	YES	NO
prediction ceiling	100%	H^2
noise	NO	1 - H^2
typical predictive accuracy	>95%*	< 10% *
main challenge	problem complexity	low power

Table 4.1 **Summary of the differences between typical image classification and genetic prediction tasks.** *n* is the number of observations and *p* is the number of input features. H^2 is broad-sense heritability. *Accuracies taken from He et al. (2016) and Lee et al. (2018) for images and PRS, respectively.

In summary, the challenges that NNs face in genomics originate from the differences in the domains where they traditionally excel, and typical genetic data. However, given their strengths at modeling highly non-linear functions, they also offer potential for revealing the higher-order encoding of genetic information of complex traits. This attribute, taken together with the application of careful QC measures, and the recent availability of large genotyped cohorts such as the UKBB dataset, offer new hope that the challenges that held NN back in human genetics so far can be surmounted.

4.2.3 Neural-network models and data preparation

The NNs used in this chapter were all FNN based models that I described in detail in the Introduction in section 1.7.2. To ensure model stability, and to speed up training, all input data (genotype and gene-level predictors) and phenotypes were standardised to have a zero mean and unit variance based on the training set (LeCun et al., 2012).

4.2.3.1 Choosing the model architecture

A NN model's hyperparameters such as regularization, activation functions, the number, type and size of layers, are collectively known as the architecture of the NN. The ideal architecture is determined by the complexity of the function modeled (the number of input features and the degree of non-linearity between them), and the limits of the computational resources available (the RAM, GPU capabilities and time limits on a computing cluster).

I believe that, if possible, it is better to let the data guide the model hyperparameter selection, rather than to impose my own prior beliefs. Therefore, the only restriction I set on

parameter	range	type	description
first layer size	[100 - 4000]	int	number of neurons in the first layer of the network
epochs	[10 - 100]	int	number of training iterations
#hidden layers	[1 - 20]	int	number of hidden layers between input and output
dropout	[0 - 0.9]	real	percent of neurons to be deactivated at each iteration on all layers
learn rate	$[10^{-5}$ - 0.01]	real	the learning rate at which parameters are changed between epochs
activation layer	SELU / linear	logical	if the non-linear capacity of the network is enabled

Table 4.2 **Summary of the search-space covered by the hyperopt tool.** The SELU activation function is described in the Introduction in section 1.7.4.4.

the NN architecture was that each subsequent hidden layer would be half the size of the one preceding it (a common design pattern employed to reduce the number of hyperparameters of a FNN model). The rest of the hyperparameters were determined by employing a semi-random search performed by the package '*hyperopt*' (Bergstra et al., 2013). I defined a range of possible hyperparameters (Table 4.2), which I based on the limits of the computational resources available, such as the RAM and time limits on the computing cluster. Then, I defined the r^2 between predicted and observed outcomes in the validation set as the criteria for hyperopt to optimise on. Hyperopt was then set to find the best performing hyperparameters in a pre-specified number of trials, which for computational time limit considerations, I set to 50. All models were trained via the ADAM optimizer that I described in the Introduction in section 1.7.4.1 (Kingma and Ba, 2014). Finally, to reduce overfitting, I also employed the early stopping mechanism (Prechelt, 1998) by recording the epoch at which the NN performed at the highest accuracy on the validation set, and after the initial training finished I retrained the model up until this epoch. The model weights for this best performing epoch were saved and reused for the subsequent evaluation on the Test Set. Expression 4.1 summarises the overall NN architecture in a shorthand notation:

$$NN : [In, FC_1, \sigma, DO, FC_2, \sigma, DO, \ldots, FC_k, \sigma, DO, Out], \tag{4.1}$$

where FC, DO and σ denote the k fully-connected layers, the dropout layers and the SELU activation functions, respectively.

4.2.4 NN methods used in this chapter

4.2.4.1 Using NNs to evaluate the evidence for non-linearity

As previously described in section 4.2.1, a common shortcoming of projects similar to mine were assertions about the supposed non-linearity the NN models would have learned from SNP data (Montañez et al., 2018; Xu et al., 2020). However, without explicitly assessing the model for non-linearity such claims lack evidence.

To test if my own NN models have learned any non-linear effects from the data, I evaluated the following two-tier strategy. The previously described hyperopt model selection process had the option to choose between a linear and a non-linear activation function for better prediction performance on the held-out validation set. Additionally, if a non-linear solution was selected, I also evaluated the final model with and without the non-linear function enabled, by turning on and off the activation layers for the final test prediction. To demonstrate why this removes any potential non-linearity from a NN model, consider the original NN equation I presented in the Introduction (1.39):

$$Y = \sigma_k(\dots\sigma_2(\sigma_1(\mathbf{X}\mathbf{W_1})\mathbf{W_2})\dots W_k). \tag{4.2}$$

where Y, $\mathbf{X}$ and $\mathbf{W}$ denote the phenotype column vector, the SNP input and the learned model weights, respectively. Then, consider the following linear NN which may be derived from the above non-linear NN by switching off all activation functions (σ) as

$$\begin{aligned} Y &= \cancel{\sigma_k}(\dots\cancel{\sigma_2}(\cancel{\sigma_1}(\mathbf{X}\mathbf{W_1})\mathbf{W_2})\dots W_k) && (4.3)\\ &= (((\mathbf{X}\mathbf{W_1})\mathbf{W_2})\dots W_k) && (4.4)\\ &= \mathbf{X}(\mathbf{W_1}\mathbf{W_2}\dots W_k) && (4.5)\\ &= \mathbf{X}W_{all}. && (4.6) \end{aligned}$$

As the two models are identical in every other respect, including the same weights, if (4.6) is at least as accurate as (4.2), then the latter could not have learned interactions between the input features.

After some initial test runs, I found that the performance of the previously described linear NN models was poor due to two different reasons, other than the lack of non-linearity. The weights for the non-linear NN were obtained via training with non-linearity enabled; thus, they may not have been ideal for a model that never had an activation function in the first place. Additionally, my preferred choice for activation (the SELU function), in addition to providing non-linearity, also standardises the output of each layer. This latter property

addresses the internal covariate shift problem I described in the Introduction in section 1.7.4.3. Thus, removing the SELU, and not providing a substitute for its normalization capacity may produce sub-optimal linear models. Therefore, to ensure that I fit the best possible linear NN with an architecture closest to the non-linear version, I made the following two changes. I replaced the SELUs with a batch normalization layer (Ioffe and Szegedy, 2015), and I also retrained the linear NNs with early stopping applied to obtain the best possible weights for these models too.

4.2.5 Inference via neural-networks

In the field of the biomedical sciences point estimates are often inadequate, and an explanation as of why a certain prediction was made, or at what confidence level, are considered highly important. Because of the complexity of the models and the non-linear functions they learn, NNs have traditionally been thought of as 'black boxes' that are unable to satisfy this criterion. However, inference for NN based methods is an emerging area of research; thus, this perception has been slowly changing.

4.2.6 Overview of my NN inference strategy

I will provide a detailed background on each component of my approach for inference in the subsequent sections; however, I will first outline my overall strategy, so that each individual component may be understood in the greater scheme of my analysis. Inference begins by training the NN model to the highest possible prediction accuracy after which the NN is then taken forward to the association stage. Interaction association is then performed for each order of interaction (two, three and four), starting from the lowest, and proceeding upward in a three-step process that I will outline next.

The first step in inference is interaction association, where each method attempts to identify combinations of predictors that were most influential for the NN's performance on the original prediction task. This step is performed via either the NID algorithm (section 4.2.6.3) or my own NNPred algorithm (section 4.2.6.5). This is followed by the second step that consist of significance testing, where the previously identified putative interactions are evaluated against the null hypothesis of no association. For the NID approach this is performed via OLS based techniques, and for the NNPred method this is performed via the application of the dropout technique (section 4.2.6.1). Once all candidate interactions have been identified for a given order, their p-values are FDR corrected, and only those below a 0.05 threshold are taken forward. Finally, the last step consists of a common search-space reduction strategy (section 4.2.6.7), that reduces the number of tests the methods need to

perform based on the heuristic of only considering those interactions deemed to be possible, given the previously discovered associations.

4.2.6.1 Uncertainty estimation via dropout

In image classification tasks where NNs produce a list of predictions to indicate the probability that a given image belongs to a certain class, one may naively expect a prediction of low confidence to have near equal predictions for all categories. However, this does not precisely quantify the uncertainty of the model. For regression based prediction tasks the problem of lack of uncertainty estimation is even more pronounced. As for regression tasks, the model only produces a real value of a single point estimate, which does not reveal anything about how confident the model was when making that particular prediction.

As I described in the Introduction in section 1.7.4.8, the dropout technique is traditionally applied during training only, and is usually switched off for test predictions. However, it was recently proposed that applying dropout at test-time induces the NN to approximate Bayesian inference in deep Gaussian processes (Gal and Ghahramani, 2016). The intuition behind this is that the application of the binary masks that switch off random subsets of neurons may be considered to create an ensemble of NNs; thus, a prediction with dropout enabled is taken as a single observation from their distribution. The implementation of this technique is very straightforward: during test time, dropout is simply not switched off. Thus, for any given test case, instead of just producing a single prediction, potentially many thousands of predictions may be produced, which are taken as the empirical distribution for that prediction. Initially, it was believed this method would only be feasible for FNNs, and not for CNNs, due to the traditionally poor performance of dropout on convolution layers; however, later it was shown that state of the art performance may be achieved if during testing the mean of this distribution is used as the point estimate, and its standard deviation as the standard error of the estimate (Gal and Ghahramani, 2015).

Deploying dropout for uncertainty estimation may be accomplished at no extra computational cost, as the same test observation may be added repeatedly into a mini batch; thus, all observations from the distribution may be obtained via a single forward propagation pass. Whether using dropout in this manner produces a genuine approximation for Bayesian inference is still subject to debate (Osband, 2016); however, this technique has become a popular method for uncertainty estimation in practice.

4.2.6.2 Estimating the importance of input features

Another important objective of inference is to obtain a list of input features (or their combinations) that were most influential in generating the model prediction. This is a challenge, as NN models fit a non-linear model where combinations between input features are considered randomly in a non-exhaustive search to yield the best overall prediction. Additionally, the learned interactions are not stored in individual neuron weights; instead, they are distributed across the network, where each neuron learns only very small fragments of the overall task. This phenomenon is known as distributed representations (Hinton, 1984).

Approaches that permit one to query a trained NN to produce association-like results are reviewed in the following sections. These may be placed into two broad categories, examining the learned NN weights directly, and inference-via-prediction type approaches (there exist a third approach which is to build Visible NNs (Yu et al., 2018); however, as I have not used this for my own work I will not describe this here).

4.2.6.3 Examining the learned weights of the network directly

These approaches obtain inference by attempting to interpret the weights of a trained NN directly. Most relevant to my work from this class of methods is the NID algorithm by Tsang et al. (2017), which was developed to identify statistical interactions between input features. Briefly, the method proposes to learn interactions from the weighs of the first hidden layer's neurons directly. Their algorithm computes I, the interaction strength of a candidate combination, by

$$I = S_i * min(U). \tag{4.7}$$

Here, the strength of the interaction is estimated by weighting the total output influence of the neuron (S_i) by the top most important inputs to that neuron ($min(U)$). The vector S is defined as

$$S = |W^{out}|^T |\mathbf{W^j}|^T |\mathbf{W^{j-1}}|^T \dots |\mathbf{W^2}|^T. \tag{4.8}$$

Thus, S is a matrix product of all the absolute weights from the output back to the second hidden layer. This produces a vector of length a, where a is the number of neurons in the first hidden layer. Intuitively, this quantifies the total influence of the first hidden layer on the output. So in turn, the element i of the vector S isolates the output influence of neuron i. U is a set $\{1,2,\dots d\}$, defined as the top d largest absolute value elements of the vector $|\mathbf{W_i^1}|$ (the column corresponding to the ith neuron in the first layer's weights), where d is the order of interactions considered. The minimum operation is applied to U, as the total strength of an

interaction would be zero, should any individual input feature have zero weight in the first layer.

The algorithm queries each neuron in the first layer, and evidence of association is evaluated based on if a candidate interaction improves the overall model fit on a validation set. A limitation of this method is that the maximum number of interactions it may consider is capped to the number of neurons in the first layer.

4.2.6.4 My NID implementation

The only difference between my implementation and the authors' was that I used regression based models to evaluate the evidence for the strength of association instead of additional NNs. In the original implementation, a NN is fit for each candidate interaction; however, in practice, this would have been infeasible for the scale of my analyses. Also, a NN model of only a few features would not represent a real advantage over a linear model where the interactions are explicitly coded into the design matrix. Thus, my routine for finding a threshold for determining the number of candidate interactions was via forward step-wise regression. I added each candidate interaction, together with their lower-order terms, sequentially into a multiple regression OLS model, and determined the cutoff as the last putative interaction above which the r^2 of the PRS stopped improving on the validation set.

4.2.6.5 Inference-via-prediction

Methods that belong to this class achieve inference via the manipulation of the weights of the model to generate predictions that are informative of feature importance. To explain the rationale behind these techniques, I briefly return to (logistic) regression. There, the slope of the model (β_{OLS}) provides direct interpretability on feature importance. However, because of the non-linearity, the weight matrices of NNs cannot be used directly in the same manner. Instead, as it was shown by Simonyan et al. (2013), the NN analogue of the β_{OLS} is the derivative of the network with respect to the input:

$$\beta_{OLS} \sim \frac{\partial}{\partial \mathbf{W_{in}}}, \tag{4.9}$$

where the above derivative was obtained via the procedure I described in the Introduction by eq 1.41. Intuitively, this quantifies which input predictors would need to be changed the least to affect the final classification the most.

Inference can be obtained on feature importance in two different ways. One approach is called *input-centric* inference, which aims to answer the question: what in a given input contributed most to the final prediction by computing

$$\hat{y}_j = x_j \frac{\partial}{\partial \mathbf{W_{in}}}, \tag{4.10}$$

where the resulting predictions ($\hat{y}_j$), known as heat or saliency maps, are obtained by multiplying a specific input x_j by the network derivative. In the image classification domain, it was found that the clarity of predictions may be improved by adding a small amount of Gaussian noise to x_j, and then taking the average over many predictions to produce the final inference, a technique known as '*SmoothGrad*' (Smilkov et al., 2017).

The other approach is called *network-centric* inference, and this aims to reveal what the NN has learned about a prediction task in a general, input-independent way. This may be implemented by an algorithm known as '*Gradient ascent*', where the derivative of the network is iteratively added to an input over many iterations as

$$x_i = x_{i-1} + \frac{\partial}{\partial \mathbf{W_{in}}_i}. \tag{4.11}$$

In the first iteration, the input (x_0) is initialised with random noise, and the derivative is computed against the desired target class or value for each iteration. One may intuitively understand this process as generating an 'ideal case' for the classifier (this same mechanism was also responsible for the popular imagery behind 'deep dreaming' (Mordvintsev et al., 2015)). To improve on this basic formula, the derivative may be modified by altering the backpropagation algorithm by zeroing out the values of neurons whose derivative was negative. This modification is known as '*Guided Backpropagation*' (Springenberg et al., 2014), and is motivated by the fact that neurons with negative derivatives would only contribute noise to the final classification.

4.2.6.6 My inference-via-prediction implementation

As I have shown in section 4.2.6.2, NN methods that attempt to provide interpretability accomplish it mostly by either dissecting the NN to extract information directly from its weights, such as the NID algorithm, or via the manipulation of the forward/backward propagation process to induce a prediction more informative on the importance of input features, such as the 'guided backpropagation' method.

The problem with the aforementioned approaches is that NNs learn distributed representations, where the learned features are distributed across many neurons (Hinton, 1984). Thus,

focusing in on any particular neuron or neurons, hoping that they may reveal information about the reasons behind a prediction, remains challenging. This is especially true for FNNs, where neurons are not restricted to a subset of the input, as in that case all neurons learn from all input features without restriction. There have been attempts to force neurons to learn 'localist' representations by the application of a special type of regularization. Such regularization forces the NN weights to become orthogonal with respect to other neurons, and thereby 'disentangle' representations (Brock et al., 2016; Rodríguez et al., 2016); however, these techniques have not found widespread use yet. The other issue that render these methods unsuitable for my purposes is that they lack the per-predictor precision that one would expect from traditional statistical inference. To clarify, in images these methods produce a heatmap-like inference, which when overlaid on the original input picture, highlight areas that were relevant for the classification. In the image classification domain, this may be sufficient for visually determining what objects in the image were important (Smilkov et al., 2017; Springenberg et al., 2014). However, to obtain precise inference about individual predictors or their combinations, for example SNPs, such an approach would not have been feasible.

I reasoned that, for lower-order interactions, there may be a much simpler and more powerful solution. My rationale for this was the insight that the only location where a NN is forced to produce a human-interpretable result is the output itself. Thus, instead of trying to force the NN models into something that they were not designed for, I opted for obtaining inference-via-prediction by simply observing the phenotype prediction for inputs that only consisted of the candidate interactions. The implementation of my algorithm (*NNPred*) is presented below:

Algorithm 1 NNPred Interaction search algorithm

input: trained NN classifier S_c

output: list of importance scores IS

procedure INTERACTION-SEARCH(S_c, c)
    $IS \leftarrow 0$ ▷ initialise empty array for importance scores
    for i `in number of` SNP_{set} **do**
        $x \leftarrow 0$ ▷ initialise empty array in the shape of the input data
        $x[SNP_{set}[\mathrm{i}]] = 1$ ▷ set each SNP in tested SNP-set to 1
        $\hat{y} = S_c(x)$ ▷ forward propagate to obtain phenotype prediction
        $IS[i] = \hat{y}$
    return IS

The NNPred algorithm cycles through all possible interactions, and generates a synthetic individual for each, where all input features are zeroed out except the candidate interaction itself. This observation is then forward propagated to produce a phenotype prediction, which is then taken as the evidence for interaction association.

To provide an estimate of uncertainty for each association, I used the dropout method that I described in detail in section 4.2.6.1 (Gal and Ghahramani, 2016). Briefly, this entailed producing a mini-batch number of test predictions with dropout enabled, and taking these as observations from the empirical distribution of the model's prediction. I then used this distribution to obtain p-values for the NN estimate via the usual formulas:

$$\begin{aligned} \beta_{NN} &= \frac{\sum \hat{y}}{L}, \\ \sigma_{NN} &= \sqrt{\frac{\sum (\hat{y}_l - \beta_{NN})^2}{L}}, \\ t &= \beta_{NN} / \sigma_{NN}, \end{aligned} \tag{4.12}$$

where β_{NN} is the NN estimate of the interaction's effect size, and $\hat{y}$ is an individual prediction out of a total of L predictions in a minibatch. σ_{NN} represents the standard error of the estimate, which is then used to obtain t, the quantile of a normal distribution. Finally, t was used to obtain the appropriate p-value.

The central limit theorem states that the sampling distribution of sample means asymptotically tends to a normal distribution. To ensure that this held true for my datasets, I performed the following test in the simulated experiments (described under 4.3). Using a 1,000 predictions of 1,024 observations each, I performed a Kolmogorov-Smirnov test against normal distributions of the same mean and standard deviation. I obtained a median p-value of 0.588, which confirmed that the assumption of normality was appropriate.

4.2.6.7 Common search space reduction strategy

An exhaustive search for higher-order interactions may quickly become computationally infeasible. Evaluating all possible combinations from even just a 1,000 SNPs up to the fourth-order would require over 40 billion tests. Thus, to reduce the search space, I applied a heuristic I described in the Introduction in section 1.1.7. In brief, this entailed only testing Dth order interactions if all nested $D-1$th order interactions were previously found by the algorithm.

An interaction was deemed to exist based on the following criterion. After a complete search for a given order of interaction, the association test p-values were all Benjamini-Hochberg FDR corrected (the number of tests were always in the thousands; thus, the

FDR correction was appropriate). Candidate interactions were considered valid if their $FDR < 0.05$, a stringent threshold which was motivated by the above described assumption, that an interaction may only exist if all its nested interactions also exist. Also, the FDR correction and filtering step was applied to each order of interaction just once, not repeatedly to all interactions found up until that point.

4.2.6.8 OLS baseline

To serve as a frame of reference for the NN based inference approaches, I also evaluated a standard OLS regression based interaction test. This model was almost identical to the one described in the Introduction (eq 1.7). The only change in this method was that, instead of just testing for second-order interactions, I extended the same model up to the fourth-order by adding the appropriate higher-order terms.

4.3 Simulation experiments on synthetic data

To assess if a NN based strategy was capable of analysing data at the scale of the UKBB, and also to have the potential to identify interactions, I performed a set of simulation experiments. The objective of these tests was to serve as a proof of concept if NNs may be used to infer the existence of statistical epistasis, if it was present.

4.3.1 Genotype dataset

I selected the same FIS genotype panel of 955 SNPs that I used for the two-way interaction tests in Chapter 3. I chose this particular panel (as opposed to the larger height or BMI panels) due to the practical considerations of the immense computational resource requirements needed to perform simulations at the scale of the UKBB. The number of individuals used in these experiments were 137,088, 34,270 and 21,775 for the training, validation and test sets, respectively.

4.3.2 Phenotype simulation details

To obtain conclusions from simulations that may offer insight for my subsequent real data analyses, I aimed to obtain simulated phenotypes that arose from a signal comparable in magnitude to the one observed in the real FIS datasets.

To begin, I had to consider the unique properties of simulations that involve epistasis, as here, there are potentially two distinct parameters that control the way causal SNPs contribute

to the phenotype. The first parameter is the causally involved number of SNPs c, which would be the only parameter needed for simulating additive phenotypes. Here however, there may also be a second parameter v, which would control the number of interactions made up from the c SNPs. To determine the causally involved number of SNPs (c) to generate the simulated phenotypes, I evaluated three potential causal fractions of the 955 SNPs: 0.25, 0.5 and 0.95. The upper limit (0.95) for this was motivated by my QC process that involved an $FDR < 0.05$ filter, which implied that ~95% of SNPs were associated with the phenotype. After empirically evaluating three values for v, (c, $c/2$ and $c*2$), I set $v = c$, based on preliminary observations that while v had an impact on the overall accuracy, it did not seem to influence the preference between the linear and non-linear methods. Therefore, to reduce the space for my simulations, I did not pursue experiments that involved alternative choices for v. Thus, the raw genetic values (GV) for individual j were calculated as

$$GV^j = \sum_i^v X_i^j \beta_i \quad ; \quad \beta_i \sim N(0,1), \tag{4.13}$$

where β_i was drawn from a standard normal distribution, and X_i^j denotes the v randomly selected combinations of SNP genotype counts selected to be causal. I simulated four architectures that ranged from the purely additive to second, third and fourth-order interactions. Thus, X_i^j was defined as

$$X_i^j = \prod_d^{D_i} SNP_{dj}, \tag{4.14}$$

where SNP_{dj} is the genotype count for the dth SNP in the ith Dth-order interaction.

Using LDAK (Speed et al., 2012), I estimated the narrow sense SNP heritability of the 955 SNPs to be ~8.1% for the real FIS phenotype. I chose LDAK as opposed to GCTA, as the latter was incapable of working with a kinship matrix of 137,088 individuals due to RAM limitations.

To simulate phenotypes with an additive genetic architecture with a pre-defined h^2, the final phenotype (y_{sim}) is determined to be the sum of the genetic (g) and noise (e) components as

$$y_{sim} = g + e, \tag{4.15}$$

where both g and e are scaled in proportion to the desired h^2. The noise component is drawn from a standard normal distribution, with zero mean and a variance of $1 - h^2$

$$e \sim N(0, \sqrt{1-h^2}). \tag{4.16}$$

Thus, the noise contributes all the remaining variance not due to h^2. The scaled genetic value (g) is in turn defined as

$$g = GV * s, \tag{4.17}$$

where s is a scaling factor given by

$$s = \sqrt{\frac{h^2}{var(GV)}}, \tag{4.18}$$

where $var(GV)$ denotes the sample variance of the genetic values. The above would generate a simulated phenotype, arising from additive genetic effects, with a pre-specified level of h^2 (Speed et al., 2012). However, generating such an additive phenotype was not my objective; instead, I was interested in if the observed additive genetic architecture may have been generated by latent non-linear effects. Thus, the above formula would not have been expected to create a phenotype with a given narrow sense heritability, if in reality the phenotype arose from non-linear effects. To test this, I generated a phenotype as above with the desired 8.1% h^2 for all four interaction levels, and estimated the h^2 once again with LDAK. I found that their actual estimated h^2 was 0.077, 0.05, 0.03 and 0.017 for additive, second, third and fourth-order interactions, respectively. These values indicated that the apparent additive signal was rapidly diminishing at higher-orders interactions. Therefore, I decided to modify the original formula.

I reasoned that the scaling factor (s) needed to be adjusted to be proportionate to the apparent additive effect of SNPs. I attempted to adjust it by fitting a multiple linear regression model, and regressing the individual genetic values (GV) on the genotype matrix as

$$GV = \mathbf{X}'\beta' + \varepsilon, \tag{4.19}$$

where $\mathbf{X}'$, β' and ε denote the genotype matrix of the individual SNPs involved in interactions, the interaction coefficients and a random noise term, respectively. I reasoned that the fitted values from this model ($\widehat{GV}$), would represent the genetic values due to the apparent main effects of the SNPs involved in interactions. Therefore, I proceeded to alter eq 4.18 by using this new $\widehat{GV}$ to produce an adjusted scaling factor by

$$s' = \sqrt{\frac{h^2}{var(\widehat{GV})}}. \tag{4.20}$$

From this point onward, with the exception of using this new scaling factor s', the rest of the simulation steps remained identical to those previously described.

% of sample size	additive	2nd order	3rd order	4th order
10%	0.43	0.42	0.50	0.54
25%	0.27	0.27	0.31	0.55
50%	0.25	0.29	0.41	0.71
75%	0.34	0.25	0.44	0.82
100%	0.20	0.25	0.57	0.92

Table 4.3 **Fractions of experiments where a non-linear solution was found by the NN out 100 simulations with a causal fraction of 0.25 of SNPs involved in statistical epistasis.** The values in the 'additive' column represent experiments where the ground truth genetic architecture was purely additive.

To evaluate the impact of my altered process, I re-estimated the h^2 for the simulated phenotypes generated from the new formula. I found that their estimated h^2 changed to 0.077, 0.074, 0.067 and 0.058 for additive, second, third and fourth-order interactions, respectively. These were still less than the target h^2 of 0.081; however, they were closer than the ones produced by the original naive simulation formula. I note that for the fourth-order scenario the gap between the target and realised h^2 was still ~30%, which I expect would result in a corresponding decrease in power to detect interactions for that scenario. Using this protocol, I simulated a 100 phenotypes for each order of interaction, and for each of the three causal fraction of SNPs. Finally, to evaluate the effect of the sample size on accuracy, I also down sampled the full cohort to 10%, 25%, 50% and 75% of the total available. In total, this produced 6,000 simulated experiments and 300,000 NN models (as each model required 50 hyperopt trials).

4.3.3 Prediction results

The fraction of experiments where a non-linear solution was identified by the model selection procedure are summarised in Tables 4.3 and 4.4 for the causal fractions 0.25 and 0.95, respectively. The relationship between the degree of non-linearity and sample size to the final prediction accuracy of the linear and non-linear NN models are shown in Figs 4.1 and 4.2 for the causal fractions 0.25 and 0.95, respectively (the same information for the causal fraction of 0.5 can be found under A.1 in Appendix A).

It is also important to note that the non-linear results represent the possibility of non-linearity, rather than that a non-linear solution was actually identified in each instance (see Figs 4.1 and 4.2). Thus, it is a more conservative estimate of the benefit of enabling non-linearity, as it counts the experiments into the non-linear result where the best performing model was still in fact linear.

% of sample size	additive	2nd order	3rd order	4th order
10%	0.47	0.50	0.68	0.63
25%	0.25	0.35	0.34	0.50
50%	0.14	0.29	0.28	0.34
75%	0.21	0.25	0.28	0.35
100%	0.20	0.29	0.30	0.36

Table 4.4 **Fractions of experiments where a non-linear solution was found by the NN out 100 simulations with a causal fraction of 0.95 of SNPs involved in statistical epistasis.** The values in the 'additive' column represent experiments where the ground truth genetic architecture was purely additive.

4.3.4 Inference results

As the difference between the linear and non-linear models was most pronounced in the experiments using a causal fraction of 0.25 at the fourth-order, I chose this series for my inference analyses. Fig 4.3 shows the performance of the three evaluated interaction detection algorithms for the fourth-order interaction series of experiments for all the instances where a result was returned by each method.

I chose ROC curves to visualise my results, which are defined by plotting the true positive rate (TPR) against the false positive rate (FPR). TPR and FPR are in turn defined as

$$TPR = TP/(TP+FN)$$
$$FPR = FP/(TN+FP),$$

where TP and FN denote true positives and false negatives, respectively, and FP and TN denote false positives and true negatives, respectively. I also recognised partial interaction matches by defining both the TP and the putative associations to include, in addition to the fourth-order interactions, also all of their unique nested interactions. To clarify, this meant that each fourth-order interaction carried with it four third-order interactions, and each of those in turn carried three second-order interactions. Finally, I removed any overlapping nested interactions, so as to only keep unique sets of SNPs. To illustrate this with a simple example, consider the following three-way interaction: $(1,2,3)$. To account for partial matches, this interaction would also include the following two-way interactions as valid targets: $(1,2)$, $(2,3)$ and $(1,3)$ (but only if none of these were already part of other interactions).

Due to the heuristic employed to avoid exhaustive searches, I also had to manually correct the TN value by adding to it the number of tests not performed by the method. I obtained this number by subtracting from number of interaction tests possible from 955 SNPs, the number

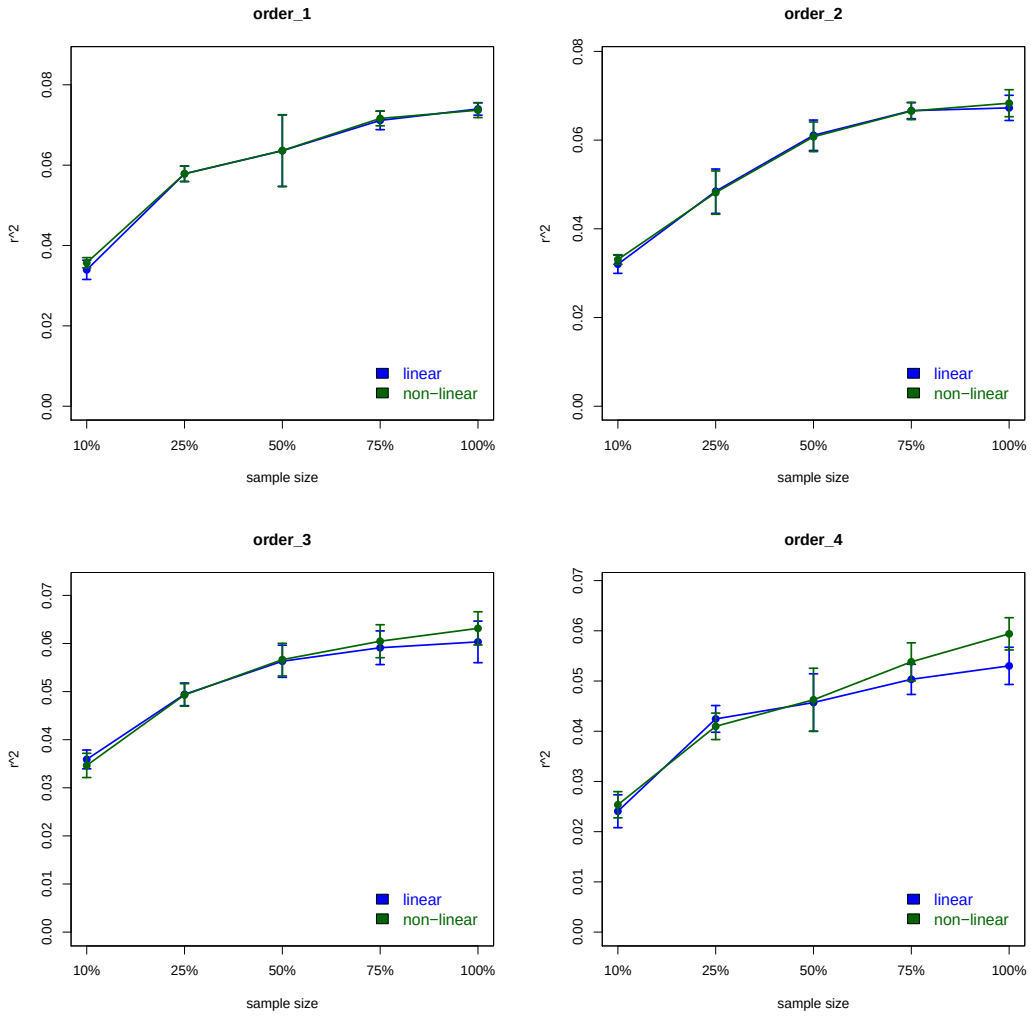


Fig. 4.1 **Neural-network performance on obtaining non-linear solutions under varying conditions for the experiments with a causal fraction of 0.25 of SNPs involved in statistical epistasis.** x-axis represents the % of sample size used and y-axis represents the r^2 of predicted vs observed phenotypes on the test set. Facets display experiments of genetic architectures that involve either additive, second, third and fourth-order interactions.

of tests the methods actually performed. This latter quantity I obtained by enumerating over the total tests performed (including the nested partial matches as I defined above) and adding to it the total number of true interactions.

The results presented in Fig 4.3 evaluate method performance conditioned on the fact that a method actually successfully returned a result. An alternative perspective of the same results is provided by considering all 100 potential experiments regardless if a result was actually obtained, and evaluating each method on that basis. Finally, I also considered evaluating method performance conditioned on the intersection of the experiments where

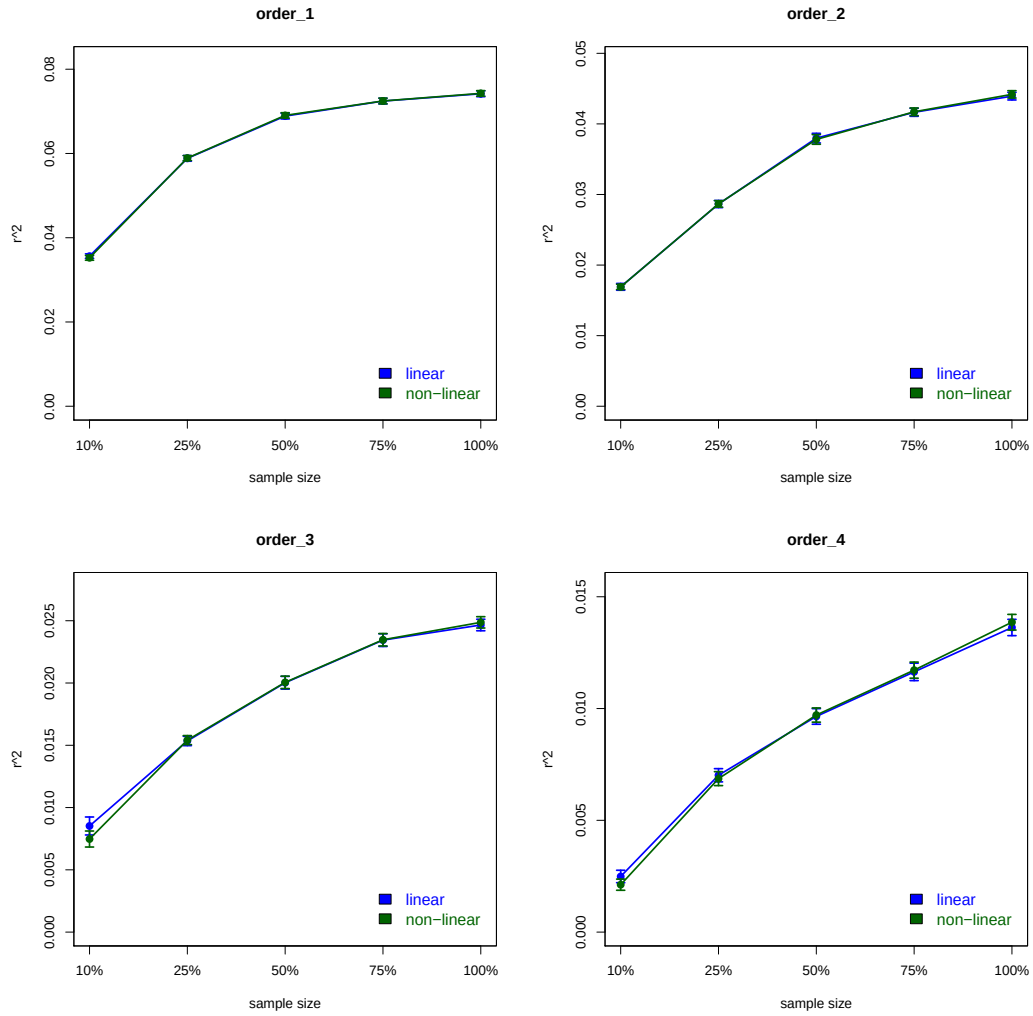
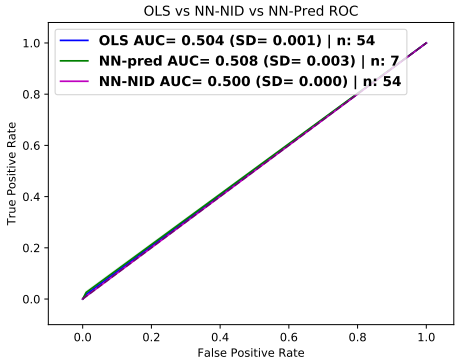


Fig. 4.2 **Neural-network performance on obtaining non-linear solutions under varying conditions for the experiments with a causal fraction of 0.95 of SNPs involved in statistical epistasis.** x-axis represents the % of sample size used and y-axis represents the r^2 of predicted vs observed phenotypes on the test set. Facets display experiments of genetic architectures that involve either additive, second, third and fourth-order interactions.

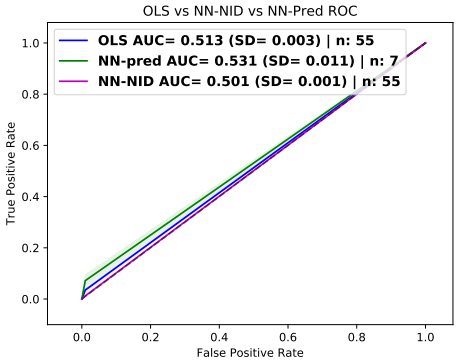
every method returned a result. Table 4.5 summarises method performance from all three perspectives.

4.3.5 Discussion of the simulation experiments

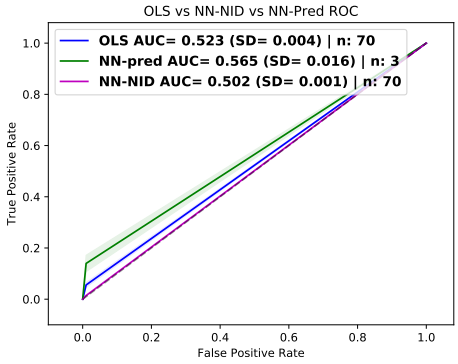
In method development, the goal of simulations is to consider plausible ranges of parameters of data to provide insight under in what scenarios would a novel method offer an advantage over conventional approaches. The level of insight is in turn proportionate to the degree



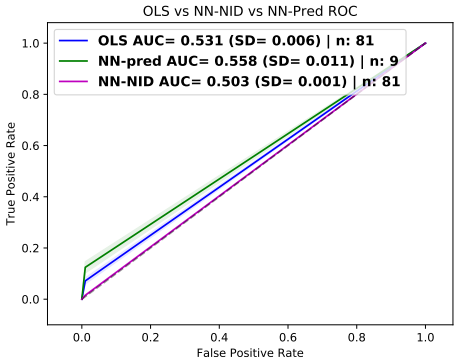
(a) 10% sample size



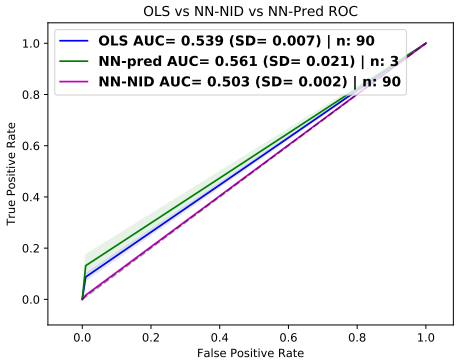
(b) 25% sample size



(c) 50% sample size



(d) 75% sample size



(e) 100% sample size

Fig. 4.3 **The performance of the three evaluated algorithms for statistical epistasis detection for the fourth-order series of experiments for the five sample sizes evaluated.** The average AUCs for OLS, NID and NNPred are shown blue, purple and green, respectively. *n* is the number of experiments from which the curves were drawn from.

that the simulations can approximate real world processes. My own simulation effort was

method	all results		successful results		intersection results	
	AUC (SE)	**n**	**AUC (SE)**	**n**	**AUC (SE)**	**n**
NNPred 10%	0.501 (0.002)	100	**0.508 (0.003)**	7	**0.508 (0.003)**	7
NID 10%	0.500 (0.000)	100	0.500 (0.000)	54	0.501 (0.001)	7
OLS 10%	**0.502 (0.002)**	100	0.504 (0.001)	54	0.505 (0.001)	7
NNPred 25%	0.502 (0.009)	100	**0.531 (0.012)**	7	**0.531 (0.012)**	7
NID 25%	0.501 (0.001)	100	0.501 (0.001)	55	0.501 (0.001)	7
OLS 25%	**0.507 (0.007)**	100	0.513 (0.003)	55	0.513 (0.002)	7
NNPred 50%	0.502 (0.012)	100	**0.565 (0.02)**	3	**0.565 (0.02)**	3
NID 50%	0.501 (0.001)	100	0.502 (0.001)	70	0.502 (0.002)	3
OLS 50%	**0.516 (0.011)**	100	0.523 (0.004)	70	0.529 (0.003)	3
NNPred 75%	0.505 (0.017)	100	**0.558 (0.012)**	9	**0.558 (0.012)**	9
NID 75%	0.502 (0.002)	100	0.503 (0.001)	81	0.503 (0.001)	9
OLS 75%	**0.525 (0.013)**	100	0.531 (0.006)	81	0.528 (0.004)	9
NNPred 100%	0.502 (0.011)	100	**0.561 (0.026)**	3	**0.561 (0.026)**	3
NID 100%	0.503 (0.002)	100	0.503 (0.002)	90	0.505 (0.002)	3
OLS 100%	**0.535 (0.014)**	100	0.539 (0.007)	90	0.532 (0.004)	3

Table 4.5 **Inference results for the simulation experiments for the three methods (NNPred, NID and OLS) at different percentages of the total sample size.** '*AUC*', '*SE*' and '*n*' denote the area under the curve, its standard error and the number of experiments the preceding values were calculated from, respectively. Values under 'all results' represent inference results from all 100 experiments. In case a method did not report a result, its accuracy was substituted by an AUC of 0.5. Values under 'successful results' represent inference results conditioned on individual methods successfully reporting a result. Values under 'intersection results' represent inference results conditioned all three methods reporting a result. Bold text highlights the best method in a given scenario.

therefore limited due to the lack of reliable evidence of what non-linear genetic architectures may comprise of in the real world.

Iterating through all potential factors, such as the number of causal interactions, the proportion of additive to epistatic effects, the ratio between various degrees of non-linear components, including both their number and effect size distribution and their relationship to MAF, would have been intractable. Therefore, my motivation was more modest, I only sought to simulate genetic architectures that covered the extreme scenarios, such as consisting entirely of a given degree of interaction. My aims were limited to illustrate general trends, such as the capacity of NNs to cope with the scale of data, and what general effects do degree of non-linearity and sample size have on expected accuracy at a plausible level of h^2. More specific conclusions, or quantifying relationships between factors (such as the sample size)

and outcomes would have been invalid, as all such relationships would have been influenced by the arbitrary decisions that were used to determine the simulation parameters.

4.3.5.1 Prediction performance

I observe the general trend that the greater the degree of non-linearity and the larger the sample size, the more likely that a non-linear solution was preferred by the model selection process. This trend was present for both causal fraction series of experiments (Tables 4.3 and 4.4). Only the lowest sample size experiments were apparently anomalous, where the model selection chose a non-linear solution at ~50% of the time, even with no genuine non-linear signal. I interpret this as an artifact of low power, as the accuracy of such models was so low to begin with that it made no difference whether a linear or a non-linear model was chosen by the hyperopt optimisation process.

With respect to comparing the series across the two causal fractions (0.25 and 0.95), I observe the following. The lower the number of SNPs involved in the interactions, the better the non-linear models performed in comparison to the linear models, an observation which was true both in terms of model selection, as well as for prediction accuracy (Figs 4.1 and 4.2). This impression was also supported by the 0.5 causal fraction series, which exhibited intermediary results between the two extremes (Figs A.1 and A.1 in Appendix 1). These results indicate that, given a fixed level of h^2, the relationship between the ability to detect non-linearity and the causal fraction is inversely related. My interpretation of this is that a smaller causal fraction requires a larger effect size per interaction to achieve the same level of signal. This observation is also consistent with theoretical results that suggest that the number of loci involved in interactions and the epistatic variance they may each explain are inversely related (Mäki-Tanila and Hill, 2014).

For the series of experiments with a causal fraction of 0.95 of SNPs, I note that despite the positive association between genuine non-linear effects and the choice of a non-linear model, even at the largest sample size and highest degree of non-linearity, a non-linear solution was selected in only 36% of the experiments (Table 4.4). Besides the aforementioned factors, the stochastic nature of the hyperopt search process and the low level of h^2 in the simulated dataset may also have contributed to the lower preference for non-linear models. During my initial exploratory analyses, I simulated a phenotype with a much greater level of h^2 of 0.5, and in that case, the same experiments yielded a non-linear solution over 95% of the time. With respect to the prediction results (Fig 4.2), I find that the linear versus non-linear models were almost identical in performance across all sample sizes and degrees of non-linearity. I performed paired t-tests, and I found that the linear and non-linear means were not significantly different. Even with a 100% of the sample size at the highest degree on

non-linearity, the gain of the non-linear NN over the linear version was not significant (paired t-test p-value = 0.361). In contrast, for the series of experiments with a causal fraction of 0.25 of SNPs (Table 4.3), I observe a greater preference for non-linearity. Here, the corresponding metric is 92% for the highest degree of non-linearity and the largest sample size, which indicates a stronger detectable effect at the stage of model selection. Additionally, I found that in the same experiment (Fig 4.1 lower right), the prediction accuracy of the non-linear model was also significantly higher than the linear version (paired t-test p-value of 0.020).

In summary, my prediction results suggest that, given the range of parameters evaluated in my simulations, NN based models are capable of inferring the presence of statistical epistasis in GWAS SNP data at the scale of UKBB.

4.3.5.2 Inference performance

From the perspective where the methods were evaluated based on the maximum number of successful results for each approach (Fig 4.3 and 'successful results' column in Table 4.5), I note that the ROC curves of all methods have a peculiar shape, there is a short curved rise near (0,0), after which there is a long straight line to (1,1). The former represents the tests performed, and the latter represents the tests not performed due to the search-space reducing heuristic I described in section 4.2.6.7. This shape provides support to the advantages of employing the aforementioned heuristic, as the long straight line suggests that my methods were much more likely to carry out tests for true positive interactions than for true negatives ones.

I observe a monotonous trend that indicates that all methods performed better with each increase in sample size that increased from 0.500 (NID at 10%) to 0.561 (NNPred at 100%). My own NNPred method exhibited the highest AUC at all % of sample sizes that ranged from 0.508 (at 10%) to 0.561 (at 100%).

Relative to the other methods, NNPred had a much lower number of scenarios where it could identify interactions. The method reported only three results for sample size scenarios 50% and 100%, and at most nine for the 75% sample size run. Due to the low number of observations for NNPred at the 50% and 100% sample size scenarios, I only performed significance tests to compare NNPred against NID and the OLS baseline in the remaining scenarios. The advantage of NNPred was significant in all of these tests, except at the comparison against OLS at the 50% sample size scenario (p-value=0.065). At the 75% sample size scenario, where NNPred had the most observations (nine), I found it to be significantly better than both NID and OLS with p-values of $7.025 * 10^{-7}$ and $1.288 * 10^{-4}$, respectively. NNPred consistently outperformed all other methods in all scenarios; however,

it had the disadvantage of being able to identify a solution only at a fraction of the time. Out of the 100 replicates, it only found interactions between three to nine times, whereas the other methods obtained results at a far higher rate, ranging between 54 to 90 times.

The OLS baseline method's performance was intermediate between NNPred and NID. OLS found solutions at the same rate as NID, and it significantly outperformed the latter at every level with p-values that decreased from $1.338 * 10^{-26}$ to $5.577 * 10^{-69}$ for the 10% and 100% sample size scenarios, respectively. I consider OLS to represent a good balance between accuracy and reliability, as although it was not as accurate as NNPred, it proved to be more robust, as it found interactions in the majority of all experiments.

The NID algorithm performed the worst out of the three evaluated methods. NID's AUCs stayed near the chance level of 0.5, only increasing slightly from ~0.500 to 0.503 at the at 10% and 100% sample size scenarios, respectively. Although these AUCs were all very low, I found that they were still all significantly different from the no skill baseline of 0.5 (the largest p-value was $4.373 * 10^{-7}$ at the 10% sample size scenario). One potential reason for NID's low performance may have been that its algorithm assumes that the strength of interactions would be well captured by neurons in the first hidden layer, rather than being more evenly dispersed across the network. However, NNs are known to learn via distributed representations (Hinton, 1984), which implies that the learned features would be distributed across the deeper layers of the network, and therefore less detectable at the hidden first layer.

Considering the alternative perspectives on method performance (Table 4.5) I make the following observations. If I evaluate method performance on all 100 experiments ('all results' column), which may be interpreted as evaluating how well a method does on a random dataset, then the OLS method emerges as the clear winner in all scenarios. NNPred's advantage over the other two methods disappears due to the low number of experiments where it returned a result. NID remained the least performant method in all but the 100% sample size scenario, where it slightly outperformed NNPred (AUCs of 0.502 vs 0.503).

The high performance but an overall low number of results returned by NNPred poses an important question about this method. That is, if it outperformed the other methods only because it returned a result in the subset of experiments where the other methods have done similarly well. This question is answered by the column 'intersection results' in Table 4.5, where results are conditioned on the intersection of experiments where all methods obtained a result. That is, this is a like-for-like comparison, where OLS and NID are evaluated on the same subset of experiments where NNPred obtained a result. I observe that the performance of the other two methods (NID and OLS) did not improve, which suggests that these experiments were a random subset, rather than the 'easy' cases where all methods

would have done well. However, this leads to a further question, which is why NNPred returned a result successfully in these instances but not in the other experiments.

To investigate why NNPred has failed to return any results for such a large number of experiments, I decided to examine if there was a difference in hyperparameter selection between NNs that succeeded or failed to return inference results for NNPred. I found that out of all the hyperparameters (listed in Table 4.2) only the learn rate, dropout and the number of layers were significantly different between failed and successful experiments. NNs that successfully returned an inference result for NNPred had a higher dropout (0.432 vs 0.646, p-value=$8.784 * 10^{-}6$), a lower average number of layers (3.849 vs 1.862, p-value=$1.951 * 10^{-}10$) and a higher learn rate (0.004 vs 0.006, p-value=0.002). The finding of a positive association between dropout and inference performance makes intuitive sense, as switching on and off a larger fraction of the network via dropout may provide a better estimate of uncertainty in predictions. The finding of a positive relationship between inference performance and a higher learn rate or a shallower NN architecture are more difficult to interpret directly. These parameters may have influenced inference performance via a combination with other factors, such as the genetic architecture in a given simulation. An auxiliary explanation for this phenomenon may be found in the reported shortcomings of dropout based uncertainty estimation for NNs, where some researchers argued that this approach does not always provide a genuine approximation of a Gaussian process (Osband, 2016). The overall conclusion I draw from this investigation is that optimising NN architectures for performance on prediction may not always yield architectures that are compatible with inference tasks.

4.4 Neural-network tests on real data

As I was interested in whether NN based approaches may offer a viable alternative to the standard methods I evaluated in Chapter 3 to infer non-linear effects, I assessed their utility on the same datasets. I applied the NN models described in section 4.2 onto the previously prepared cohorts that comprised of the four UKBB traits and the two IBD sub-phenotypes.

4.4.1 Data preparation and model selection

I used the same set of predictors and subsets of individuals that I finished with in Chapter 3. For the SNP datasets I had 955, 1,732, 1,671 and 450 SNPs for FIS, height, BMI and asthma, respectively. For the protein score datasets I used 99, 781, 317 and 38 protein scores for FIS, height, BMI and asthma, respectively. For the asthma phenotype, I also

used the three TWAS tissue gene expression datasets of 264, 218 and 253 gene-scores for monocytes, neutrophils and T-cells, respectively. Finally, I also applied the NN models onto the concatenated cross-domain datasets that integrated SNPs, protein scores and TWAS scores (for asthma), which comprised of 966, 1,805, 1,695 and 579 predictors for FIS, height, BMI and asthma, respectively.

I brought the IBD datasets in line with the UKBB datasets by processing them through the same filtering steps. I only kept the LD-clumped top FDR < 0.05 corrected SNPs from additive associations, and I also filtered out all variants that were within the same recombination block. This process left 308 and 285 SNPs for CD and UC, respectively.

After determining the model architecture via hyperopt on the first bootstrap sample, I trained NN models for all 20 bootstrap samples using the same hyperparameters with early stopping enabled. The number of epochs used for this was $+20$, relative to the ideal number of epochs identified in the first sample. I added this redundancy to allow for slight variations in the ideal number of epochs between bootstrap samples, which was expected, given that each sample is a different observation from the same distribution.

4.4.2 Prediction results on real data

A non-linear solution was preferred for only the following experiments (given in the format of phenotype/domain): BMI/SNP, asthma/SNP, asthma/neutrophils and asthma/cross-domain, together with the two IBD sub-phenotypes/SNP. The results from these are presented in Fig 4.4. Among these, only the asthma cross-domain test was significant with a paired t-test p-value of $1.366 * 10^{-9}$.

4.4.3 Inference results on the asthma cross-domain data

I performed NN based interaction association tests in all 20 bootstrap samples for the asthma cross-domain data. Among the three evaluated methods, only the NID and OLS methods reported putative interactions. To assess why NNPred did not return any inference results for this cohort, I examined the hyperparameters of the model used for this analysis. I found that this model's hyperparameters were closer to those models that did not return an inference result in the simulations (detailed in section 4.3.5.2) than to those which did, with dropout of 0.517, six hidden layers and a learn rate of 0.005.

On average, NID and OLS found 706 and 3,357 associations across all bootstrap samples, respectively, and 1,100 and 749 of associations were present in at least half of the bootstrap samples for NID and OLS, respectively. I reasoned that as the different bootstrap samples are the resampling of the same single original pool of observations, the strongest associations

should manifest through all of them. Thus, to reduce the potential for false positives generated by the stochastic nature of bootstrap sampling, I intersected the 20 association results, and only considered candidate interactions further that manifested across all 20 samples. This left no results for OLS and six candidate interactions for NID.

Genuine associations are expected to show consistent signal across different sources of evidence; therefore, I run the following diagnostic tests to assess the credibility of these six pairs of interactions. I retrieved the association p-values for the six pairs of two-way interaction tests from my earlier interaction tests in Chapter 3. Then, I performed new regression based two-way interaction tests (described in Chapter 3 by eq 3.3) in an attempt to replicate the six pairs of associations in the Test Set. Finally, I also performed cases only tests to obtain an additional source of support. This test (described in the Introduction in section 1.1.7) evaluates the hypothesis that cases that carry the interacting alleles at both loci should be over-represented relative to those that only carry a single copy (Vittinghoff and Bauer, 2006). However, unlike standard SNP based tests where the relationship between allele counts may be evaluated in a contingency table, I was also dealing with continuous values for the gene-based predictors. Evaluating correlation between the continuous predictor and the allele counts within cases may be considered an analogous test; therefore, I performed correlation tests for those pairs. Table 4.6 presents the results from these diagnostics, which includes the NID importance scores, p-values from the original two-way tests, p-values from the interaction tests from the Test Set, and finally, the p-values from the cases only tests.

All standard statistical tests unanimously indicated that none of the six pairs were genuine interactions. The p-values for the interaction terms from both the Training Set and the Test Set, together with the correlation test were all non-significant. Given their non-linearity, NNs are better at capturing structure in the data than linear models, which I thought may explain this apparent discrepancy between their results and of those reported by standard methods. Therefore, I decided to examine the raw data more closely to inspect it for factors that may indicate anomalies, such as outliers or structure.

Table 4.7 summarises the diagnostic statistics that I obtained for these predictors, which included their genomic location and MAF (where applicable). Next, to examine the data more closely, I plotted the genotype counts for cases and controls separately for the three pairs involving only SNPs. Finally, to obtain an analogue of the same information that involved continuous predictors, I created a scatter plot for the two gene-scores (ENSG00000238818 vs SPARC) and box-plots the SNP/gene-score pairs (rs16858573 vs GLMN and rs2970932 vs ENSG00000232528). Visual examination of the data (Fig 4.5 and Table 4.8), did not reveal any anomalies, such as outliers or groupings, that may have explained the discrepancy

predictors	$\mathbf{NN_{IS}}$	$\mathbf{p_{train}}$	$\mathbf{p_{test}}$	$\mathbf{p^{cases}_{testCorr}}$
ENSG00000238818, SPARC	0.169	0.792	0.261	0.096
GLMN, rs16858573	0.167	0.412	0.580	0.770
rs903361, rs2112535	0.172	0.164	0.572	0.636
rs2970932, ENSG00000232528	0.185	0.856	0.823	0.828
rs3813308, rs2830962	0.177	0.385	0.315	0.184
rs2492419, rs2832662	0.173	0.392	0.345	0.435

Table 4.6 **Comparison between the significance metrics of the NN and standard statistical methods for the variants identified as potentially interacting.** $\mathbf{NN_{IS}}$ is the importance score produced by the NID algorithm (arbitrary scale). $\mathbf{p_{train}}$ is the raw interaction p-value from Chapter 3 that considered all predictors which survived the filtering process in the Training Set. $\mathbf{p_{test}}$ is the raw interaction p-value for the same pairs in the Test set. $\mathbf{p^{cases}_{testCorr}}$ is the p-value of the correlation between the predictors in the cases only test in the Test Set.

between the NN and the standard methods. All plots appeared to support the standard formal test results that indicated no difference between cases and controls for the pairs investigated.

4.4.4 Summary and limitations

The results from the IBD datasets resemble the results from the small sample size simulation experiments (Fig 4.4). That is, a non-linear solution may have been preferred in situations where the NN had a very low power. A non-linear solution may have been chosen due to chance, or potentially due to the overfitting on the validation set with the help of non-linearity, which then ultimately fell short on the predictions for the Test set. Indeed, I found this to be the case for both CD and UC, as the non-linear models were significantly worse than their linear versions, indicated by their t-test p-values of $1.071 * 10^{-12}$ and $2.314 * 10^{-6}$ for CD and UC, respectively. This finding underlines why NNs require large sample sizes, and that any positive results that originate from smaller cohorts have to be treated with caution.

In the UKBB, only the asthma cross-domain experiment showed a genuine non-linear effect (paired t-test p-value $1.366 * 10^{-9}$). One possible explanation for this may have been that interactions exist across SNP and gene-level predictors. To see if I could localise this non-linear advantage, I performed NN based association tests which identified six putative pairs that were particularly relevant for this improved non-linear prediction. All of the gene-level variants originated from the TWAS panels, two from monocytes (ENSG00000232528, SPARC) and two from neutrophils (FOXK1 and ENSG00000238818). Among the genes, SPARC was the only one with an established link to asthma (Wong and Sukkar, 2017).

predictor	**chr**	**bp**	**MAF_{case}**	**$MAF_{control}$**
ENSG00000238818.1	1	15237862	-	-
SPARC	5	150043046	-	-
GLMN	1	91712200	-	-
rs16858573	2	143875725	0.119 (0.003)	0.125 ($9 * 10^{-4}$)
rs903361	1	203091274	0.327 (0.007)	0.338 (0.001)
rs2112535	5	176531075	0.252 (0.003)	0.245 (0.001)
rs2970932	2	162858200	0.415 (0.004)	0.423 (0.001)
ENSG00000232528.3	20	748610	-	-
rs3813308	5	118690781	0.444 (0.004)	0.433 (0.001)
rs2830962	21	28844948	0.172 (0.003)	0.166 (0.001)
rs2492419	6	83407023	0.42 (0.004)	0.428 (0.001)
rs2832662	21	31596876	0.012 ($8 * 10^{-4}$)	0.01 ($3 * 10^{-4}$)

Table 4.7 **Diagnostic statistics for each variant potentially involved in interactions.** The values in parentheses in the 'MAF' columns are the standard error of the mean.

Among the SNPs, three have been previously associated with asthma or respiratory diseases in the GWAS catalogue (rs16858573, rs903361 and rs3813308).

I attempted to find support for the putative interactions for the asthma phenotype by comparing these results to those obtained for the same pairs in Chapter 3. I also performed additional tests on the Test Set to obtain an orthogonal source of evidence. None of these putative associations found any support from the standard statistical methods. All standard statistical tests yielded results consistent with the null hypothesis.

There are several additional reservations that further reduce the credibility of these alleged interactions. The asthma phenotype was the one UKBB trait where I have not used a different chip for the Test Set due to its prior links with asthma, and had to evaluate it on an independent subset of the main UK Biobank Axiom™ chip which may have given rise to a non-linear batch effect. Additionally, given that asthma is a binary phenotype, putative interactions are susceptible to the thresholding artefact I described in the Introduction in section 1.1.6.3. Finally, the NID search method is an experimental algorithm with no proven empirical track record. From my own experience in the simulation experiments it actually proved to be consistently the worst among all evaluated methods (section 4.3.5). Thus, NID may have made false positive associations due to low power, and the algorithm's unrealistic assumptions that interactions would be well captured by the first hidden layer's weights, rather than be more widely distributed across the network (Hinton, 1984).

In conclusion, I deemed that there is insufficient evidence to consider any of these six interactions to be real. Given how easy it is to fit a convincing biological narrative onto false

cases

rs903361 \ rs2112535	0	1	2
0	0.256	0.171	0.028
1	0.242	0.165	0.028
2	0.060	0.041	0.008

controls

rs903361 \ rs2112535	0	1	2
0	0.250	0.162	0.026
1	0.255	0.165	0.027
2	0.065	0.043	0.007

cases

rs3813308 \ rs2830962	0	1	2
0	0.256	0.171	0.028
1	0.242	0.165	0.028
2	0.060	0.041	0.008

controls

rs3813308 \ rs2830962	0	1	2
0	0.250	0.162	0.026
1	0.255	0.165	0.027
2	0.065	0.043	0.007

cases

rs2492419 \ rs2832662	0	1	2
0	0.256	0.171	0.028
1	0.242	0.165	0.028
2	0.060	0.041	0.008

controls

rs2492419 \ rs2832662	0	1	2
0	0.250	0.162	0.026
1	0.255	0.165	0.027
2	0.065	0.043	0.007

Table 4.8 **Training Set genotype fraction tables for the asthma phenotype for cases and controls for the three putative two-way interactions that involved SNP pairs.**

positive associations (Biedrzycki et al., 2019), I refrain from further speculation about how these interactions could have contributed to the pathogenesis of asthma.

4.4.5 The outlook of NNs for building PRS

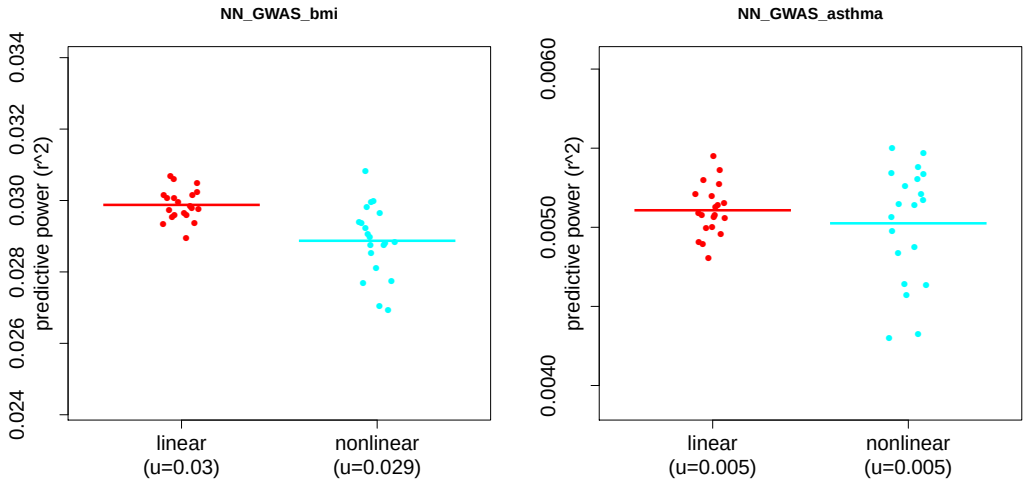
My project of using NNs to build more accurate PRS offered numerous improvements over the comparable efforts on humans by Bellot et al. (2018) and Xu et al. (2020). My contribution included a more rigorous treatment to eliminate haplotype effects, an explicit test if NNs learned genuine non-linear genetic effects, the novelty of considering interactions across domains, and finally, the application of NN based inference methods to identify individual interactions. However, despite these advances, I found that my results are consistent with the aforementioned studies, both of which found no consistent or substantial contributions from non-linear effects to phenotypic variance. Our results are in stark contrast with agricultural applications, where similar efforts have been much more successful (Ma et al., 2017; Pook

et al., 2020). I speculate that this may be due to the fact that the plant and animal subjects of those experiments were not natural populations, but rather products of recent, artificial breeding programmes. Such recent crosses between distantly related populations (or inbred lines) may 'convert' functional epistasis into statistical epistasis by the creation of novel heterozygotes at loci previously only involved in functional epistasis (Mackay, 2014).

FNNs face numerous additional practical limitations due to the fact that they require genotype level data to build PRS. My simulations suggested that the sample size required for NNs to reliably obtain a non-linear solution is likely to be far beyond non-biobank-scale cohorts. Although biobanks are increasing in size and popularity (GEL, 2020; The All of Us Research Program Investigators, 2019), many traits still rely on the pooling of smaller cohorts in meta-analyses. This presents an additional problem for NNs, as building NN derived PRS directly from SNP data requires the integration of cohorts on the genotype level. This is incompatible with the normal practice of meta-analyses, where QC and the association step are performed within each cohort, and results are only integrated via summary statistics. The only way NNs could work with such data would be if QC would be performed on the basis of the 'lowest common denominator' (removing all variants that failed in any of the datasets), and the cohort batch effects regressed out from the phenotypes. This would result in a loss of many variants, and potentially in the imperfect controlling of batch effects. Finally, even if such technical integration was feasible, genotype level access may not be granted on the basis of legal or privacy considerations.

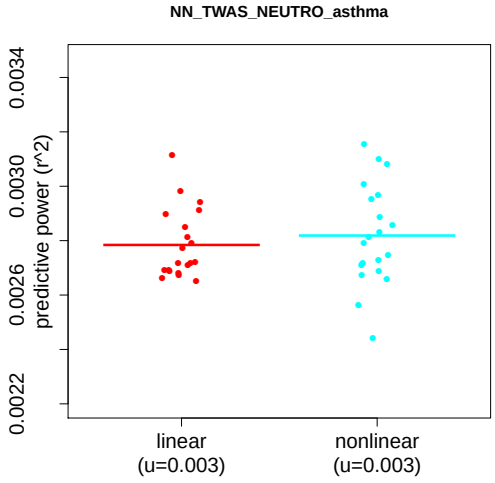
There are also computational challenges that would have to be addressed before NNs could become a viable option for phenotype prediction. Even if statistical interactions did contribute to phenotypic variance substantially, most of the phenotypic variance would still be due to additive effects. Most state-of-the-art PRS consists of ~500K or more SNPs (Khera et al., 2018), a dimensionality that would seem insurmountable for fully connected NNs trained on GPUs and sample sizes of the current generation.

In the future, if single cohort biobanks on the scale of the millions become available with perfect coverage to eliminate haplotype effects, together with GPUs powerful enough to fit the number of SNPs comparable to that of the state-of-the-art additive PRS, experiments similar to mine may be repeated with the hope of a more positive verdict. For the time being however, I see limited real-world applications for NNs to build PRS directly from GWAS SNP data.

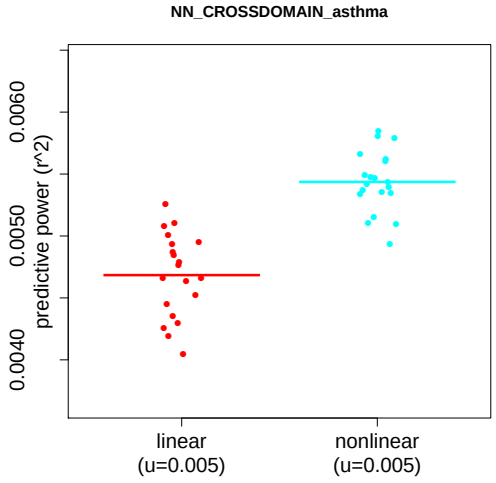


(a) BMI - SNP

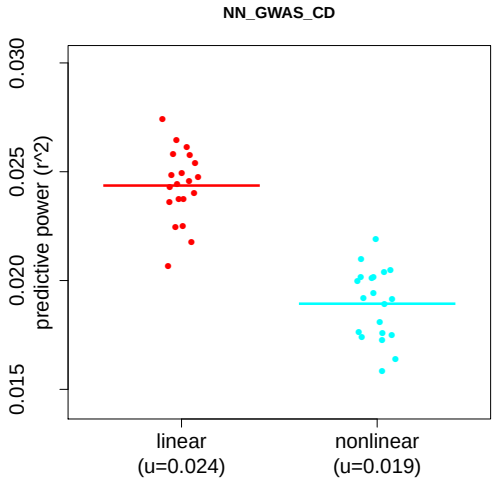
(b) Asthma - SNP



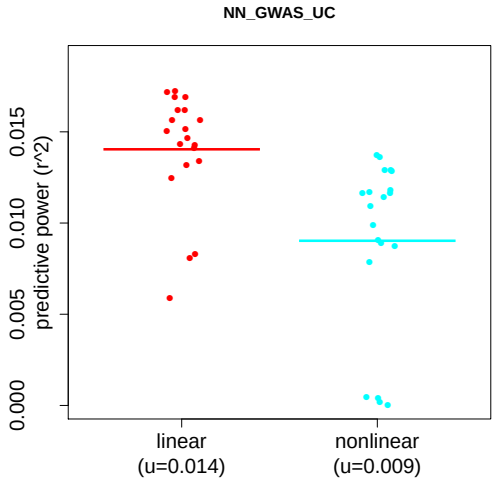
(c) Asthma - TWAS neutrophils



(d) Asthma - cross-domain

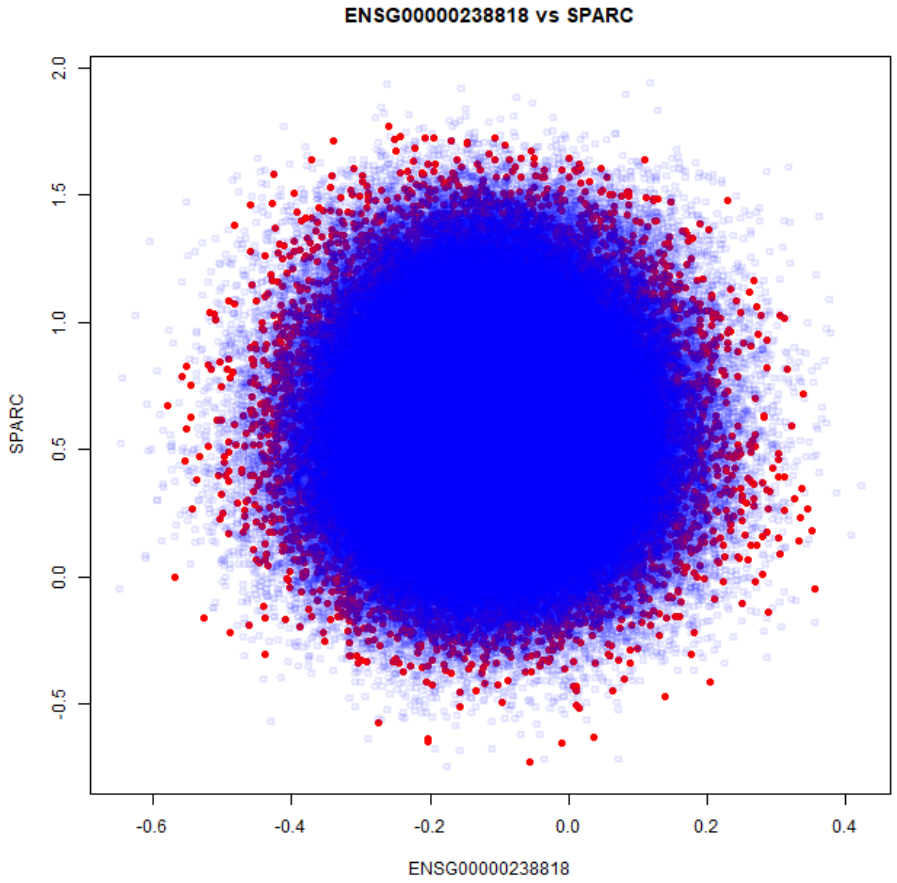


(e) CD - SNP

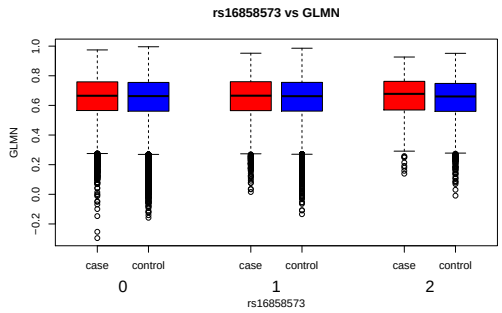


(f) UC - SNP

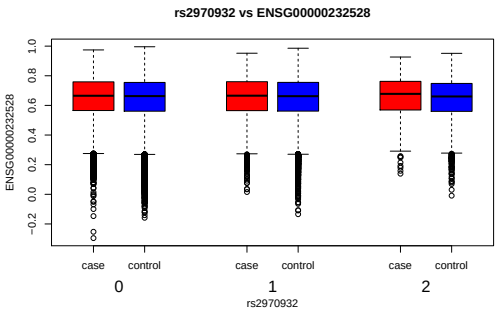
Fig. 4.4 **NN performance in the six experiments where a non-linear solution was preferred.** Results given in the format of *phenotype - domain*. y-axis represents r^2 of predicted vs observed phenotypes on the Test Set. For CD and UC the Test Sets were GWAS1 and GWAS2, respectively.



(a) ENSG00000238818 vs SPARC



(b) rs16858573 vs GLMN



(c) rs2970932 vs ENSG00000232528

Fig. 4.5 **Diagnostic plots for the asthma cross-domain experiments for putative interaction pairs involving gene-level predictors.** Red and blue represent cases and controls, respectively.

Chapter 5

Conclusion

5.1 Overview and limitations

The main motivation for the work in this thesis was to find evidence for how or if non-linear encoding of genetic information contributes to phenotypic variance. After applying quality control measures and establishing the additive association baselines in Chapter 2, I searched for statistical epistasis using standard statistical methods and NN approaches in Chapter 3 and Chapter 4, respectively. Like the two parallel efforts that were comparable in scope to my work (Bellot et al., 2018; Xu et al., 2020), neither the standard nor the NN approaches produced evidence for contributions of statistical epistasis to phenotypic variance.

I believe that the greatest practical limitation of my work was that I restricted myself to only perform recombination block level tests, which precluded the possibility of detecting interactions within blocks. I thought that this was necessary due to the potential for a perfect overlap between genuine statistical epistasis and haplotype effects to exist (Wood et al., 2014). Such haplotype effects are a physical property of the DNA molecule, whereas I was interested in non-linear effects that describe information encoding. However, taking this highly conservative approach meant that I no longer had the ability to identify (the biologically potentially more plausible) local interactions, and this may have contributed to the overall negative results of my analyses. In the future, once WGS data becomes standard, large-scale fine mapping databases (such as the CausalDB (Wang et al., 2019)) and methods that could handle multiple causal signals (Wang et al., 2020) become more widely used, interaction tests that involve fine mapped causal loci may be performed without the danger to be mistaken for pure haplotype effects.

5.2 Reflections on non-linear genetic effects

Current estimates of the fraction of the human genome that is truly functional range from 8.2% to 80% (Dunham et al., 2012; Rands et al., 2014). All functional areas of the genome would be expected to work in concert to produce a genome-wide phenotype, which would then arise as an emergent property from the activity of all active parts of the base sequence. From this perspective, non-linearity appears to be an inevitable property that arises from the compression of information to produce complex biological systems.

When I began my work, I started with the very sensible, although now I what believe to be erroneous, intuition that to achieve the aforementioned non-linear encoding of genetic information, the process that generates or maintains trait variance should also be non-linear. There were numerous supporters of this view who made plausible arguments based on either simulations (Carter et al., 2005) and theoretical grounds (Mackay and Moore, 2014) or by citing examples from model organisms (Hansen, 2013). However, after reflecting on my findings, and revisiting some of the same literature that shaped my initial views, I have now come to believe that my initial views were misguided.

After working with and developing methods for real biological data for several years, I find simulations and theoretical arguments less convincing than before, as biology is a science of what is, rather than what could be. Arguments based on how the apparent additive profile of traits could also be explained by alternatively parameterised models that involve epistasis, such as those made by Huang and Mackay (2016) that I covered in the Introduction under section 1.1.5.1, now leave me unimpressed, as these theories cannot be proved or disproved by current statistical frameworks or datasets. State-of-the-art evidence from a very recent study by Hivert et al. (2020) that performed large-scale variance component analysis in the UKBB across 70 complex traits, found no significant contribution to phenotypic variance from epistatic effects.

I now believe that those who think that evidence from model organisms or artificial populations imply that statistical epistasis may be relevant to humans may have made the same conceptual mistake I did. This thinking ultimately stems from conflating functional with statistical epistasis or alternatively phrased, the trait with the variance in a trait. To clarify, I will illustrate this with a quote from Mackay and Moore (2014), where the authors stated that "*quantitative variation in phenotypes and disease risk must result in part from the perturbation of highly dynamic, interconnected and non-linear networks [...] by multiple genetic variants; thus, gene-gene interactions are likely."* Here, they reasoned that because the trait itself is a product of complex non-linear systems, trait variation would also require non-linear effects, whereas these two do not necessarily imply each other. For another illustrative example of this thinking, see the study by Kuzmin et al. (2018). Here, after finding that

artificially introduced variation generated abundant higher-order interactions in yeast, the authors suggested that epistasis may explain missing heritability in humans. However, I now posit that this non-naturally segregating variation that they introduced via mutagenesis merely exposed latent functional epistasis by generating artificial heterozygosity at normally non-polymorphic loci. This in turn caused the disruption of the normal functionality of the genome (such artificial variation is almost always deleterious), rather than provided insights on normal trait variation. While such deleterious mutations may also occur in natural populations, as they are deleterious, they would not persist or be present in great enough numbers to affect population variance. Chance (rare) mutations, which selection have not had time to eliminate yet, in the constrained part of the genome could also conceivably manifest as statistical epistasis in humans; however, these would be either transient, or make up a very small fraction of the total trait variance. On the other hand, (missing) heritability is a property of normal trait variance due to naturally segregating variation in the base sequence. Considering this alternative explanation, I now do not believe that the evidence presented by the authors supports the conclusion that statistical interactions may be relevant to humans.

Finally, as for the few demonstrated cases of epistasis in humans (for example in rheumatoid arthritis (Dang et al., 2016; Génin et al., 2013; The Australo-Anglo-American Spondyloarthritis Consortium (TASC) et al., 2011)), these only support my current view that such effects are scarce, and in the larger landscape of phenotypic variation they account for little relative to additive associations. Indeed, the GWAS Catalog recorded an exponentially increasing number of additive associations in the last 10 years (Buniello et al., 2019), whereas there has been no comparable progress in epistasis detection.

The apparent lack of a direct contribution of non-linear genetic effects that would impact trait variance may appear puzzling at first, especially given the almost unimaginably complex encoding of information in the genome. However, this paradox may be resolved by Fisher's original explanation for this phenomena that I first described in the Introduction in section 1.1.5.2. Fisher proposed that non-linear information encoding may be achieved by one change at a time through purely additive processes. Under this model the probability of a new allele's frequency rising or falling is conditioned on the (potentially) fixed parts of the genome. Fisher remarked on this topic in his book The Genetical Theory of Natural Selection, where he wrote:

'[...] the effects by which any gene-substitution is recognized depend on the results of interactions with, possibly, all other ingredients of the germ plasm [...]' (p52, 2nd ed.).

In conclusion, perhaps the strongest argument against the importance of epistasis to trait variance is that it is simply not necessary. Given that nature tends to prefer parsimonious

solutions where possible, if non-linear information encoding can occur from purely additive processes, then there is no need to introduce or even to maintain non-linear population genetic variance.

5.3 Outlook and future work

Even if non-linearity does not (substantially) directly contribute to phenotypic variance in a population, the way information is stored is still a crucial attribute of the genome, and decoding it will be essential to deepen our understanding of how genetic variation impacts complex traits and disease risk. Thus, with the benefit of hindsight, I feel that I spent my time looking for non-linearity in the wrong place, between polymorphic loci, rather than where it resides in abundance, in the rest of the genome. Therefore, my research interests now turn towards considering the non-linearity within fixed areas of the genome as a prior, and finding ways to connect that back to phenotypic variance.

Relating fixed parts of the genome to phenotypic variance may seem like an impossibility at first, as loci which do not vary in the population, by definition, cannot contribute to trait variance; thus, their function may appear inscrutable. However, sequence analysis is about relating the different parts of the genome against each other, loci which do not vary in the population still vary with respect to other regions of the genome; thus, invariant sequence context may be used to infer the effects of polymorphic loci. Therefore, examining the (local) sequence context of causal loci may reveal information about what makes, say, a height SNP a height SNP. If this information can be learned, then this may be used to predict a prior for polymorphic loci elsewhere; thus, accomplishing the goal to relate non-linear genetic effects in the fixed parts of the genome to phenotypic variance. Considering the wider field of how the NN framework is applied in genomics, I see a trend converging towards this goal.

The prevailing trend in most successful NN projects so far was to link narrow molecular phenotypes, such as TF binding, to local sequence context of ~1000bps. The scope of more recent sequence analyses have been gradually expanded to encompass larger and larger areas of the genome, which grew from 1000bp to ~131Kb, to consider more distal regulatory features (Kelley et al., 2018), even up to ~1Mb to study genome folding (Fudenberg et al., 2019). The complexity of the target phenotypes have also been increasing. Early efforts aimed to predict basic molecular phenomena, such as the presence of regulatory features; however, more recent studies have realised more ambitious goals, such as relating sequence features to gene expression via the integration of many smaller models over 40Kbs (Zhou et al., 2018). However, none have yet managed to explicitly tie non-linear genetic effects directly to genome-wide phenotypes, such as complex traits and diseases. Thus, I expect

that connecting non-linearity in sequence data to phenotypic variance may be the next major challenge to be overcome in the coming years.

At this junction, it is also necessary to re-examine the ceiling of the maximum level of functional inference possible from NN based sequence analysis. As I covered in the Introduction in section 1.7, NNs perform best under a large data regime, where the outcome depends on non-linear combinations of the input features. To explore this argument further, I need to introduce a new concept which I will refer to as the 'self-containedness' of the problem being modeled. To clarify, this concept describes the observation that the class label of an image only depends on the pixels in the image, or that for games like GO (Silver et al., 2016) all the relevant information is included on the game board. For these types of prediction tasks a NN based model may achieve near perfect accuracy in prediction, as all the elements that contribute to the outcome are present in the training data. However, some may argue that biology is different, as biological systems potentially depend on input from external sources. Inference in this context would be equivalent to training from and then predicting trait SNP coefficients, such as those obtained from a GWAS. The model from which the SNP coefficients are obtained from include (covariates and) a noise term; thus in expectation, a SNP coefficient is the pure genetic effect driven by the base sequence alone, and is therefore predictable from the base sequence. Of course, the more complex the trait, the wider the context that would need to be considered; however, as the ultimate source of causality is still the base sequence, predicting SNP coefficients should also remain possible in theory. The overall trait inference possible from the sequence alone is quantified by heritability, which also represents a direct measurement of this aforementioned 'self-containedness' of the system. Recent heritability analyses revealed that for a great many traits the nucleotide base sequence is the ultimate origin of causality for the majority of trait variance (Polderman et al., 2015); thus in theory, the limits of trait inference from pure sequence data are also correspondingly high. From this perspective, the phenotype may be viewed as a non-linear transformation of the base sequence, up to level of broad-sense heritability. This view also implies that all intermediate stages such as cell, tissue and organ differentiation, expression levels, micro-biome (or at least the parts of these systems that are relevant to the traits), are also in turn determined by the base sequence. Thus, at least in theory, there would have to exist a direct non-linear map from the base sequence to the genetic component of the phenotype which does not depend on any further information from biological samples or environmental covariates. Given certain parallel developments in genomic studies described below, this suggests that modelling non-linearity may become increasingly important in the future.

The size of GWAS cohorts have been steadily increasing. Back in the 2000s studies typically numbered in the low thousands of individuals, whereas today meta-analyses have reached the ~1 million watermark (Lee et al., 2018). This trend is going to continue in the future, with biobank scale efforts in the UK alone set to reach ~5 million individuals with the *5 million genomes project* in 2023 (GEL, 2020). With other countries following suit, it seems highly likely that within a decade meta-analyses will reach cohort sizes on the order of tens of millions. This brings me to one of the rarely appreciated advantages of the GWAS design, which is the way its resource costs scale relative to studies that rely on more intrusive biological samples (such as specific cells or tissues). The biological data required for a GWAS is minimal, a saliva sample is sufficient, which may be collected during routine visits to one's GP. As electronic health records are becoming common (which may serve the the target phenotypes), GWAS may be considered as a mostly information based study that lends itself to large-scale automation, which could encompass entire populations in the near future.

Let us now consider studies that require more involved biological samples, such as biopsies of tissue samples, single-cell sequencing or microbiome data etc. These types of studies scale linearly with the number of sample donors, as they rely on manual and often labour intensive sample collection procedures. Also, the ceiling for cohort sizes would be limited to the fraction of the population willing to undergo such invasive procedures. Thus, it may be reasonably expected that while the costs of GWAS-like studies scale less than linearly with sample size, so these will likely to reach tens or even hundreds of millions of individuals, studies that require biological samples will grow in size at a far lower rate. I believe that this difference in scaling up may also mean that the relative importance of GWAS-like studies will grow disproportionately in the long term. This likely increase in the importance and size of GWAS type studies may also create more opportunities for methods that could provide mechanistic insights into the function of the genome based primarily on sequence information. Much of statistical genetics today is about recovering a faint signal from a very noisy source, whereas NNs excel in the task of modelling a highly complex non-linear signal when sample size is no longer a limiting factor. As the volume and the resolution of available genomic data increases, the field of genetics may start to resemble more closely the domains where NNs traditionally excel; hence, I expect the areas where NNs are applied in genomics to increase in the near future.

The current paradigm to infer mechanism relies on empirical evidence from the experimental manipulation of the genome that may yield insights on functionality, which could then be used to identify drug targets for example. While traditional GWAS is limited to identify one-to-one associations between individual genetic variants and phenotypic outcomes, the previously described trends, which will result in an orders of magnitude increase in genetic

information, may open up opportunities for alternative approaches that could offer more mechanistic insights via advanced computational approaches. Methods such as NNs and their algorithmic descendants, which employ non-linear modelling of genetic effects, are uniquely suited to extract functional insights purely from genetic information by the virtue of the learned non-linearity. To illustrate why this is the case with a general example, consider a (fine mapped) GWAS SNP. Because of the additive nature of the GWAS association, it cannot reveal anything about its genomic context by itself. However, if associations would be made via implicating SNPs together with their relevant sequence context (that may include potentially non-polymorphic regulatory targets), then each association could also become biologically informative. Therefore, despite my own negative results in this work, I am cautiously optimistic about the future applicability of non-linear methods to genetic data, and I see a potential future where large biobank-scale GWAS and NNs are applied in complementary roles, as the former would generate the data and additive associations, and the latter could provide inference on mechanism of effect.

References

Gad Abraham, Yixuan Qiu, and Michael Inouye. FlashPCA2: principal component analysis of Biobank-scale genotype datasets. *Bioinformatics*, 33(17):2776–2778, September 2017. ISSN 1367-4803, 1460-2059. doi: 10.1093/bioinformatics/btx299. URL https://academic.oup.com/bioinformatics/article/33/17/2776/3798630.

Christopher I. Amos, Joe Dennis, Zhaoming Wang, Jinyoung Byun, Fredrick R. Schumacher, Simon A. Gayther, Graham Casey, David J. Hunter, Thomas A. Sellers, Stephen B. Gruber, Alison M. Dunning, Kyriaki Michailidou, Laura Fachal, Kimberly Doheny, Amanda B. Spurdle, Yafang Li, Xiangjun Xiao, Jane Romm, Elizabeth Pugh, Gerhard A. Coetzee, Dennis J. Hazelett, Stig E. Bojesen, Charlisse Caga-Anan, Christopher A. Haiman, Ahsan Kamal, Craig Luccarini, Daniel Tessier, Daniel Vincent, François Bacot, David J. Van Den Berg, Stefanie Nelson, Stephen Demetriades, David E. Goldgar, Fergus J. Couch, Judith L. Forman, Graham G. Giles, David V. Conti, Heike Bickeböller, Angela Risch, Melanie Waldenberger, Irene Brüske-Hohlfeld, Belynda D. Hicks, Hua Ling, Lesley McGuffog, Andrew Lee, Karoline Kuchenbaecker, Penny Soucy, Judith Manz, Julie M. Cunningham, Katja Butterbach, Zsofia Kote-Jarai, Peter Kraft, Liesel FitzGerald, Sara Lindström, Marcia Adams, James D. McKay, Catherine M. Phelan, Sara Benlloch, Linda E. Kelemen, Paul Brennan, Marjorie Riggan, Tracy A. O'Mara, Hongbing Shen, Yongyong Shi, Deborah J. Thompson, Marc T. Goodman, Sune F. Nielsen, Andrew Berchuck, Sylvie Laboissiere, Stephanie L. Schmit, Tameka Shelford, Christopher K. Edlund, Jack A. Taylor, John K. Field, Sue K. Park, Kenneth Offit, Mads Thomassen, Rita Schmutzler, Laura Ottini, Rayjean J. Hung, Jonathan Marchini, Ali Amin Al Olama, Ulrike Peters, Rosalind A. Eeles, Michael F. Seldin, Elizabeth Gillanders, Daniela Seminara, Antonis C. Antoniou, Paul D.P. Pharoah, Georgia Chenevix-Trench, Stephen J. Chanock, Jacques Simard, and Douglas F. Easton. The OncoArray Consortium: A Network for Understanding the Genetic Architecture of Common Cancers. *Cancer Epidemiology Biomarkers & Prevention*, 26(1):126–135, January 2017. ISSN 1055-9965, 1538-7755. doi: 10.1158/1055-9965.EPI-16-0106. URL http://cebp.aacrjournals.org/lookup/doi/10.1158/1055-9965.EPI-16-0106.

Carl A Anderson, Fredrik H Pettersson, Geraldine M Clarke, Lon R Cardon, Andrew P Morris, and Krina T Zondervan. Data quality control in genetic case-control association studies. *Nature Protocols*, 5(9):1564–1573, September 2010. ISSN 1754-2189, 1750-2799. doi: 10.1038/nprot.2010.116. URL http://www.nature.com/doifinder/10.1038/nprot.2010.116.

Christof Angermueller, Tanel Pärnamaa, Leopold Parts, and Oliver Stegle. Deep learning for computational biology. *Molecular Systems Biology*, 12(7):878, July 2016. ISSN

1744-4292, 1744-4292, 1744-4292. doi: 10.15252/msb.20156651. URL http://msb.embopress.org/lookup/doi/10.15252/msb.20156651.

William J. Astle, Heather Elding, Tao Jiang, Dave Allen, Dace Ruklisa, Alice L. Mann, Daniel Mead, Heleen Bouman, Fernando Riveros-Mckay, Myrto A. Kostadima, John J. Lambourne, Suthesh Sivapalaratnam, Kate Downes, Kousik Kundu, Lorenzo Bomba, Kim Berentsen, John R. Bradley, Louise C. Daugherty, Olivier Delaneau, Kathleen Freson, Stephen F. Garner, Luigi Grassi, Jose Guerrero, Matthias Haimel, Eva M. Janssen-Megens, Anita Kaan, Mihir Kamat, Bowon Kim, Amit Mandoli, Jonathan Marchini, Joost H.A. Martens, Stuart Meacham, Karyn Megy, Jared O'Connell, Romina Petersen, Nilofar Sharifi, Simon M. Sheard, James R. Staley, Salih Tuna, Martijn van der Ent, Klaudia Walter, Shuang-Yin Wang, Eleanor Wheeler, Steven P. Wilder, Valentina Iotchkova, Carmel Moore, Jennifer Sambrook, Hendrik G. Stunnenberg, Emanuele Di Angelantonio, Stephen Kaptoge, Taco W. Kuijpers, Enrique Carrillo-de Santa-Pau, David Juan, Daniel Rico, Alfonso Valencia, Lu Chen, Bing Ge, Louella Vasquez, Tony Kwan, Diego Garrido-Martín, Stephen Watt, Ying Yang, Roderic Guigo, Stephan Beck, Dirk S. Paul, Tomi Pastinen, David Bujold, Guillaume Bourque, Mattia Frontini, John Danesh, David J. Roberts, Willem H. Ouwehand, Adam S. Butterworth, and Nicole Soranzo. The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease. *Cell*, 167(5):1415–1429.e19, November 2016. ISSN 00928674. doi: 10.1016/j.cell.2016.10.042. URL https://linkinghub.elsevier.com/retrieve/pii/S0092867416314635.

David P. Baker, Paul J. Eslinger, Martin Benavides, Ellen Peters, Nathan F. Dieckmann, and Juan Leon. The cognitive impact of the education revolution: A possible cause of the Flynn Effect on population IQ. *Intelligence*, 49:144–158, March 2015. ISSN 01602896. doi: 10.1016/j.intell.2015.01.003. URL https://linkinghub.elsevier.com/retrieve/pii/S0160289615000082.

Sara Balduzzi, Gerta Rücker, and Guido Schwarzer. How to perform a meta-analysis with R: a practical tutorial. *Evidence Based Mental Health*, 22(4):153–160, November 2019. ISSN 1362-0347, 1468-960X. doi: 10.1136/ebmental-2019-300117. URL http://ebmh.bmj.com/lookup/doi/10.1136/ebmental-2019-300117.

Jeffrey C Barrett, James C Lee, Charles W Lees, Natalie J Prescott, Carl A Anderson, Anne Phillips, Emma Wesley, Kirstie Parnell, Hu Zhang, Hazel Drummond, Elaine R Nimmo, Dunecan Massey, Kasia Blaszczyk, Timothy Elliott, Lynn Cotterill, Helen Dallal, Alan J Lobo, Craig Mowat, Jeremy D Sanderson, Derek P Jewell, William G Newman, Cathryn Edwards, Tariq Ahmad, John C Mansfield, Jack Satsangi, Miles Parkes, Christopher G Mathew, Peter Donnelly, Leena Peltonen, Jenefer M Blackwell, Elvira Bramon, Matthew A Brown, Juan P Casas, Aiden Corvin, Nicholas Craddock, Panos Deloukas, Audrey Duncanson, Janusz Jankowski, Hugh S Markus, Christopher G Mathew, Mark I McCarthy, Colin N A Palmer, Robert Plomin, Anna Rautanen, Stephen J Sawcer, Nilesh Samani, Richard C Trembath, Ananth C Viswanathan, Nicholas Wood, Chris C A Spencer, Jeffrey C Barrett, Céline Bellenguez, Daniel Davison, Colin Freeman, Amy Strange, Peter Donnelly, Cordelia Langford, Sarah E Hunt, Sarah Edkins, Rhian Gwilliam, Hannah Blackburn, Suzannah J Bumpstead, Serge Dronov, Matthew Gillman, Emma Gray, Naomi Hammond, Alagurevathi Jayakumar, Owen T McCann, Jennifer Liddle, Marc L Perez, Simon C Potter, Radhi Ravindrarajah, Michelle Ricketts, Matthew Waller, Paul Weston, Sara Widaa, Pamela Whittaker, Panos Deloukas,

Leena Peltonen, Christopher G Mathew, Jenefer M Blackwell, Matthew A Brown, Aiden Corvin, Mark I McCarthy, Chris C A Spencer, Antony P Attwood, Jonathan Stephens, Jennifer Sambrook, Willem H Ouwehand, Wendy L McArdle, Susan M Ring, and David P Strachan. Genome-wide association study of ulcerative colitis identifies three new susceptibility loci, including the HNF4A region. *Nature Genetics*, 41(12): 1330–1334, December 2009. ISSN 1061-4036, 1546-1718. doi: 10.1038/ng.483. URL http://www.nature.com/doifinder/10.1038/ng.483.

W. Bateson. *The Progress of Genetics Since the Rediscovery of Mendel's Papers*. Progressus rei botanicae. G. Fischer, 1906. URL https://books.google.co.uk/books?id=nz8ZAAAAYAAJ.

Olivia Belbin, Kevin Morgan, Chris Medway, Donald Warden, Mario Cortina-Borja, Cornelia M. van Duijn, Hieab H.H. Adams, Ana Frank-Garcia, Keeley Brookes, Pascual Sánchez-Juan, Victoria Alvarez, Reinhard Heun, Heike Kölsch, Eliecer Coto, Patrick G. Kehoe, Eloy Rodriguez-Rodriguez, Maria J Bullido, M. Arfan Ikram, A. David Smith, and Donald J. Lehmann. The Epistasis Project: A Multi-Cohort Study of the Effects of BDNF, DBH, and SORT1 Epistasis on Alzheimer's Disease Risk. *Journal of Alzheimer's Disease*, 68(4):1535–1547, April 2019. ISSN 13872877, 18758908. doi: 10.3233/JAD-181116. URL https://www.medra.org/servlet/aliasResolver?alias=iospress&doi=10.3233/JAD-181116.

Pau Bellot, Gustavo de Los Campos, and Miguel Pérez-Enciso. Can deep learning improve genomic prediction of complex human traits? *Genetics*, 210(3):809–819, 2018.

Yoshua Bengio, Yann LeCun, et al. Scaling learning algorithms towards ai. 2007.

James Bergstra, Daniel Yamins, and David Cox. Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures. In *International conference on machine learning*, pages 115–123, 2013.

Richard J. Biedrzycki, Ashley E. Sier, Dongjing Liu, Erika N. Dreikorn, and Daniel E. Weeks. Spinning convincing stories for both true and false association signals. *Genetic Epidemiology*, 43(4):356–364, June 2019. ISSN 0741-0395, 1098-2272. doi: 10.1002/gepi.22189. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/gepi.22189.

Mariusz Bojarski, Davide Del Testa, Daniel Dworakowski, Bernhard Firner, Beat Flepp, Prasoon Goyal, Lawrence D Jackel, Mathew Monfort, Urs Muller, Jiakai Zhang, et al. End to end learning for self-driving cars. *arXiv preprint arXiv:1604.07316*, 2016.

Evan A. Boyle, Yang I. Li, and Jonathan K. Pritchard. An Expanded View of Complex Traits: From Polygenic to Omnigenic. *Cell*, 169(7):1177–1186, June 2017. ISSN 00928674. doi: 10.1016/j.cell.2017.05.038. URL https://linkinghub.elsevier.com/retrieve/pii/S0092867417306293.

Nadav Brandes, Nathan Linial, and Michal Linial. PWAS: Proteome-Wide Association Study. preprint, Bioinformatics, October 2019a. URL http://biorxiv.org/lookup/doi/10.1101/812289.

Nadav Brandes, Nathan Linial, and Michal Linial. Quantifying gene selection in cancer through protein functional alteration bias. *Nucleic Acids Research*, 47(13):6642–6655, July 2019b. ISSN 0305-1048, 1362-4962. doi: 10.1093/nar/gkz546. URL https://academic.oup.com/nar/article/47/13/6642/5523008.

Andrew Brock, Theodore Lim, James M Ritchie, and Nick Weston. Neural photo editing with introspective adversarial networks. *arXiv preprint arXiv:1609.07093*, 2016.

Sharon R. Browning and Brian L. Browning. Haplotype phasing: existing methods and new developments. *Nature Reviews Genetics*, 12(10):703–714, October 2011. ISSN 1471-0056, 1471-0064. doi: 10.1038/nrg3054. URL http://www.nature.com/articles/nrg3054.

Brendan K Bulik-Sullivan, Po-Ru Loh, Hilary K Finucane, Stephan Ripke, Jian Yang, Nick Patterson, Mark J Daly, Alkes L Price, and Benjamin M Neale. Ld score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature genetics*, 47(3):291–295, 2015.

Annalisa Buniello, Jacqueline A L MacArthur, Maria Cerezo, Laura W Harris, James Hayhurst, Cinzia Malangone, Aoife McMahon, Joannella Morales, Edward Mountjoy, Elliot Sollis, Daniel Suveges, Olga Vrousgou, Patricia L Whetzel, Ridwan Amode, Jose A Guillen, Harpreet S Riat, Stephen J Trevanion, Peggy Hall, Heather Junkins, Paul Flicek, Tony Burdett, Lucia A Hindorff, Fiona Cunningham, and Helen Parkinson. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Research*, 47(D1):D1005–D1012, January 2019. ISSN 0305-1048, 1362-4962. doi: 10.1093/nar/gky1120. URL https://academic.oup.com/nar/article/47/D1/D1005/5184712.

Oliver S. Burren, Hui Guo, and Chris Wallace. VSEAMS: a pipeline for variant set enrichment analysis using summary GWAS data identifies IKZF3, BATF and ESRRA as key transcription factors in type 1 diabetes. *Bioinformatics*, 30(23):3342–3348, December 2014. ISSN 1460-2059, 1367-4803. doi: 10.1093/bioinformatics/btu571. URL https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btu571.

Oliver S Burren, Guillermo Reales, Limy Wong, John Bowes, James C Lee, Anne Barton, Paul A Lyons, Kenneth GC Smith, Wendy Thomson, Paul DW Kirk, and Chris Wallace. Informed dimension reduction of clinically-related genome-wide association summary data characterises cross-trait axes of genetic risk. preprint, Genetics, January 2020. URL http://biorxiv.org/lookup/doi/10.1101/2020.01.14.905869.

Clare Bycroft, Colin Freeman, Desislava Petkova, Gavin Band, Lloyd T Elliott, Kevin Sharp, Allan Motyer, Damjan Vukcevic, Olivier Delaneau, Jared O'Connell, Adrian Cortes, Samantha Welsh, Gil McVean, Stephen Leslie, Peter Donnelly, and Jonathan Marchini. Genome-wide genetic data on ~500,000 UK Biobank participants. July 2017. doi: 10.1101/166298. URL http://biorxiv.org/lookup/doi/10.1101/166298.

Ashley J.R. Carter, Joachim Hermisson, and Thomas F. Hansen. The role of epistatic gene interactions in the response to selection and the evolution of evolvability. *Theoretical Population Biology*, 68(3):179–196, November 2005. ISSN 00405809. doi: 10.1016/j.tpb.2005.05.002. URL https://linkinghub.elsevier.com/retrieve/pii/S0040580905000638.

Stephane E. Castel, Alejandra Cervera, Pejman Mohammadi, François Aguet, Ferran Reverter, Aaron Wolman, Roderic Guigo, Ivan Iossifov, Ana Vasileva, and Tuuli Lappalainen. Modified penetrance of coding variants by cis-regulatory variation contributes to disease risk. *Nature Genetics*, 50(9):1327–1334, September 2018. ISSN 1061-4036, 1546-1718. doi: 10.1038/s41588-018-0192-y. URL http://www.nature.com/articles/s41588-018-0192-y.

Casimiro Castillejo-López, Angélica M Delgado-Vega, Jerome Wojcik, Sergey V Kozyrev, Elangovan Thavathiru, Ying-Yu Wu, Elena Sánchez, David Pöllmann, Juan R López-Egido, Serena Fineschi, Nicolás Domínguez, Rufei Lu, Judith A James, Joan T Merrill, Jennifer A Kelly, Kenneth M Kaufman, Kathy L Moser, Gary Gilkeson, Johan Frostegård, Bernardo A Pons-Estel, Sandra D'Alfonso, Torsten Witte, José Luis Callejas, John B Harley, Patrick M Gaffney, Javier Martin, Joel M Guthridge, and Marta E Alarcón-Riquelme. Genetic and physical interaction of the B-cell systemic lupus erythematosus-associated genes *BANK1* and *BLK*. *Annals of the Rheumatic Diseases*, 71(1):136–142, January 2012. ISSN 0003-4967, 1468-2060. doi: 10.1136/annrheumdis-2011-200085. URL http://ard.bmj.com/lookup/doi/10.1136/annrheumdis-2011-200085.

Lu Chen, Bing Ge, Francesco Paolo Casale, Louella Vasquez, Tony Kwan, Diego Garrido-Martín, Stephen Watt, Ying Yan, Kousik Kundu, Simone Ecker, Avik Datta, David Richardson, Frances Burden, Daniel Mead, Alice L. Mann, Jose Maria Fernandez, Sophia Rowlston, Steven P. Wilder, Samantha Farrow, Xiaojian Shao, John J. Lambourne, Adriana Redensek, Cornelis A. Albers, Vyacheslav Amstislavskiy, Sofie Ashford, Kim Berentsen, Lorenzo Bomba, Guillaume Bourque, David Bujold, Stephan Busche, Maxime Caron, Shu-Huang Chen, Warren Cheung, Oliver Delaneau, Emmanouil T. Dermitzakis, Heather Elding, Irina Colgiu, Frederik O. Bagger, Paul Flicek, Ehsan Habibi, Valentina Iotchkova, Eva Janssen-Megens, Bowon Kim, Hans Lehrach, Ernesto Lowy, Amit Mandoli, Filomena Matarese, Matthew T. Maurano, John A. Morris, Vera Pancaldi, Farzin Pourfarzad, Karola Rehnstrom, Augusto Rendon, Thomas Risch, Nilofar Sharifi, Marie-Michelle Simon, Marc Sultan, Alfonso Valencia, Klaudia Walter, Shuang-Yin Wang, Mattia Frontini, Stylianos E. Antonarakis, Laura Clarke, Marie-Laure Yaspo, Stephan Beck, Roderic Guigo, Daniel Rico, Joost H.A. Martens, Willem H. Ouwehand, Taco W. Kuijpers, Dirk S. Paul, Hendrik G. Stunnenberg, Oliver Stegle, Kate Downes, Tomi Pastinen, and Nicole Soranzo. Genetic Drivers of Epigenetic and Transcriptional Variation in Human Immune Cells. *Cell*, 167(5):1398–1414.e24, November 2016. ISSN 00928674. doi: 10.1016/j.cell.2016.10.026. URL https://linkinghub.elsevier.com/retrieve/pii/S0092867416314465.

Travers Ching, Daniel S Himmelstein, Brett K Beaulieu-Jones, Alexandr A Kalinin, Brian T Do, Gregory P Way, Enrico Ferrero, Paul-Michael Agapow, Michael Zietz, Michael M Hoffman, et al. Opportunities and obstacles for deep learning in biology and medicine. *Journal of The Royal Society Interface*, 15(141):20170387, 2018.

Pietro Chiurazzi and Filomena Pirozzi. Advances in understanding–genetic basis of intellectual disability. *F1000Research*, 5, 2016.

Shing Wan Choi, Timothy Shin Heng Mak, and Paul O'Reilly. A guide to performing Polygenic Risk Score analyses. *bioRxiv*, September 2018. doi: 10.1101/416545. URL http://biorxiv.org/lookup/doi/10.1101/416545.

David G. Clayton. Prediction and Interaction in Complex Disease Genetics: Experience in Type 1 Diabetes. *PLoS Genetics*, 5(7):e1000540, July 2009. ISSN 1553-7404. doi: 10.1371/journal.pgen.1000540. URL http://dx.plos.org/10.1371/journal.pgen.1000540.

Isabelle Cleynen, Gabrielle Boucher, Luke Jostins, L Philip Schumm, Sebastian Zeissig, Tariq Ahmad, Vibeke Andersen, Jane M Andrews, Vito Annese, Stephan Brand, Steven R Brant, Judy H Cho, Mark J Daly, Marla Dubinsky, Richard H Duerr, Lynnette R Ferguson, Andre Franke, Richard B Gearry, Philippe Goyette, Hakon Hakonarson, Jonas Halfvarson, Johannes R Hov, Hailang Huang, Nicholas A Kennedy, Limas Kupcinskas, Ian C Lawrance, James C Lee, Jack Satsangi, Stephan Schreiber, Emilie Théâtre, Andrea E van der Meulen-de Jong, Rinse K Weersma, David C Wilson, Miles Parkes, Severine Vermeire, John D Rioux, John Mansfield, Mark S Silverberg, Graham Radford-Smith, Dermot P B McGovern, Jeffrey C Barrett, and Charlie W Lees. Inherited determinants of Crohn's disease and ulcerative colitis phenotypes: a genetic association study. *The Lancet*, 387(10014):156–167, January 2016. ISSN 01406736. doi: 10.1016/S0140-6736(15)00465-1. URL https://linkinghub.elsevier.com/retrieve/pii/S0140673615004651.

COPDGene Investigators, ECLIPSE Investigators, LifeLines Investigators, SPIROMICS Research Group, International COPD Genetics Network Investigators, UK BiLEVE Investigators, International COPD Genetics Consortium, Brian D Hobbs, Kim de Jong, Maxime Lamontagne, Yohan Bossé, Nick Shrine, María Soler Artigas, Louise V Wain, Ian P Hall, Victoria E Jackson, Annah B Wyss, Stephanie J London, Kari E North, Nora Franceschini, David P Strachan, Terri H Beaty, John E Hokanson, James D Crapo, Peter J Castaldi, Robert P Chase, Traci M Bartz, Susan R Heckbert, Bruce M Psaty, Sina A Gharib, Pieter Zanen, Jan W Lammers, Matthijs Oudkerk, H J Groen, Nicholas Locantore, Ruth Tal-Singer, Stephen I Rennard, Jørgen Vestbo, Wim Timens, Peter D Paré, Jeanne C Latourelle, Josée Dupuis, George T O'Connor, Jemma B Wilk, Woo Jin Kim, Mi Kyeong Lee, Yeon-Mok Oh, Judith M Vonk, Harry J de Koning, Shuguang Leng, Steven A Belinsky, Yohannes Tesfaigzi, Ani Manichaikul, Xin-Qun Wang, Stephen S Rich, R Graham Barr, David Sparrow, Augusto A Litonjua, Per Bakke, Amund Gulsvik, Lies Lahousse, Guy G Brusselle, Bruno H Stricker, André G Uitterlinden, Elizabeth J Ampleford, Eugene R Bleecker, Prescott G Woodruff, Deborah A Meyers, Dandi Qiao, David A Lomas, Jae-Joon Yim, Deog Kyeom Kim, Iwona Hawrylkiewicz, Pawel Sliwinski, Megan Hardin, Tasha E Fingerlin, David A Schwartz, Dirkje S Postma, William MacNee, Martin D Tobin, Edwin K Silverman, H Marike Boezen, and Michael H Cho. Genetic loci associated with chronic obstructive pulmonary disease overlap with loci for lung function and pulmonary fibrosis. *Nature Genetics*, 49 (3):426–432, March 2017. ISSN 1061-4036, 1546-1718. doi: 10.1038/ng.3752. URL http://www.nature.com/articles/ng.3752.

H. J. Cordell. Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Human Molecular Genetics*, 11(20):2463–2468, October 2002. ISSN 14602083. doi: 10.1093/hmg/11.20.2463. URL https://academic.oup.com/hmg/article-lookup/doi/10.1093/hmg/11.20.2463.

Heather J Cordell. Detecting gene–gene interactions that underlie human diseases. *Nature Reviews Genetics*, 10(6):392–404, June 2009. ISSN 1471-0056, 1471-0064. doi: 10.1038/nrg2579. URL http://www.nature.com/articles/nrg2579.

Jerry A. Coyne, Nicholas H. Barton, and Michael Turelli. PERSPECTIVE: A CRITIQUE OF SEWALL WRIGHT'S SHIFTING BALANCE THEORY OF EVOLUTION. *Evolution*, 51(3):643–671, June 1997. ISSN 00143820. doi: 10.1111/j.1558-5646.1997.tb03650.x. URL http://doi.wiley.com/10.1111/j.1558-5646.1997.tb03650.x.

José Crossa, Gustavo de Los Campos, Paulino Pérez, Daniel Gianola, Juan Burgueño, José Luis Araus, Dan Makumbi, Ravi P. Singh, Susanne Dreisigacker, Jianbing Yan, Vivi Arief, Marianne Banziger, and Hans-Joachim Braun. Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. *Genetics*, 186(2):713–724, October 2010. ISSN 1943-2631. doi: 10.1534/genetics.110.118521.

José Crossa, Yoseph Beyene, Semagn Kassa, Paulino Pérez, John M. Hickey, Charles Chen, Gustavo de los Campos, Juan Burgueño, Vanessa S. Windhausen, Ed Buckler, Jean-Luc Jannink, Marco A. Lopez Cruz, and Raman Babu. Genomic Prediction in Maize Breeding Populations with Genotyping-by-Sequencing. *G3: Genes|Genomes|Genetics*, 3(11):1903–1926, November 2013. ISSN 2160-1836. doi: 10.1534/g3.113.008227. URL http://g3journal.org/lookup/doi/10.1534/g3.113.008227.

James F Crow. On epistasis: why it is unimportant in polygenic directional selection. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 365(1544): 1241–1244, 2010.

George Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4):303–314, 1989.

Jie Dang, Jiangxia Li, Qian Xin, Shan Shan, Xianli Bian, Qianqian Yuan, Na Liu, Xiaochun Ma, Yan Li, and Qiji Liu. Gene-gene interaction of ATG5, ATG7, BLK and BANK1 in systemic lupus erythematosus. *International Journal of Rheumatic Diseases*, 19(12): 1284–1293, December 2016. ISSN 17561841. doi: 10.1111/1756-185X.12768. URL http://doi.wiley.com/10.1111/1756-185X.12768.

Katrina M de Lange, Loukas Moutsianas, James C Lee, Christopher A Lamb, Yang Luo, Nicholas A Kennedy, Luke Jostins, Daniel L Rice, Javier Gutierrez-Achury, Sun-Gou Ji, Graham Heap, Elaine R Nimmo, Cathryn Edwards, Paul Henderson, Craig Mowat, Jeremy Sanderson, Jack Satsangi, Alison Simmons, David C Wilson, Mark Tremelling, Ailsa Hart, Christopher G Mathew, William G Newman, Miles Parkes, Charlie W Lees, Holm Uhlig, Chris Hawkey, Natalie J Prescott, Tariq Ahmad, John C Mansfield, Carl A Anderson, and Jeffrey C Barrett. Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease. *Nature Genetics*, 49(2):256–261, February 2017. ISSN 1061-4036, 1546-1718. doi: 10.1038/ng.3760. URL http://www.nature.com/articles/ng.3760.

Ronald de Vlaming and Patrick J. F. Groenen. The Current and Future Use of Ridge Regression for Prediction in Quantitative Genetics. *BioMed Research International*, 2015:1–18, 2015. ISSN 2314-6133, 2314-6141. doi: 10.1155/2015/143712. URL http://www.hindawi.com/journals/bmri/2015/143712/.

Olivier Delaneau, Cedric Coulonges, and Jean-Francois Zagury. Shape-IT: new rapid and accurate algorithm for haplotype inference. *BMC Bioinformatics*, 9(1):540, 2008. ISSN 1471-2105. doi: 10.1186/1471-2105-9-540. URL http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-9-540.

Olivier Delaneau, Jonathan Marchini, and Jean-François Zagury. A linear complexity phasing method for thousands of genomes. *Nature Methods*, 9(2):179–181, February 2012. ISSN 1548-7091, 1548-7105. doi: 10.1038/nmeth.1785. URL http://www.nature.com/articles/nmeth.1785.

Frank Dudbridge and Arief Gusnanto. Estimation of significance thresholds for genomewide association scans. *Genetic Epidemiology*, 32(3):227–234, April 2008. ISSN 07410395, 10982272. doi: 10.1002/gepi.20297. URL http://doi.wiley.com/10.1002/gepi.20297.

Ian Dunham, Ewan Birney, Bryan R Lajoie, Amartya Sanyal, Xianjun Dong, Melissa Greven, Xinying Lin, Jie Wang, Troy W Whitfield, Jiali Zhuang, et al. An integrated encyclopedia of dna elements in the human genome. 2012.

Claudia Durand and Gudrun A Rappold. Height matters—from monogenic disorders to normal variation. *Nature Reviews Endocrinology*, 9(3):171–177, 2013.

Cathy E. Elks, Marcel den Hoed, Jing Hua Zhao, Stephen J. Sharp, Nicholas J. Wareham, Ruth J. F. Loos, and Ken K. Ong. Variability in the Heritability of Body Mass Index: A Systematic Review and Meta-Regression. *Frontiers in Endocrinology*, 3, 2012. ISSN 1664-2392. doi: 10.3389/fendo.2012.00029. URL http://journal.frontiersin.org/article/10.3389/fendo.2012.00029/abstract.

Jeffrey B. Endelman. Ridge Regression and Other Kernels for Genomic Selection with R Package rrBLUP. *The Plant Genome Journal*, 4(3):250, 2011. ISSN 1940-3372. doi: 10.3835/plantgenome2011.08.0024. URL https://www.crops.org/publications/tpg/abstracts/4/3/250.

Luke M Evans, Rasool Tahmasbi, Scott I Vrieze, Gonçalo R Abecasis, Sayantan Das, Steven Gazal, Douglas W Bjelland, Teresa R De Candia, Michael E Goddard, Benjamin M Neale, et al. Comparison of methods that use whole genome data to estimate the heritability and genetic architecture of complex traits. *Nature genetics*, 50(5):737–745, 2018.

R E Everts, J Rothuizen, and B A Oost. Identification of a premature stop codon in the melanocyte-stimulating hormone receptor gene (MC1R) in Labrador and Golden retrievers with yellow coat colour. *Animal Genetics*, 31(3):194–199, June 2000. ISSN 0268-9146, 1365-2052. doi: 10.1046/j.1365-2052.2000.00639.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1046/j.1365-2052.2000.00639.x.

Kyle Kai-How Farh, Alexander Marson, Jiang Zhu, Markus Kleinewietfeld, William J. Housley, Samantha Beik, Noam Shoresh, Holly Whitton, Russell J. H. Ryan, Alexander A. Shishkin, Meital Hatan, Marlene J. Carrasco-Alfonso, Dita Mayer, C. John Luckey, Nikolaos A. Patsopoulos, Philip L. De Jager, Vijay K. Kuchroo, Charles B. Epstein, Mark J. Daly, David A. Hafler, and Bradley E. Bernstein. Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature*, 518(7539):337–343, February 2015. ISSN 0028-0836, 1476-4687. doi: 10.1038/nature13835. URL http://www.nature.com/articles/nature13835.

Katherine A Fawcett and Inês Barroso. The genetics of obesity: FTO leads the way. *Trends in genetics : TIG*, 26(6):266–274, June 2010. ISSN 0168-9525. doi: 10.1016/j.tig.

2010.02.006. URL https://pubmed.ncbi.nlm.nih.gov/20381893. Edition: 2010/04/08 Publisher: Elsevier Trends Journals.

Chloe Fawns-Ritchie and Ian J. Deary. Reliability and validity of the UK Biobank cognitive tests. *PLOS ONE*, 15(4):e0231627, April 2020. ISSN 1932-6203. doi: 10.1371/journal.pone.0231627. URL https://dx.plos.org/10.1371/journal.pone.0231627.

FinnGen. FinnGen documentation of r3 release, 2020. URL https://finngen.gitbook.io/documentation/.

Alexandra E. Fish, John A. Capra, and William S. Bush. Are Interactions between cis -Regulatory Variants Evidence for Biological Epistasis or Statistical Artifacts? *The American Journal of Human Genetics*, 99(4):817–830, October 2016. ISSN 00029297. doi: 10.1016/j.ajhg.2016.07.022. URL https://linkinghub.elsevier.com/retrieve/pii/S0002929716303238.

RA Fisher. The correlation between relatives on the supposition of mendelian inheritance. *Transactions of the Royal Society of Edinburgh*, 52:399–433, 1918.

Ronald Aylmer Fisher. *The genetical theory of natural selection*. The Clarendon Press, 1930.

Anna Fry, Thomas J Littlejohns, Cathie Sudlow, Nicola Doherty, Ligia Adamska, Tim Sprosen, Rory Collins, and Naomi E Allen. Comparison of Sociodemographic and Health-Related Characteristics of UK Biobank Participants With Those of the General Population. *American Journal of Epidemiology*, 186(9):1026–1034, November 2017. ISSN 0002-9262, 1476-6256. doi: 10.1093/aje/kwx246. URL https://academic.oup.com/aje/article/186/9/1026/3883629.

Geoff Fudenberg, David R. Kelley, and Katherine S. Pollard. Predicting 3D genome folding from DNA sequence. preprint, Genomics, October 2019. URL http://biorxiv.org/lookup/doi/10.1101/800060.

Kunihiko Fukushima and Sei Miyake. Neocognitron: A new algorithm for pattern recognition tolerant of deformations and shifts in position. *Pattern Recognition*, 15(6):455–469, January 1982. ISSN 00313203. doi: 10.1016/0031-3203(82)90024-3. URL https://linkinghub.elsevier.com/retrieve/pii/0031320382900243.

Terrence S. Furey, Praveen Sethupathy, and Shehzad Z. Sheikh. Redefining the IBDs using genome-scale molecular phenotyping. *Nature Reviews Gastroenterology & Hepatology*, 16(5):296–311, May 2019. ISSN 1759-5045, 1759-5053. doi: 10.1038/s41575-019-0118-x. URL http://www.nature.com/articles/s41575-019-0118-x.

Yarin Gal and Zoubin Ghahramani. Bayesian convolutional neural networks with bernoulli approximate variational inference. *arXiv preprint arXiv:1506.02158*, 2015.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059, 2016.

Michael D. Gallagher and Alice S. Chen-Plotkin. The Post-GWAS Era: From Association to Function. *The American Journal of Human Genetics*, 102(5):717–730, May 2018. ISSN 00029297. doi: 10.1016/j.ajhg.2018.04.002. URL https://linkinghub.elsevier.com/retrieve/pii/S0002929718301344.

GEL. 5 million genomes project, 2020. URL https://www.weka.io/wp-content/uploads/2020/01/GEL-CaseStudy_W03CS202001.pdf.

Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Deep sparse rectifier neural networks. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 315–323, 2011.

GTEx Consortium, Eric R Gamazon, Heather E Wheeler, Kaanan P Shah, Sahar V Mozaffari, Keston Aquino-Michaels, Robert J Carroll, Anne E Eyler, Joshua C Denny, Dan L Nicolae, Nancy J Cox, and Hae Kyung Im. A gene-based association method for mapping traits using reference transcriptome data. *Nature Genetics*, 47(9):1091–1098, September 2015. ISSN 1061-4036, 1546-1718. doi: 10.1038/ng.3367. URL http://www.nature.com/articles/ng.3367.

Alexander Gusev, Arthur Ko, Huwenbo Shi, Gaurav Bhatia, Wonil Chung, Brenda W J H Penninx, Rick Jansen, Eco J C de Geus, Dorret I Boomsma, Fred A Wright, Patrick F Sullivan, Elina Nikkola, Marcus Alvarez, Mete Civelek, Aldons J Lusis, Terho Lehtimäki, Emma Raitoharju, Mika Kähönen, Ilkka Seppälä, Olli T Raitakari, Johanna Kuusisto, Markku Laakso, Alkes L Price, Päivi Pajukanta, and Bogdan Pasaniuc. Integrative approaches for large-scale transcriptome-wide association studies. *Nature Genetics*, 48(3):245–252, March 2016. ISSN 1061-4036, 1546-1718. doi: 10.1038/ng.3506. URL http://www.nature.com/articles/ng.3506.

Jean Louis Guénet, Fernando Benavides, Jean-Jacques Panthier, and Xavier Montagutelli. *Genetics of the Mouse*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2015. ISBN 978-3-662-44286-9 978-3-662-44287-6. doi: 10.1007/978-3-662-44287-6. URL http://link.springer.com/10.1007/978-3-662-44287-6.

Emmanuelle Génin, Baptiste Coustet, Yannick Allanore, Ikue Ito, Maria Teruel, Arnaud Constantin, Thierry Schaeverbeke, Adeline Ruyssen-Witrand, Shigeto Tohma, Alain Cantagrel, Olivier Vittecoq, Thomas Barnetche, Xavier Le Loët, Patrice Fardellone, Hiroshi Furukawa, Olivier Meyer, Benjamin Fernández-Gutiérrez, Alejandro Balsa, Miguel A. González-Gay, Gilles Chiocchia, Naoyuki Tsuchiya, Javier Martin, and Philippe Dieudé. Epistatic Interaction between BANK1 and BLK in Rheumatoid Arthritis: Results from a Large Trans-Ethnic Meta-Analysis. *PLoS ONE*, 8(4):e61044, April 2013. ISSN 1932-6203. doi: 10.1371/journal.pone.0061044. URL https://dx.plos.org/10.1371/journal.pone.0061044.

Thomas F. Hansen. WHY EPISTASIS IS IMPORTANT FOR SELECTION AND ADAPTATION: PERSPECTIVE. *Evolution*, 67(12):3501–3511, December 2013. ISSN 00143820. doi: 10.1111/evo.12214. URL http://doi.wiley.com/10.1111/evo.12214.

Frank E Harrell Jr. *rms: Regression Modeling Strategies*, 2019. URL https://CRAN.R-project.org/package=rms. R package version 5.1-4.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Gibran Hemani, Konstantin Shakhbazov, Harm-Jan Westra, Tonu Esko, Anjali K. Henders, Allan F. McRae, Jian Yang, Greg Gibson, Nicholas G. Martin, Andres Metspalu, Lude Franke, Grant W. Montgomery, Peter M. Visscher, and Joseph E. Powell. Detection and replication of epistasis influencing transcription in humans. *Nature*, 508(7495): 249–253, February 2014. ISSN 0028-0836, 1476-4687. doi: 10.1038/nature13005. URL http://www.nature.com/doifinder/10.1038/nature13005.

C. R. Henderson. Estimation of genetic parameters. *Biometrics - vol. 6*, pages 186–187, 1950. event-place: Washington, USA.

Natalia Hernandez-Pacheco, Maria Pino-Yanes, and Carlos Flores. Genomic Predictors of Asthma Phenotypes and Treatment Response. *Frontiers in Pediatrics*, 7:6, February 2019. ISSN 2296-2360. doi: 10.3389/fped.2019.00006. URL https://www.frontiersin.org/article/10.3389/fped.2019.00006/full.

Joel Hestness, Sharan Narang, Newsha Ardalani, Gregory Diamos, Heewoo Jun, Hassan Kianinejad, Md Mostofa Ali Patwary, Yang Yang, and Yanqi Zhou. Deep Learning Scaling is Predictable, Empirically. *arXiv:1712.00409 [cs, stat]*, December 2017. URL http://arxiv.org/abs/1712.00409. arXiv: 1712.00409.

W. David Hill, Ruben C. Arslan, Charley Xia, Michelle Luciano, Carmen Amador, Pau Navarro, Caroline Hayward, Reka Nagy, David J. Porteous, Andrew M. McIntosh, Ian J. Deary, Chris S. Haley, and Lars Penke. Genomic analysis of family data reveals additional genetic effects on intelligence and personality. *Molecular Psychiatry*, 23(12):2347–2362, December 2018. ISSN 1359-4184, 1476-5578. doi: 10.1038/s41380-017-0005-1. URL http://www.nature.com/articles/s41380-017-0005-1.

L. A. Hindorff, P. Sethupathy, H. A. Junkins, E. M. Ramos, J. P. Mehta, F. S. Collins, and T. A. Manolio. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences*, 106(23): 9362–9367, June 2009. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.0903103106. URL http://www.pnas.org/cgi/doi/10.1073/pnas.0903103106.

Geoffrey E Hinton. Distributed representations. 1984.

Valentin Hivert, Julia Sidorenko, Florian Rohart, Michael E Goddard, Jian Yang, Naomi R Wray, Loic Yengo, and Peter M Visscher. Estimation of non-additive genetic variance in human complex traits from a large sample of unrelated individuals. preprint, Genetics, November 2020. URL http://biorxiv.org/lookup/doi/10.1101/2020.11.09.375501.

Adela Hruby and Frank B. Hu. The Epidemiology of Obesity: A Big Picture. *PharmacoEconomics*, 33(7):673–689, July 2015. ISSN 1170-7690, 1179-2027. doi: 10.1007/s40273-014-0243-x. URL http://link.springer.com/10.1007/s40273-014-0243-x.

Hailiang Huang, Ming Fang, Luke Jostins, Maša Umićević Mirkov, Gabrielle Boucher, Carl A. Anderson, Vibeke Andersen, Isabelle Cleynen, Adrian Cortes, François Crins, Mauro D'Amato, Valérie Deffontaine, Julia Dmitrieva, Elisa Docampo, Mahmoud Elansary, Kyle Kai-How Farh, Andre Franke, Ann-Stephan Gori, Philippe Goyette, Jonas Halfvarson, Talin Haritunians, Jo Knight, Ian C. Lawrance, Charlie W. Lees, Edouard Louis, Rob Mariman, Theo Meuwissen, Myriam Mni, Yukihide Momozawa, Miles Parkes, Sarah L. Spain, Emilie Théâtre, Gosia Trynka, Jack Satsangi, Suzanne van Sommeren, Severine Vermeire, Ramnik J. Xavier, Rinse K. Weersma, Richard H. Duerr, Christopher G. Mathew, John D. Rioux, Dermot P. B. McGovern, Judy H. Cho, Michel Georges, Mark J. Daly, and International Inflammatory Bowel Disease Genetics Consortium Barrett, Jeffrey C. Fine-mapping inflammatory bowel disease loci to single-variant resolution. *Nature*, 547(7662):173–178, July 2017. ISSN 0028-0836, 1476-4687. doi: 10.1038/nature22969. URL http://www.nature.com/articles/nature22969.

Wen Huang and Trudy F. C. Mackay. The Genetic Architecture of Quantitative Traits Cannot Be Inferred from Variance Component Analysis. *PLOS Genetics*, 12(11): e1006421, November 2016. ISSN 1553-7404. doi: 10.1371/journal.pgen.1006421. URL http://dx.plos.org/10.1371/journal.pgen.1006421.

D. H. Hubel and T. N. Wiesel. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of Physiology*, 160(1):106–154, January 1962. ISSN 00223751. doi: 10.1113/jphysiol.1962.sp006837. URL http://doi.wiley.com/10.1113/jphysiol.1962.sp006837.

International Inflammatory Bowel Disease Genetics Consortium, Philippe Goyette, Gabrielle Boucher, Dermot Mallon, Eva Ellinghaus, Luke Jostins, Hailiang Huang, Stephan Ripke, Elena S Gusareva, Vito Annese, Stephen L Hauser, Jorge R Oksenberg, Ingo Thomsen, Stephen Leslie, Mark J Daly, Kristel Van Steen, Richard H Duerr, Jeffrey C Barrett, Dermot P B McGovern, L Philip Schumm, James A Traherne, Mary N Carrington, Vasilis Kosmoliaptsis, Tom H Karlsen, Andre Franke, and John D Rioux. High-density mapping of the MHC identifies a shared role for HLA-DRB1*01:03 in inflammatory bowel diseases and heterozygous advantage in ulcerative colitis. *Nature Genetics*, 47(2): 172–179, February 2015. ISSN 1061-4036, 1546-1718. doi: 10.1038/ng.3176. URL http://www.nature.com/articles/ng.3176.

International Multiple Sclerosis Genetics Consortium, International IBD Genetics Consortium, Jimmy Z Liu, Suzanne van Sommeren, Hailiang Huang, Siew C Ng, Rudi Alberts, Atsushi Takahashi, Stephan Ripke, James C Lee, Luke Jostins, Tejas Shah, Shifteh Abedian, Jae Hee Cheon, Judy Cho, Naser E Daryani, Lude Franke, Yuta Fuyuno, Ailsa Hart, Ramesh C Juyal, Garima Juyal, Won Ho Kim, Andrew P Morris, Hossein Poustchi, William G Newman, Vandana Midha, Timothy R Orchard, Homayon Vahedi, Ajit Sood, Joseph J Y Sung, Reza Malekzadeh, Harm-Jan Westra, Keiko Yamazaki, Suk-Kyun Yang, Jeffrey C Barrett, Andre Franke, Behrooz Z Alizadeh, Miles Parkes, Thelma B K, Mark J Daly, Michiaki Kubo, Carl A Anderson, and Rinse K Weersma. Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nature Genetics*, 47(9): 979–986, September 2015. ISSN 1061-4036, 1546-1718. doi: 10.1038/ng.3359. URL http://www.nature.com/articles/ng.3359.

Sergey Ioffe and Christian Szegedy. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *arXiv:1502.03167 [cs]*, February 2015. URL http://arxiv.org/abs/1502.03167. arXiv: 1502.03167.

A. G. Ivakhnenko. Polynomial Theory of Complex Systems. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-1(4):364–378, October 1971. ISSN 0018-9472, 2168-2909. doi: 10.1109/TSMC.1971.4308320. URL http://ieeexplore.ieee.org/document/4308320/.

Arija G. Jansen, Sabine E. Mous, Tonya White, Danielle Posthuma, and Tinca J. C. Polderman. What Twin Studies Tell Us About the Heritability of Brain Development, Morphology, and Function: A Review. *Neuropsychology Review*, 25(1):27–46, March 2015. ISSN 1040-7308, 1573-6660. doi: 10.1007/s11065-015-9278-9. URL http://link.springer.com/10.1007/s11065-015-9278-9.

Yong Jiang and Jochen C. Reif. Modeling Epistasis in Genomic Selection. *Genetics*, 201 (2):759–768, October 2015. ISSN 0016-6731, 1943-2631. doi: 10.1534/genetics.115.177907. URL http://www.genetics.org/lookup/doi/10.1534/genetics.115.177907.

Åsa Johansson, Mathias Rask-Andersen, Torgny Karlsson, and Weronica E Ek. Genome-wide association analysis of 350 000 Caucasians from the UK Biobank identifies novel loci for asthma, hay fever and eczema. *Human Molecular Genetics*, 28(23):4022–4041, December 2019. ISSN 0964-6906, 1460-2083. doi: 10.1093/hmg/ddz175. URL https://academic.oup.com/hmg/article/28/23/4022/5540983.

Hyun Min Kang, Noah A. Zaitlen, Claire M. Wade, Andrew Kirby, David Heckerman, Mark J. Daly, and Eleazar Eskin. Efficient Control of Population Structure in Model Organism Association Mapping. *Genetics*, 178(3):1709–1723, March 2008. ISSN 0016-6731, 1943-2631. doi: 10.1534/genetics.107.080101. URL http://www.genetics.org/lookup/doi/10.1534/genetics.107.080101.

David R. Kelley, Yakir A. Reshef, Maxwell Bileschi, David Belanger, Cory Y. McLean, and Jasper Snoek. Sequential regulatory activity prediction across chromosomes with convolutional neural networks. *Genome Research*, 28(5):739–750, May 2018. ISSN 1088-9051, 1549-5469. doi: 10.1101/gr.227819.117. URL http://genome.cshlp.org/lookup/doi/10.1101/gr.227819.117.

Amit V. Khera, Mark Chaffin, Krishna G. Aragam, Mary E. Haas, Carolina Roselli, Seung Hoan Choi, Pradeep Natarajan, Eric S. Lander, Steven A. Lubitz, Patrick T. Ellinor, and Sekar Kathiresan. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nature Genetics*, 50(9):1219–1224, September 2018. ISSN 1061-4036, 1546-1718. doi: 10.1038/s41588-018-0183-z. URL http://www.nature.com/articles/s41588-018-0183-z.

Hwasoon Kim, Alexander Grueneberg, Ana I. Vazquez, Stephen Hsu, and Gustavo de los Campos. Will Big Data Close the Missing Heritability Gap? *Genetics*, 207(3):1135–1145, November 2017. ISSN 0016-6731, 1943-2631. doi: 10.1534/genetics.117.300271. URL http://www.genetics.org/lookup/doi/10.1534/genetics.117.300271.

Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. *arXiv:1412.6980 [cs]*, December 2014. URL http://arxiv.org/abs/1412.6980. arXiv: 1412.6980.

Günter Klambauer, Thomas Unterthiner, Andreas Mayr, and Sepp Hochreiter. Self-normalizing neural networks. In *Advances in neural information processing systems*, pages 971–980, 2017.

R. J. Klein. Complement Factor H Polymorphism in Age-Related Macular Degeneration. *Science*, 308(5720):385–389, April 2005. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.1109557. URL https://www.sciencemag.org/lookup/doi/10.1126/science.1109557.

Evan Koch, Mickey Ristroph, and Mark Kirkpatrick. Long Range Linkage Disequilibrium across the Human Genome. *PLoS ONE*, 8(12):e80754, December 2013. ISSN 1932-6203. doi: 10.1371/journal.pone.0080754. URL https://dx.plos.org/10.1371/journal.pone.0080754.

Augustine Kong, Michael L. Frigge, Gudmar Thorleifsson, Hreinn Stefansson, Alexander I. Young, Florian Zink, Gudrun A. Jonsdottir, Aysu Okbay, Patrick Sulem, Gisli Masson, Daniel F. Gudbjartsson, Agnar Helgason, Gyda Bjornsdottir, Unnur Thorsteinsdottir, and Kari Stefansson. Selection against variants in the genome associated with educational attainment. *Proceedings of the National Academy of Sciences*, 114(5):E727–E732, January 2017. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1612113114. URL http://www.pnas.org/lookup/doi/10.1073/pnas.1612113114.

Augustine Kong, Gudmar Thorleifsson, Michael L. Frigge, Bjarni J. Vilhjalmsson, Alexander I. Young, Thorgeir E. Thorgeirsson, Stefania Benonisdottir, Asmundur Oddsson, Bjarni V. Halldorsson, Gisli Masson, Daniel F. Gudbjartsson, Agnar Helgason, Gyda Bjornsdottir, Unnur Thorsteinsdottir, and Kari Stefansson. The nature of nurture: Effects of parental genotypes. *Science*, 359(6374):424–428, January 2018. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.aan6877. URL https://www.sciencemag.org/lookup/doi/10.1126/science.aan6877.

Elena Kuzmin, Benjamin VanderSluis, Wen Wang, Guihong Tan, Raamesh Deshpande, Yiqun Chen, Matej Usaj, Attila Balint, Mojca Mattiazzi Usaj, Jolanda van Leeuwen, Elizabeth N. Koch, Carles Pons, Andrius J. Dagilis, Michael Pryszlak, Jason Zi Yang Wang, Julia Hanchard, Margot Riggi, Kaicong Xu, Hamed Heydari, Bryan-Joseph San Luis, Ermira Shuteriqi, Hongwei Zhu, Nydia Van Dyk, Sara Sharifpoor, Michael Costanzo, Robbie Loewith, Amy Caudy, Daniel Bolnick, Grant W. Brown, Brenda J. Andrews, Charles Boone, and Chad L. Myers. Systematic analysis of complex genetic interactions. *Science*, 360(6386):eaao1729, April 2018. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.aao1729. URL http://www.sciencemag.org/lookup/doi/10.1126/science.aao1729.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521:436, May 2015. URL http://dx.doi.org/10.1038/nature14539.

Yann A LeCun, Léon Bottou, Genevieve B Orr, and Klaus-Robert Müller. Efficient backprop. In *Neural networks: Tricks of the trade*, pages 9–48. Springer, 2012.

James J. Lee, Robbee Wedow, Aysu Okbay, Edward Kong, Omeed Maghzian, Meghan Zacher, Tuan Anh Nguyen-Viet, Peter Bowers, Julia Sidorenko, Richard Karlsson Linnér, Mark Alan Fontana, Tushar Kundu, Chanwook Lee, Hui Li, Ruoxi Li, Rebecca Royer, Pascal N. Timshel, Raymond K. Walters, Emily A. Willoughby, Loïc Yengo, Michelle Agee, Babak Alipanahi, Adam Auton, Robert K. Bell, Katarzyna Bryc, Sarah L. Elson, Pierre Fontanillas, David A. Hinds, Jennifer C. McCreight., Karen E. Huber, Nadia K. Litterman, Matthew H. McIntyre, Joanna L. Mountain, Elizabeth S. Noblin, Carrie A. M. Northover, Steven J. Pitts, J. Fah Sathirapongsasuti, Olga V. Sazonova, Janie F. Shelton, Suyash Shringarpure, Chao Tian, Vladimir Vacic, Catherine H. Wilson, Aysu Okbay, Jonathan P. Beauchamp, Mark Alan Fontana, James J. Lee, Tune H. Pers, Cornelius A. Rietveld, Patrick Turley, Guo-Bo Chen, Valur Emilsson, S. Fleur W. Meddens, Sven Oskarsson, Joseph K. Pickrell, Kevin Thom, Pascal Timshel, Ronald de Vlaming, Abdel Abdellaoui, Tarunveer S. Ahluwalia, Jonas Bacelis, Clemens Baumbach, Gyda Bjornsdottir, Johannes H. Brandsma, Maria Pina Concas, Jaime Derringer, Nicholas A. Furlotte, Tessel E. Galesloot, Giorgia Girotto, Richa Gupta, Leanne M. Hall, Sarah E. Harris, Edith Hofer, Momoko Horikoshi, Jennifer E. Huffman, Kadri Kaasik, Ioanna P. Kalafati, Robert Karlsson, Augustine Kong, Jari Lahti, Sven J. van der Lee, Christiaan de Leeuw, Penelope A. Lind, Karl-Oskar Lindgren, Tian Liu, Massimo Mangino, Jonathan Marten, Evelin Mihailov, Michael B. Miller, Peter J. van der Most, Christopher Oldmeadow, Antony Payton, Natalia Pervjakova, Wouter J. Peyrot, Yong Qian, Olli Raitakari, Rico Rueedi, Erika Salvi, Börge Schmidt, Katharina E. Schraut, Jianxin Shi, Albert V. Smith, Raymond A. Poot, Beate St Pourcain, Alexander Teumer, Gudmar Thorleifsson, Niek Verweij, Dragana Vuckovic, Juergen Wellmann, Harm-Jan Westra, Jingyun Yang, Wei Zhao, Zhihong Zhu, Behrooz Z. Alizadeh, Najaf Amin, Andrew Bakshi, Sebastian E. Baumeister, Ginevra Biino, Klaus Bønnelykke, Patricia A. Boyle, Harry Campbell, Francesco P. Cappuccio, Gail Davies, Jan-Emmanuel De Neve, Panos Deloukas, Ilja Demuth, Jun Ding, Peter Eibich, Lewin Eisele, Niina Eklund, David M. Evans, Jessica D. Faul, Mary F. Feitosa, Andreas J. Forstner, Ilaria Gandin, Bjarni Gunnarsson, Bjarni V. Halldórsson, Tamara B. Harris, Andrew C. Heath, Lynne J. Hocking, Elizabeth G. Holliday, Georg Homuth, Michael A. Horan, Jouke-Jan Hottenga, Philip L. de Jager, Peter K. Joshi, Astanand Jugessur, Marika A. Kaakinen, Mika Kähönen, Stavroula Kanoni, Liisa Keltigangas-Järvinen, Lambertus A. L. M. Kiemeney, Ivana Kolcic, Seppo Koskinen, Aldi T. Kraja, Martin Kroh, Zoltan Kutalik, Antti Latvala, Lenore J. Launer, Maël P. Lebreton, Douglas F. Levinson, Paul Lichtenstein, Peter Lichtner, David C. M. Liewald, LifeLines Cohort Study Loukola, Anu, Pamela A. Madden, Reedik Mägi, Tomi Mäki-Opas, Riccardo E. Marioni, Pedro Marques-Vidal, Gerardus A. Meddens, George McMahon, Christa Meisinger, Thomas Meitinger, Yusplitri Milaneschi, Lili Milani, Grant W. Montgomery, Ronny Myhre, Christopher P. Nelson, Dale R. Nyholt, William E. R. Ollier, Aarno Palotie, Lavinia Paternoster, Nancy L. Pedersen, Katja E. Petrovic, David J. Porteous, Katri Räikkönen, Susan M. Ring, Antonietta Robino, Olga Rostapshova, Igor Rudan, Aldo Rustichini, Veikko Salomaa, Alan R. Sanders, Antti-Pekka Sarin, Helena Schmidt, Rodney J. Scott, Blair H. Smith, Jennifer A. Smith, Jan A. Staessen, Elisabeth Steinhagen-Thiessen, Konstantin Strauch, Antonio Terracciano, Martin D. Tobin, Sheila Ulivi, Simona Vaccargiu, Lydia Quaye, Frank J. A. van Rooij, Cristina Venturini, Anna A. E. Vinkhuyzen, Uwe Völker, Henry Völzke, Judith M. Vonk, Diego Vozzi, Johannes Waage, Erin B. Ware, Gonneke Willemsen, John R. Attia, David A. Bennett, Klaus Berger, Lars Bertram, Hans Bisgaard, Dorret I. Boomsma, Ingrid B. Borecki, Ute Bültmann, Christopher F.

Chabris, Francesco Cucca, Daniele Cusi, Ian J. Deary, George V. Dedoussis, Cornelia M. van Duijn, Johan G. Eriksson, Barbara Franke, Lude Franke, Paolo Gasparini, Pablo V. Gejman, Christian Gieger, Hans-Jörgen Grabe, Jacob Gratten, Patrick J. F. Groenen, Vilmundur Gudnason, Pim van der Harst, Caroline Hayward, David A. Hinds, Wolfgang Hoffmann, Elina Hyppönen, William G. Iacono, Bo Jacobsson, Marjo-Riitta Järvelin, Karl-Heinz Jöckel, Jaakko Kaprio, Sharon L. R. Kardia, Terho Lehtimäki, Steven F. Lehrer, Patrik K. E. Magnusson, Nicholas G. Martin, Matt McGue, Andres Metspalu, Neil Pendleton, Brenda W. J. H. Penninx, Markus Perola, Nicola Pirastu, Mario Pirastu, Ozren Polasek, Danielle Posthuma, Christine Power, Michael A. Province, Nilesh J. Samani, David Schlessinger, Reinhold Schmidt, Thorkild I. A. Sørensen, Tim D. Spector, Kari Stefansson, Unnur Thorsteinsdottir, A. Roy Thurik, Nicholas J. Timpson, Henning Tiemeier, Joyce Y. Tung, André G. Uitterlinden, Veronique Vitart, Peter Vollenweider, David R. Weir, James F. Wilson, Alan F. Wright, Dalton C. Conley, Robert F. Krueger, George Davey Smith, Albert Hofman, David I. Laibson, Sarah E. Medland, Michelle N. Meyer, Jian Yang, Magnus Johannesson, Peter M. Visscher, Tõnu Esko, Philipp D. Koellinger, David Cesarini, 23andMe Research Team, COGENT (Cognitive Genomics Consortium), and Social Science Genetic Association Consortium. Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. *Nature Genetics*, 50(8): 1112–1121, August 2018. ISSN 1546-1718. doi: 10.1038/s41588-018-0147-3. URL https://doi.org/10.1038/s41588-018-0147-3.

Phil H. Lee, Verneri Anttila, Hyejung Won, Yen-Chen A. Feng, Jacob Rosenthal, Zhaozhong Zhu, Elliot M. Tucker-Drob, Michel G. Nivard, Andrew D. Grotzinger, Danielle Posthuma, Meg M.-J. Wang, Dongmei Yu, Eli A. Stahl, Raymond K. Walters, Richard J.L. Anney, Laramie E. Duncan, Tian Ge, Rolf Adolfsson, Tobias Banaschewski, Sintia Belangero, Edwin H. Cook, Giovanni Coppola, Eske M. Derks, Pieter J. Hoekstra, Jaakko Kaprio, Anna Keski-Rahkonen, George Kirov, Henry R. Kranzler, Jurjen J. Luykx, Luis A. Rohde, Clement C. Zai, Esben Agerbo, M.J. Arranz, Philip Asherson, Marie Bækvad-Hansen, Gísli Baldursson, Mark Bellgrove, Richard A. Belliveau, Jan Buitelaar, Christie L. Burton, Jonas Bybjerg-Grauholm, Miquel Casas, Felecia Cerrato, Kimberly Chambert, Claire Churchhouse, Bru Cormand, Jennifer Crosbie, Søren Dalsgaard, Ditte Demontis, Alysa E. Doyle, Ashley Dumont, Josephine Elia, Jakob Grove, Olafur O. Gudmundsson, Jan Haavik, Hakon Hakonarson, Christine S. Hansen, Catharina A. Hartman, Ziarih Hawi, Amaia Hervás, David M. Hougaard, Daniel P. Howrigan, Hailiang Huang, Jonna Kuntsi, Kate Langley, Klaus-Peter Lesch, Patrick W.L. Leung, Sandra K. Loo, Joanna Martin, Alicia R. Martin, James J. McGough, Sarah E. Medland, Jennifer L. Moran, Ole Mors, Preben B. Mortensen, Robert D. Oades, Duncan S. Palmer, Carsten B. Pedersen, Marianne G. Pedersen, Triinu Peters, Timothy Poterba, Jesper B. Poulsen, Josep Antoni Ramos-Quiroga, Andreas Reif, Marta Ribasés, Aribert Rothenberger, Paula Rovira, Cristina Sánchez-Mora, F. Kyle Satterstrom, Russell Schachar, Maria Soler Artigas, Stacy Steinberg, Hreinn Stefansson, Patrick Turley, G. Bragi Walters, Thomas Werge, Tetyana Zayats, Dan E. Arking, Francesco Bettella, Joseph D. Buxbaum, Jane H. Christensen, Ryan L. Collins, Hilary Coon, Silvia De Rubeis, Richard Delorme, Dorothy E. Grice, Thomas F. Hansen, Peter A. Holmans, Sigrun Hope, Christina M. Hultman, Lambertus Klei, Christine Ladd-Acosta, Pall Magnusson, Terje Nærland, Mette Nyegaard, Dalila Pinto, Per Qvist, Karola Rehnström, Abraham Reichenberg, Jennifer Reichert, Kathryn Roeder, Guy A. Rouleau, Evald Saemundsen, Stephan J. Sanders, Sven Sandin, Beate St Pourcain, Kari Stefansson, James S. Sutcliffe,

Michael E. Talkowski, Lauren A. Weiss, A. Jeremy Willsey, Ingrid Agartz, Huda Akil, Diego Albani, Martin Alda, Thomas D. Als, Adebayo Anjorin, Lena Backlund, Nicholas Bass, Michael Bauer, Bernhard T. Baune, Frank Bellivier, Sarah E. Bergen, Wade H. Berrettini, Joanna M. Biernacka, Douglas H.R. Blackwood, Erlend Bøen, Monika Budde, William Bunney, Margit Burmeister, William Byerley, Enda M. Byrne, Sven Cichon, Toni-Kim Clarke, Jonathan R.I. Coleman, Nicholas Craddock, David Curtis, Piotr M. Czerski, Anders M. Dale, Nina Dalkner, Udo Dannlowski, Franziska Degenhardt, Arianna Di Florio, Torbjørn Elvsåshagen, Bruno Etain, Sascha B. Fischer, Andreas J. Forstner, Liz Forty, Josef Frank, Mark Frye, Janice M. Fullerton, Katrin Gade, Héléna A. Gaspar, Elliot S. Gershon, Michael Gill, Fernando S. Goes, Scott D. Gordon, Katherine Gordon-Smith, Melissa J. Green, Tiffany A. Greenwood, Maria Grigoroiu-Serbanescu, José Guzman-Parra, Joanna Hauser, Martin Hautzinger, Urs Heilbronner, Stefan Herms, Per Hoffmann, Dominic Holland, Stéphane Jamain, Ian Jones, Lisa A. Jones, Radhika Kandaswamy, John R. Kelsoe, James L. Kennedy, Oedegaard Ketil Joachim, Sarah Kittel-Schneider, Manolis Kogevinas, Anna C. Koller, Catharina Lavebratt, Cathryn M. Lewis, Qingqin S. Li, Jolanta Lissowska, Loes M.O. Loohuis, Susanne Lucae, Anna Maaser, Ulrik F. Malt, Nicholas G. Martin, Lina Martinsson, Susan L. McElroy, Francis J. McMahon, Andrew McQuillin, Ingrid Melle, Andres Metspalu, Vincent Millischer, Philip B. Mitchell, Grant W. Montgomery, Gunnar Morken, Derek W. Morris, Bertram Müller-Myhsok, Niamh Mullins, Richard M. Myers, Caroline M. Nievergelt, Merete Nordentoft, Annelie Nordin Adolfsson, Markus M. Nöthen, Roel A. Ophoff, Michael J. Owen, Sara A. Paciga, Carlos N. Pato, Michele T. Pato, Roy H. Perlis, Amy Perry, James B. Potash, Céline S. Reinbold, Marcella Rietschel, Margarita Rivera, Mary Roberson, Martin Schalling, Peter R. Schofield, Thomas G. Schulze, Laura J. Scott, Alessandro Serretti, Engilbert Sigurdsson, Olav B. Smeland, Eystein Stordal, Fabian Streit, Jana Strohmaier, Thorgeir E. Thorgeirsson, Jens Treutlein, Gustavo Turecki, Arne E. Vaaler, Eduard Vieta, John B. Vincent, Yunpeng Wang, Stephanie H. Witt, Peter Zandi, Roger A.H. Adan, Lars Alfredsson, Tetsuya Ando, Harald Aschauer, Jessica H. Baker, Vladimir Bencko, Andrew W. Bergen, Andreas Birgegård, Vesna Boraska Perica, Harry Brandt, Roland Burghardt, Laura Carlberg, Matteo Cassina, Maurizio Clementi, Philippe Courtet, Steven Crawford, Scott Crow, James J. Crowley, Unna N. Danner, Oliver S.P. Davis, Daniela Degortes, Janiece E. DeSocio, Danielle M. Dick, Christian Dina, Elisa Docampo, Karin Egberts, Stefan Ehrlich, Thomas Espeseth, Fernando Fernández-Aranda, Manfred M. Fichter, Lenka Foretova, Monica Forzan, Giovanni Gambaro, Ina Giegling, Fragiskos Gonidakis, Philip Gorwood, Monica Gratacos Mayora, Yiran Guo, Katherine A. Halmi, Konstantinos Hatzikotoulas, Johannes Hebebrand, Sietske G. Helder, Beate Herpertz-Dahlmann, Wolfgang Herzog, Anke Hinney, Hartmut Imgart, Susana Jiménez-Murcia, Craig Johnson, Jennifer Jordan, Antonio Julià, Deborah Kaminská, Leila Karhunen, Andreas Karwautz, Martien J.H. Kas, Walter H. Kaye, Martin A. Kennedy, Youl-Ri Kim, Lars Klareskog, Kelly L. Klump, Gun Peggy S. Knudsen, Mikael Landén, Stephanie Le Hellard, Robert D. Levitan, Dong Li, Paul Lichtenstein, Mario Maj, Sara Marsal, Sara McDevitt, James Mitchell, Palmiero Monteleone, Alessio Maria Monteleone, Melissa A. Munn-Chernoff, Benedetta Nacmias, Marie Navratilova, Julie K. O'Toole, Leonid Padyukov, Jacques Pantel, Hana Papezova, Raquel Rabionet, Anu Raevuori, Nicolas Ramoz, Ted Reichborn-Kjennerud, Valdo Ricca, Marion Roberts, Dan Rujescu, Filip Rybakowski, André Scherag, Ulrike Schmidt, Jochen Seitz, Lenka Slachtova, Margarita C.T. Slof-Op't Landt, Agnieszka Slopien, Sandro Sorbi, Lorraine Southam, Michael Strober, Alfonso Tortorella, Federica Tozzi,

Janet Treasure, Konstantinos Tziouvas, Annemarie A. van Elburg, Tracey D. Wade, Gudrun Wagner, Esther Walton, Hunna J. Watson, H-Erich Wichmann, D. Blake Woodside, Eleftheria Zeggini, Stephanie Zerwas, Stephan Zipfel, Mark J. Adams, Till F.M. Andlauer, Klaus Berger, Elisabeth B. Binder, Dorret I. Boomsma, Enrique Castelao, Lucía Colodro-Conde, Nese Direk, Anna R. Docherty, Enrico Domenici, Katharina Domschke, Erin C. Dunn, Jerome C. Foo, E.J.C. de. Geus, Hans J. Grabe, Steven P. Hamilton, Carsten Horn, Jouke-Jan Hottenga, David Howard, Marcus Ising, Stefan Kloiber, Douglas F. Levinson, Glyn Lewis, Patrik K.E. Magnusson, Hamdi Mbarek, Christel M. Middeldorp, Sara Mostafavi, Dale R. Nyholt, Brenda WJH. Penninx, Roseann E. Peterson, Giorgio Pistis, David J. Porteous, Martin Preisig, Jorge A. Quiroz, Catherine Schaefer, Eva C. Schulte, Jianxin Shi, Daniel J. Smith, Pippa A. Thomson, Henning Tiemeier, Rudolf Uher, Sandra van der Auwera, Myrna M. Weissman, Madeline Alexander, Martin Begemann, Elvira Bramon, Nancy G. Buccola, Murray J. Cairns, Dominique Campion, Vaughan J. Carr, C. Robert Cloninger, David Cohen, David A. Collier, Aiden Corvin, Lynn E. DeLisi, Gary Donohoe, Frank Dudbridge, Jubao Duan, Robert Freedman, Pablo V. Gejman, Vera Golimbet, Stephanie Godard, Hannelore Ehrenreich, Annette M. Hartmann, Frans A. Henskens, Masashi Ikeda, Nakao Iwata, Assen V. Jablensky, Inge Joa, Erik G. Jönsson, Brian J. Kelly, Jo Knight, Bettina Konte, Claudine Laurent-Levinson, Jimmy Lee, Todd Lencz, Bernard Lerer, Carmel M. Loughland, Anil K. Malhotra, Jacques Mallet, Colm McDonald, Marina Mitjans, Bryan J. Mowry, Kieran C. Murphy, Robin M. Murray, F. Anthony O'Neill, Sang-Yun Oh, Aarno Palotie, Christos Pantelis, Ann E. Pulver, Tracey L. Petryshen, Digby J. Quested, Brien Riley, Alan R. Sanders, Ulrich Schall, Sibylle G. Schwab, Rodney J. Scott, Pak C. Sham, Jeremy M. Silverman, Kang Sim, Agnes A. Steixner, Paul A. Tooney, Jim van Os, Marquis P. Vawter, Dermot Walsh, Mark Weiser, Dieter B. Wildenauer, Nigel M. Williams, Brandon K. Wormley, Fuquan Zhang, Christos Androutsos, Paul D. Arnold, Cathy L. Barr, Csaba Barta, Katharina Bey, O. Joseph Bienvenu, Donald W. Black, Lawrence W. Brown, Cathy Budman, Danielle Cath, Keun-Ah Cheon, Valentina Ciullo, Barbara J. Coffey, Daniele Cusi, Lea K. Davis, Damiaan Denys, Christel Depienne, Andrea Dietrich, Valsamma Eapen, Peter Falkai, Thomas V. Fernandez, Blanca Garcia-Delgar, Daniel A. Geller, Donald L. Gilbert, Marco A. Grados, Erica Greenberg, Edna Grünblatt, Julie Hagstrøm, Gregory L. Hanna, Andreas Hartmann, Tammy Hedderly, Gary A. Heiman, Isobel Heyman, Hyun Ju Hong, Alden Huang, Chaim Huyser, Laura Ibanez-Gomez, Ekaterina A. Khramtsova, Young Key Kim, Young-Shin Kim, Robert A. King, Yun-Joo Koh, Anastasios Konstantinidis, Sodahm Kook, Samuel Kuperman, Bennett L. Leventhal, Christine Lochner, Andrea G. Ludolph, Marcos Madruga-Garrido, Irene Malaty, Athanasios Maras, James T. McCracken, Inge A. Meijer, Pablo Mir, Astrid Morer, Kirsten R. Müller-Vahl, Alexander Münchau, Tara L. Murphy, Allan Naarden, Peter Nagy, Gerald Nestadt, Paul S. Nestadt, Humberto Nicolini, Erika L. Nurmi, Michael S. Okun, Peristera Paschou, Fabrizio Piras, Federica Piras, Christopher Pittenger, Kerstin J. Plessen, Margaret A. Richter, Renata Rizzo, Mary Robertson, Veit Roessner, Stephan Ruhrmann, Jack F. Samuels, Paul Sandor, Monika Schlögelhofer, Eun-Young Shin, Harvey Singer, Dong-Ho Song, Jungeun Song, Gianfranco Spalletta, Dan J. Stein, S Evelyn Stewart, Eric A. Storch, Barbara Stranger, Manfred Stuhrmann, Zsanett Tarnok, Jay A. Tischfield, Jennifer Tübing, Frank Visscher, Nienke Vulink, Michael Wagner, Susanne Walitza, Sina Wanderer, Martin Woods, Yulia Worbe, Gwyneth Zai, Samuel H. Zinner, Patrick F. Sullivan, Barbara Franke, Mark J. Daly, Cynthia M. Bulik, Cathryn M. Lewis, Andrew M. McIntosh, Michael C. O'Donovan,

Amanda Zheutlin, Ole A. Andreassen, Anders D. Børglum, Gerome Breen, Howard J. Edenberg, Ayman H. Fanous, Stephen V. Faraone, Joel Gelernter, Carol A. Mathews, Manuel Mattheisen, Karen S. Mitchell, Michael C. Neale, John I. Nurnberger, Stephan Ripke, Susan L. Santangelo, Jeremiah M. Scharf, Murray B. Stein, Laura M. Thornton, James T.R. Walters, Naomi R. Wray, Daniel H. Geschwind, Benjamin M. Neale, Kenneth S. Kendler, and Jordan W. Smoller. Genomic Relationships, Novel Loci, and Pleiotropic Mechanisms across Eight Psychiatric Disorders. *Cell*, 179(7):1469–1482.e11, December 2019. ISSN 00928674. doi: 10.1016/j.cell.2019.11.020. URL https://linkinghub.elsevier.com/retrieve/pii/S0092867419312760.

Sang Hong Lee, Naomi R. Wray, Michael E. Goddard, and Peter M. Visscher. Estimating Missing Heritability for Disease from Genome-wide Association Studies. *The American Journal of Human Genetics*, 88(3):294–305, March 2011. ISSN 00029297. doi: 10.1016/j.ajhg.2011.02.002. URL https://linkinghub.elsevier.com/retrieve/pii/S0002929711000206.

R. C. Lewontin. The Interaction of Selection and Linkage. I. General Considerations; Heterotic Models. *Genetics*, 49(1):49–67, January 1964. ISSN 0016-6731.

Na Li and Matthew Stephens. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics*, 165(4):2213–2233, December 2003. ISSN 0016-6731.

Dan-Yu Lin and Patrick F. Sullivan. Meta-Analysis of Genome-wide Association Studies with Overlapping Subjects. *The American Journal of Human Genetics*, 85(6):862–872, December 2009. ISSN 00029297. doi: 10.1016/j.ajhg.2009.11.001. URL https://linkinghub.elsevier.com/retrieve/pii/S0002929709005151.

Peng Lin, Sarah M. Hartz, Zhehao Zhang, Scott F. Saccone, Jia Wang, Jay A. Tischfield, Howard J. Edenberg, John R. Kramer, Alison M.Goate, Laura J. Bierut, John P. Rice, and for the COGA Collaborators COGEND Collaborators, GENEVA. A New Statistic to Evaluate Imputation Reliability. *PLoS ONE*, 5(3):e9697, March 2010. ISSN 1932-6203. doi: 10.1371/journal.pone.0009697. URL https://dx.plos.org/10.1371/journal.pone.0009697.

Christoph Lippert, Jennifer Listgarten, Robert I. Davidson, Jeff Baxter, Hoifung Poon, Carl M. Kadie, and David Heckerman. An Exhaustive Epistatic SNP Association Analysis on Expanded Wellcome Trust Data. *Scientific Reports*, 3(1):1099, December 2013. ISSN 2045-2322. doi: 10.1038/srep01099. URL http://www.nature.com/articles/srep01099.

Po-Ru Loh, Gleb Kichaev, Steven Gazal, Armin P. Schoech, and Alkes L. Price. Mixed-model association for biobank-scale datasets. *Nature Genetics*, 50(7):906–908, July 2018. ISSN 1061-4036, 1546-1718. doi: 10.1038/s41588-018-0144-6. URL http://www.nature.com/articles/s41588-018-0144-6.

Yang Luo, Katrina M de Lange, Luke Jostins, Loukas Moutsianas, Joshua Randall, Nicholas A Kennedy, Christopher A Lamb, Shane McCarthy, Tariq Ahmad, Cathryn Edwards, Eva Goncalves Serra, Ailsa Hart, Chris Hawkey, John C Mansfield, Craig Mowat, William G Newman, Sam Nichols, Martin Pollard, Jack Satsangi, Alison Simmons, Mark Tremelling, Holm Uhlig, David C Wilson, James C Lee, Natalie J

Prescott, Charlie W Lees, Christopher G Mathew, Miles Parkes, Jeffrey C Barrett, and Carl A Anderson. Exploring the genetic architecture of inflammatory bowel disease by whole-genome sequencing identifies association at ADCY7. *Nature Genetics*, 49(2): 186–192, February 2017. ISSN 1061-4036, 1546-1718. doi: 10.1038/ng.3761. URL http://www.nature.com/articles/ng.3761.

Wenlong Ma, Zhixu Qiu, Jie Song, Qian Cheng, and Chuang Ma. DeepGS: Predicting phenotypes from genotypes using Deep Learning. *bioRxiv*, December 2017. doi: 10.1101/241414. URL http://biorxiv.org/lookup/doi/10.1101/241414.

Trudy F. C. Mackay. Epistasis and quantitative traits: using model organisms to study gene–gene interactions. *Nature Reviews Genetics*, 15(1):22–33, January 2014. ISSN 1471-0056, 1471-0064. doi: 10.1038/nrg3627. URL http://www.nature.com/articles/nrg3627.

Trudy FC Mackay and Jason H Moore. Why epistasis is important for tackling complex human disease genetics. *Genome Medicine*, 6(6):125, 2014. ISSN 1756-994X. doi: 10.1186/gm561. URL http://genomemedicine.biomedcentral.com/articles/10.1186/gm561.

Timothy Shin Heng Mak, Robert Milan Porsch, Shing Wan Choi, Xueya Zhou, and Pak Chung Sham. Polygenic scores via penalized regression on summary statistics: MAK et al. *Genetic Epidemiology*, 41(6):469–480, September 2017. ISSN 07410395. doi: 10.1002/gepi.22050. URL http://doi.wiley.com/10.1002/gepi.22050.

Asko Mäki-Tanila and William G Hill. Influence of gene interaction on complex trait variation with multilocus models. *Genetics*, 198(1):355–367, 2014.

Jonathan Marchini and Bryan Howie. Genotype imputation for genome-wide association studies. *Nature Reviews Genetics*, 11(7):499–511, July 2010. ISSN 1471-0056, 1471-0064. doi: 10.1038/nrg2796. URL http://www.nature.com/articles/nrg2796.

Jonathan Marchini, Peter Donnelly, and Lon R Cardon. Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nature Genetics*, 37(4):413–417, April 2005. ISSN 1061-4036, 1546-1718. doi: 10.1038/ng1537. URL http://www.nature.com/articles/ng1537.

Jonathan Marchini, David Cutler, Nick Patterson, Matthew Stephens, Eleazar Eskin, Eran Halperin, Shin Lin, Zhaohui S. Qin, Heather M. Munro, Gonçalo R. Abecasis, and Peter Donnelly. A Comparison of Phasing Algorithms for Trios and Unrelated Individuals. *The American Journal of Human Genetics*, 78(3):437–450, March 2006. ISSN 00029297. doi: 10.1086/500808. URL https://linkinghub.elsevier.com/retrieve/pii/S0002929707623830.

Gabor T. Marth, Ian Korf, Mark D. Yandell, Raymond T. Yeh, Zhijie Gu, Hamideh Zakeri, Nathan O. Stitziel, LaDeana Hillier, Pui-Yan Kwok, and Warren R. Gish. A general approach to single-nucleotide polymorphism discovery. *Nature Genetics*, 23(4):452–456, December 1999. ISSN 1061-4036, 1546-1718. doi: 10.1038/70570. URL http://www.nature.com/articles/ng1299_452.

Alicia R. Martin, Masahiro Kanai, Yoichiro Kamatani, Yukinori Okada, Benjamin M. Neale, and Mark J. Daly. Clinical use of current polygenic risk scores may exacerbate health disparities. *Nature Genetics*, 51(4):584–591, April 2019. ISSN 1061-4036, 1546-1718. doi: 10.1038/s41588-019-0379-x. URL http://www.nature.com/articles/s41588-019-0379-x.

Nasim Mavaddat, Kyriaki Michailidou, Joe Dennis, Michael Lush, Laura Fachal, Andrew Lee, Jonathan P. Tyrer, Ting-Huei Chen, Qin Wang, Manjeet K. Bolla, Xin Yang, Muriel A. Adank, Thomas Ahearn, Kristiina Aittomäki, Jamie Allen, Irene L. Andrulis, Hoda Anton-Culver, Natalia N. Antonenkova, Volker Arndt, Kristan J. Aronson, Paul L. Auer, Päivi Auvinen, Myrto Barrdahl, Laura E. Beane Freeman, Matthias W. Beckmann, Sabine Behrens, Javier Benitez, Marina Bermisheva, Leslie Bernstein, Carl Blomqvist, Natalia V. Bogdanova, Stig E. Bojesen, Bernardo Bonanni, Anne-Lise Børresen-Dale, Hiltrud Brauch, Michael Bremer, Hermann Brenner, Adam Brentnall, Ian W. Brock, Angela Brooks-Wilson, Sara Y. Brucker, Thomas Brüning, Barbara Burwinkel, Daniele Campa, Brian D. Carter, Jose E. Castelao, Stephen J. Chanock, Rowan Chlebowski, Hans Christiansen, Christine L. Clarke, J. Margriet Collée, Emilie Cordina-Duverger, Sten Cornelissen, Fergus J. Couch, Angela Cox, Simon S. Cross, Kamila Czene, Mary B. Daly, Peter Devilee, Thilo Dörk, Isabel dos Santos-Silva, Martine Dumont, Lorraine Durcan, Miriam Dwek, Diana M. Eccles, Arif B. Ekici, A. Heather Eliassen, Carolina Ellberg, Christoph Engel, Mikael Eriksson, D. Gareth Evans, Peter A. Fasching, Jonine Figueroa, Olivia Fletcher, Henrik Flyger, Asta Försti, Lin Fritschi, Marike Gabrielson, Manuela Gago-Dominguez, Susan M. Gapstur, José A. García-Sáenz, Mia M. Gaudet, Vassilios Georgoulias, Graham G. Giles, Irina R. Gilyazova, Gord Glendon, Mark S. Goldberg, David E. Goldgar, Anna González-Neira, Grethe I. Grenaker Alnæs, Mervi Grip, Jacek Gronwald, Anne Grundy, Pascal Guénel, Lothar Haeberle, Eric Hahnen, Christopher A. Haiman, Niclas Håkansson, Ute Hamann, Susan E. Hankinson, Elaine F. Harkness, Steven N. Hart, Wei He, Alexander Hein, Jane Heyworth, Peter Hillemanns, Antoinette Hollestelle, Maartje J. Hooning, Robert N. Hoover, John L. Hopper, Anthony Howell, Guanmengqian Huang, Keith Humphreys, David J. Hunter, Milena Jakimovska, Anna Jakubowska, Wolfgang Janni, Esther M. John, Nichola Johnson, Michael E. Jones, Arja Jukkola-Vuorinen, Audrey Jung, Rudolf Kaaks, Katarzyna Kaczmarek, Vesa Kataja, Renske Keeman, Michael J. Kerin, Elza Khusnutdinova, Johanna I. Kiiski, Julia A. Knight, Yon-Dschun Ko, Veli-Matti Kosma, Stella Koutros, Vessela N. Kristensen, Ute Krüger, Tabea Kühl, Diether Lambrechts, Loic Le Marchand, Eunjung Lee, Flavio Lejbkowicz, Jenna Lilyquist, Annika Lindblom, Sara Lindström, Jolanta Lissowska, Wing-Yee Lo, Sibylle Loibl, Jirong Long, Jan Lubiński, Michael P. Lux, Robert J. MacInnis, Tom Maishman, Enes Makalic, Ivana Maleva Kostovska, Arto Mannermaa, Siranoush Manoukian, Sara Margolin, John W.M. Martens, Maria Elena Martinez, Dimitrios Mavroudis, Catriona McLean, Alfons Meindl, Usha Menon, Pooja Middha, Nicola Miller, Fernando Moreno, Anna Marie Mulligan, Claire Mulot, Victor M. Muñoz-Garzon, Susan L. Neuhausen, Heli Nevanlinna, Patrick Neven, William G. Newman, Sune F. Nielsen, Børge G. Nordestgaard, Aaron Norman, Kenneth Offit, Janet E. Olson, Håkan Olsson, Nick Orr, V. Shane Pankratz, Tjoung-Won Park-Simon, Jose I.A. Perez, Clara Pérez-Barrios, Paolo Peterlongo, Julian Peto, Mila Pinchev, Dijana Plaseska-Karanfilska, Eric C. Polley, Ross Prentice, Nadege Presneau, Darya Prokofyeva, Kristen Purrington, Katri Pylkäs, Brigitte Rack, Paolo Radice, Rohini Rau-Murthy, Gad Rennert, Hedy S. Rennert, Valerie Rhenius, Mark Robson, Atocha Romero, Kathryn J. Ruddy, Matthias Ruebner, Emmanouil Saloustros, Dale P. Sandler, Elinor J. Sawyer, Daniel F. Schmidt, Rita K. Schmutzler, Andreas Schneeweiss, Minouk J. Schoemaker,

Fredrick Schumacher, Peter Schürmann, Lukas Schwentner, Christopher Scott, Rodney J. Scott, Caroline Seynaeve, Mitul Shah, Mark E. Sherman, Martha J. Shrubsole, Xiao-Ou Shu, Susan Slager, Ann Smeets, Christof Sohn, Penny Soucy, Melissa C. Southey, John J. Spinelli, Christa Stegmaier, Jennifer Stone, Anthony J. Swerdlow, Rulla M. Tamimi, William J. Tapper, Jack A. Taylor, Mary Beth Terry, Kathrin Thöne, Rob A.E.M. Tollenaar, Ian Tomlinson, Thérèse Truong, Maria Tzardi, Hans-Ulrich Ulmer, Michael Untch, Celine M. Vachon, Elke M. van Veen, Joseph Vijai, Clarice R. Weinberg, Camilla Wendt, Alice S. Whittemore, Hans Wildiers, Walter Willett, Robert Winqvist, Alicja Wolk, Xiaohong R. Yang, Drakoulis Yannoukakos, Yan Zhang, Wei Zheng, Argyrios Ziogas, Alison M. Dunning, Deborah J. Thompson, Georgia Chenevix-Trench, Jenny Chang-Claude, Marjanka K. Schmidt, Per Hall, Roger L. Milne, Paul D.P. Pharoah, Antonis C. Antoniou, Nilanjan Chatterjee, Peter Kraft, Montserrat García-Closas, Jacques Simard, and Douglas F. Easton. Polygenic Risk Scores for Prediction of Breast Cancer and Breast Cancer Subtypes. *The American Journal of Human Genetics*, 104(1):21–34, January 2019. ISSN 00029297. doi: 10.1016/j.ajhg.2018.11.002. URL https://linkinghub.elsevier.com/retrieve/pii/S0002929718304051.

Warren S. McCulloch and Walter Pitts. A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, 5(4):115–133, December 1943. ISSN 0007-4985, 1522-9602. doi: 10.1007/BF02478259. URL http://link.springer.com/10.1007/BF02478259.

T. H. Meuwissen, B. J. Hayes, and M. E. Goddard. Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, 157(4):1819–1829, April 2001. ISSN 0016-6731.

Melinda C Mills and Charles Rahal. A scientometric review of genome-wide association studies. *Communications biology*, 2(1):1–11, 2019.

Yukihide Momozawa, Julia Dmitrieva, Emilie Théâtre, Valérie Deffontaine, Souad Rahmouni, Benoît Charloteaux, François Crins, Elisa Docampo, Mahmoud Elansary, Ann-Stephan Gori, Christelle Lecut, Rob Mariman, Myriam Mni, Cécile Oury, Ilya Altukhov, Dmitry Alexeev, Yuri Aulchenko, Leila Amininejad, Gerd Bouma, Frank Hoentjen, Mark Löwenberg, Bas Oldenburg, Marieke J. Pierik, Andrea E. vander Meulen-de Jong, C. Janneke van der Woude, Marijn C. Visschedijk, Mark Lathrop, Jean-Pierre Hugot, Rinse K. Weersma, Martine De Vos, Denis Franchimont, Severine Vermeire, Michiaki Kubo, Edouard Louis, and The International IBD Genetics Consortium Georges, Michel. IBD risk loci are enriched in multigenic regulatory modules encompassing putative causative genes. *Nature Communications*, 9(1): 2427, December 2018. ISSN 2041-1723. doi: 10.1038/s41467-018-04365-8. URL http://www.nature.com/articles/s41467-018-04365-8.

Casimiro A Curbelo Montañez, Paul Fergus, Carl Chalmers, and Jade Hind. Analysis of extremely obese individuals using deep learning stacked autoencoders and genome-wide genetic data. In *International Meeting on Computational Intelligence Methods for Bioinformatics and Biostatistics*, pages 262–276. Springer, 2018.

Jason H. Moore and Scott M. Williams. Traversing the conceptual divide between biological and statistical epistasis: systems biology and a more modern synthesis. *BioEssays*, 27

(6):637–646, June 2005. ISSN 0265-9247, 1521-1878. doi: 10.1002/bies.20236. URL http://doi.wiley.com/10.1002/bies.20236.

Alexander Mordvintsev, Christopher Olah, and Mike Tyka. Deepdream-a code example for visualizing neural networks. *Google Research*, 2(5), 2015.

Gota Morota and Daniel Gianola. Kernel-based whole-genome prediction of complex traits: a review. *Frontiers in Genetics*, 5, October 2014. ISSN 1664-8021. doi: 10.3389/fgene.2014.00363. URL http://journal.frontiersin.org/article/10.3389/fgene.2014.00363/abstract.

Gerhard Moser, Bruce Tier, Ron E Crump, Mehar S Khatkar, and Herman W Raadsma. A comparison of five methods to predict genomic breeding values of dairy bulls from genome-wide SNP markers. *Genetics Selection Evolution*, 41(1), December 2009. ISSN 1297-9686. doi: 10.1186/1297-9686-41-56. URL https://gsejournal.biomedcentral.com/articles/10.1186/1297-9686-41-56.

Hakhamanesh Mostafavi, Arbel Harpak, Ipsita Agarwal, Dalton Conley, Jonathan K Pritchard, and Molly Przeworski. Variable prediction accuracy of polygenic scores within an ancestry group. *eLife*, 9:e48376, January 2020. ISSN 2050-084X. doi: 10.7554/eLife.48376. URL https://elifesciences.org/articles/48376.

Alison A Motsinger-Reif, Scott M Dudek, Lance W Hahn, and Marylyn D Ritchie. Comparison of approaches for machine-learning optimization of neural networks for detecting gene-gene interactions in genetic epidemiology. *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society*, 32(4):325–340, 2008.

Siew C Ng, Hai Yun Shi, Nima Hamidi, Fox E Underwood, Whitney Tang, Eric I Benchimol, Remo Panaccione, Subrata Ghosh, Justin C Y Wu, Francis K L Chan, Joseph J Y Sung, and Gilaad G Kaplan. Worldwide incidence and prevalence of inflammatory bowel disease in the 21st century: a systematic review of population-based studies. *The Lancet*, 390(10114):2769–2778, December 2017. ISSN 01406736. doi: 10.1016/S0140-6736(17)32448-0. URL https://linkinghub.elsevier.com/retrieve/pii/S0140673617324480.

Magnus Nordborg and Simon Tavaré. Linkage disequilibrium: what history has to tell us. *Trends in Genetics*, 18(2):83–90, February 2002. ISSN 01689525. doi: 10.1016/S0168-9525(02)02557-X. URL https://linkinghub.elsevier.com/retrieve/pii/S016895250202557X.

Ian Osband. Risk versus uncertainty in deep learning: Bayes, bootstrap and the dangers of dropout. 2016.

Kouichi Ozaki, Yozo Ohnishi, Aritoshi Iida, Akihiko Sekine, Ryo Yamada, Tatsuhiko Tsunoda, Hiroshi Sato, Hideyuki Sato, Masatsugu Hori, Yusuke Nakamura, and Toshihiro Tanaka. Functional SNPs in the lymphotoxin-a gene that are associated with susceptibility to myocardial infarction. *Nature Genetics*, 32(4):650–654, December 2002. ISSN 1061-4036, 1546-1718. doi: 10.1038/ng1047. URL http://www.nature.com/articles/ng1047.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

Anna Peeters, Emma Gearon, Kathryn Backholer, and Bendix Carstensen. Trends in the skewness of the body mass index distribution among urban Australian adults, 1980 to 2007. *Annals of Epidemiology*, 25(1):26–33, January 2015. ISSN 10472797. doi: 10.1016/j.annepidem.2014.10.008. URL https://linkinghub.elsevier.com/retrieve/pii/S1047279714004505.

Roseann E. Peterson, Karoline Kuchenbaecker, Raymond K. Walters, Chia-Yen Chen, Alice B. Popejoy, Sathish Periyasamy, Max Lam, Conrad Iyegbe, Rona J. Strawbridge, Leslie Brick, Caitlin E. Carey, Alicia R. Martin, Jacquelyn L. Meyers, Jinni Su, Junfang Chen, Alexis C. Edwards, Allan Kalungi, Nastassja Koen, Lerato Majara, Emanuel Schwarz, Jordan W. Smoller, Eli A. Stahl, Patrick F. Sullivan, Evangelos Vassos, Bryan Mowry, Miguel L. Prieto, Alfredo Cuellar-Barboza, Tim B. Bigdeli, Howard J. Edenberg, Hailiang Huang, and Laramie E. Duncan. Genome-wide Association Studies in Ancestrally Diverse Populations: Opportunities, Methods, Pitfalls, and Recommendations. *Cell*, 179(3):589–603, October 2019. ISSN 00928674. doi: 10.1016/j.cell.2019.08.051. URL https://linkinghub.elsevier.com/retrieve/pii/S0092867419310025.

Patrick C. Phillips. Epistasis — the essential role of gene interactions in the structure and evolution of genetic systems. *Nature Reviews Genetics*, 9(11):855–867, November 2008. ISSN 1471-0056, 1471-0064. doi: 10.1038/nrg2452. URL http://www.nature.com/articles/nrg2452.

R Plomin and I J Deary. Genetics and intelligence differences: five special findings. *Molecular Psychiatry*, 20(1):98–108, February 2015. ISSN 1359-4184, 1476-5578. doi: 10.1038/mp.2014.105. URL http://www.nature.com/articles/mp2014105.

Robert Plomin and Sophie von Stumm. The new genetics of intelligence. *Nature Reviews Genetics*, 19(3):148, 2018.

Tinca J C Polderman, Beben Benyamin, Christiaan A de Leeuw, Patrick F Sullivan, Arjen van Bochoven, Peter M Visscher, and Danielle Posthuma. Meta-analysis of the heritability of human traits based on fifty years of twin studies. *Nature Genetics*, 47(7):702–709, July 2015. ISSN 1061-4036, 1546-1718. doi: 10.1038/ng.3285. URL http://www.nature.com/articles/ng.3285.

Christian Ponte-Fernández, Jorge González-Domínguez, and María J Martín. Fast search of third-order epistatic interactions on CPU and GPU clusters. *The International Journal of High Performance Computing Applications*, 34(1):20–29, January 2020. ISSN 1094-3420, 1741-2846. doi: 10.1177/1094342019852128. URL http://journals.sagepub.com/doi/10.1177/1094342019852128.

Torsten Pook, Jan Freudenthal, Arthur Korte, and Henner Simianer. Using local convolutional neural networks for genomic prediction. preprint, Genetics, May 2020. URL http://biorxiv.org/lookup/doi/10.1101/2020.05.12.090118.

Alice B Popejoy and Stephanie M Fullerton. Genomics is failing on diversity. *Nature News*, 538(7624):161, 2016.

Lutz Prechelt. Early stopping-but when? In *Neural Networks: Tricks of the trade*, pages 55–69. Springer, 1998.

Florian Privé, Julyan Arbel, and Bjarni J. Vilhjálmsson. LDpred2: better, faster, stronger. preprint, Genetics, April 2020. URL http://biorxiv.org/lookup/doi/10.1101/2020.04.28.066720.

Sara L. Pulit, Sera A. J. de With, and Paul I. W. de Bakker. Resetting the bar: Statistical significance in whole-genome sequencing-based association studies of global populations: P ulit et al . *Genetic Epidemiology*, 41(2):145–151, February 2017. ISSN 07410395. doi: 10.1002/gepi.22032. URL http://doi.wiley.com/10.1002/gepi.22032.

Chris M. Rands, Stephen Meader, Chris P. Ponting, and Gerton Lunter. 8.2% of the Human Genome Is Constrained: Variation in Rates of Turnover across Functional Element Classes in the Human Lineage. *PLoS Genetics*, 10(7):e1004525, July 2014. ISSN 1553-7404. doi: 10.1371/journal.pgen.1004525. URL https://dx.plos.org/10.1371/journal.pgen.1004525.

James C Raven. Mental tests used in genetic studies: The performance of related individuals on tests mainly educative and mainly reproductive. *Unpublished master's thesis, University of London*, 1936.

JC Raven et al. Raven manual: Section 4, advanced progressive matrices, 1988 edition, 1988.

Herve Rhinn, Ryousuke Fujita, Liang Qiang, Rong Cheng, Joseph H. Lee, and Asa Abeliovich. Integrative genomics identifies APOE e4 effectors in Alzheimer's disease. *Nature*, 500(7460):45–50, August 2013. ISSN 0028-0836, 1476-4687. doi: 10.1038/nature12415. URL http://www.nature.com/articles/nature12415.

Herve Rhinn, Ryousuke Fujita, Liang Qiang, Rong Chen, Joseph H. Lee, and Asa Abeliovich. Retraction Note: Integrative genomics identifies APOE e4 effectors in Alzheimer's disease. *Nature*, 523(7562):626–626, July 2015. ISSN 0028-0836, 1476-4687. doi: 10.1038/nature14591. URL http://www.nature.com/articles/nature14591.

Samuli Ripatti, Emmi Tikkanen, Marju Orho-Melander, Aki S Havulinna, Kaisa Silander, Amitabh Sharma, Candace Guiducci, Markus Perola, Antti Jula, Juha Sinisalo, Marja-Liisa Lokki, Markku S Nieminen, Olle Melander, Veikko Salomaa, Leena Peltonen, and Sekar Kathiresan. A multilocus genetic risk score for coronary heart disease: case-control and prospective cohort analyses. *The Lancet*, 376(9750):1393–1400, October 2010. ISSN 01406736. doi: 10.1016/S0140-6736(10)61267-6. URL https://linkinghub.elsevier.com/retrieve/pii/S0140673610612676.

Jared C. Roach, Gustavo Glusman, Robert Hubley, Stephen Z. Montsaroff, Alisha K. Holloway, Denise E. Mauldin, Deepak Srivastava, Vidu Garg, Katherine S. Pollard, David J. Galas, Leroy Hood, and Arian F.A. Smit. Chromosomal Haplotypes by Genetic Phasing of Human Families. *The American Journal of Human Genetics*, 89(3):382–397, September 2011. ISSN 00029297. doi: 10.1016/j.ajhg.2011.07.023. URL https://linkinghub.elsevier.com/retrieve/pii/S0002929711003181.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

Pau Rodríguez, Jordi Gonzalez, Guillem Cucurull, Josep M Gonfaus, and Xavier Roca. Regularizing cnns with locally constrained decorrelations. *arXiv preprint arXiv:1611.01967*, 2016.

F. Rosenblatt. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6):386–408, 1958. ISSN 1939-1471, 0033-295X. doi: 10.1037/h0042519. URL http://doi.apa.org/getdoi.cfm?doi=10.1037/h0042519.

Jaya M. Satagopan and Robert C. Elston. Evaluation of removable statistical interaction for binary traits. *Statistics in Medicine*, 32(7):1164–1190, March 2013. ISSN 02776715. doi: 10.1002/sim.5628. URL http://doi.wiley.com/10.1002/sim.5628.

Jeanne E. Savage, Philip R. Jansen, Sven Stringer, Kyoko Watanabe, Julien Bryois, Christiaan A. de Leeuw, Mats Nagel, Swapnil Awasthi, Peter B. Barr, Jonathan R. I. Coleman, Katrina L. Grasby, Anke R. Hammerschlag, Jakob A. Kaminski, Robert Karlsson, Eva Krapohl, Max Lam, Marianne Nygaard, Chandra A. Reynolds, Joey W. Trampush, Hannah Young, Delilah Zabaneh, Sara Hägg, Narelle K. Hansell, Ida K. Karlsson, Sten Linnarsson, Grant W. Montgomery, Ana B. Muñoz-Manchado, Erin B. Quinlan, Gunter Schumann, Nathan G. Skene, Bradley T. Webb, Tonya White, Dan E. Arking, Dimitrios Avramopoulos, Robert M. Bilder, Panos Bitsios, Katherine E. Burdick, Tyrone D. Cannon, Ornit Chiba-Falek, Andrea Christoforou, Elizabeth T. Cirulli, Eliza Congdon, Aiden Corvin, Gail Davies, Ian J. Deary, Pamela DeRosse, Dwight Dickinson, Srdjan Djurovic, Gary Donohoe, Emily Drabant Conley, Johan G. Eriksson, Thomas Espeseth, Nelson A. Freimer, Stella Giakoumaki, Ina Giegling, Michael Gill, David C. Glahn, Ahmad R. Hariri, Alex Hatzimanolis, Matthew C. Keller, Emma Knowles, Deborah Koltai, Bettina Konte, Jari Lahti, Stephanie Le Hellard, Todd Lencz, David C. Liewald, Edythe London, Astri J. Lundervold, Anil K. Malhotra, Ingrid Melle, Derek Morris, Anna C. Need, William Ollier, Aarno Palotie, Antony Payton, Neil Pendleton, Russell A. Poldrack, Katri Räikkönen, Ivar Reinvang, Panos Roussos, Dan Rujescu, Fred W. Sabb, Matthew A. Scult, Olav B. Smeland, Nikolaos Smyrnis, John M. Starr, Vidar M. Steen, Nikos C. Stefanis, Richard E. Straub, Kjetil Sundet, Henning Tiemeier, Aristotle N. Voineskos, Daniel R. Weinberger, Elisabeth Widen, Jin Yu, Goncalo Abecasis, Ole A. Andreassen, Gerome Breen, Lene Christiansen, Birgit Debrabant, Danielle M. Dick, Andreas Heinz, Jens Hjerling-Leffler, M. Arfan Ikram, Kenneth S. Kendler, Nicholas G. Martin, Sarah E. Medland, Nancy L. Pedersen, Robert Plomin, Tinca J. C. Polderman, Stephan Ripke, Sophie van der Sluis, Patrick F. Sullivan, Scott I. Vrieze, Margaret J. Wright, and Danielle Posthuma. Genome-wide association meta-analysis in 269,867 individuals identifies new genetic and functional links to intelligence. *Nature Genetics*, 50 (7):912–919, July 2018. ISSN 1061-4036, 1546-1718. doi: 10.1038/s41588-018-0152-6. URL http://www.nature.com/articles/s41588-018-0152-6.

Jürgen Schmidhuber. Deep learning in neural networks: An overview. *Neural Networks*, 61:85–117, January 2015. ISSN 08936080. doi: 10.1016/j.neunet.2014.09.003. URL https://linkinghub.elsevier.com/retrieve/pii/S0893608014002135.

Martin Schumacher, Reinhard Roßner, and Werner Vach. Neural networks and logistic regression: Part I. *Computational Statistics & Data Analysis*, 21(6):661–682, June 1996.

ISSN 01679473. doi: 10.1016/0167-9473(95)00032-1. URL https://linkinghub.elsevier.com/retrieve/pii/0167947395000321.

Shailja C. Shah, Hamed Khalili, Corinne Gower-Rousseau, Ola Olen, Eric I. Benchimol, Elsebeth Lynge, Kári R. Nielsen, Paul Brassard, Maria Vutcovici, Alain Bitton, Charles N. Bernstein, Desmond Leddin, Hala Tamim, Tryggvi Stefansson, Edward V. Loftus, Bjørn Moum, Whitney Tang, Siew C. Ng, Richard Gearry, Brankica Sincic, Sally Bell, Bruce E. Sands, Peter L. Lakatos, Zsuzsanna Végh, Claudia Ott, Gilaad G. Kaplan, Johan Burisch, and Jean-Frederic Colombel. Sex-Based Differences in Incidence of Inflammatory Bowel Diseases—Pooled Analysis of Population-Based Studies From Western Countries. *Gastroenterology*, 155(4):1079–1089.e3, October 2018. ISSN 00165085. doi: 10.1053/j.gastro.2018.06.043. URL https://linkinghub.elsevier.com/retrieve/pii/S0016508518346857.

P.C. Sham and S. Purcell. Equivalence between Haseman-Elston and Variance-Components Linkage Analyses for Sib Pairs. *The American Journal of Human Genetics*, 68(6):1527–1532, June 2001. ISSN 00029297. doi: 10.1086/320593. URL https://linkinghub.elsevier.com/retrieve/pii/S0002929707610659.

Jianxin Shi, Douglas F. Levinson, Jubao Duan, Alan R. Sanders, Yonglan Zheng, Itsik Pe'er, Frank Dudbridge, Peter A. Holmans, Alice S. Whittemore, Bryan J. Mowry, Ann Olincy, Farooq Amin, C. Robert Cloninger, Jeremy M. Silverman, Nancy G. Buccola, William F. Byerley, Donald W. Black, Raymond R. Crowe, Jorge R. Oksenberg, Daniel B. Mirel, Kenneth S. Kendler, Robert Freedman, and Pablo V. Gejman. Common variants on chromosome 6p22.1 are associated with schizophrenia. *Nature*, 460(7256):753–757, August 2009. ISSN 0028-0836, 1476-4687. doi: 10.1038/nature08192. URL http://www.nature.com/articles/nature08192.

David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. *arXiv:1312.6034 [cs]*, December 2013. URL http://arxiv.org/abs/1312.6034. arXiv: 1312.6034.

Montgomery Slatkin. Linkage disequilibrium — understanding the evolutionary past and mapping the medical future. *Nature Reviews Genetics*, 9(6):477–485, June 2008. ISSN 1471-0056, 1471-0064. doi: 10.1038/nrg2361. URL http://www.nature.com/articles/nrg2361.

Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017.

Daria Sorokina, Rich Caruana, Mirek Riedewald, and Daniel Fink. Detecting statistical interactions with additive groves of trees. In *Proceedings of the 25th international conference on Machine learning*, pages 1000–1007, 2008.

Sarah L. Spain and Jeffrey C. Barrett. Strategies for fine-mapping complex traits. *Human Molecular Genetics*, 24(R1):R111–R119, October 2015. ISSN 0964-6906, 1460-2083.

doi: 10.1093/hmg/ddv260. URL https://academic.oup.com/hmg/article-lookup/doi/10.1093/hmg/ddv260.

Doug Speed and David J. Balding. MultiBLUP: improved SNP-based prediction for complex traits. *Genome Research*, 24(9):1550–1557, September 2014. ISSN 1088-9051, 1549-5469. doi: 10.1101/gr.169375.113. URL http://genome.cshlp.org/lookup/doi/10.1101/gr.169375.113.

Doug Speed, Gibran Hemani, Michael R. Johnson, and David J. Balding. Improved Heritability Estimation from Genome-wide SNPs. *The American Journal of Human Genetics*, 91(6):1011–1021, December 2012. ISSN 00029297. doi: 10.1016/j.ajhg.2012.10.010. URL http://linkinghub.elsevier.com/retrieve/pii/S0002929712005332.

Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*, 2014.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

M. Stephens, N. J. Smith, and P. Donnelly. A new statistical method for haplotype reconstruction from population data. *American Journal of Human Genetics*, 68(4):978–989, April 2001. ISSN 0002-9297. doi: 10.1086/319501.

Gert Stulp, Louise Barrett, Felix C. Tropf, and Melinda Mills. Does natural selection favour taller stature among the tallest people on earth? *Proceedings of the Royal Society B: Biological Sciences*, 282(1806):20150211, May 2015. ISSN 0962-8452, 1471-2954. doi: 10.1098/rspb.2015.0211. URL https://royalsocietypublishing.org/doi/10.1098/rspb.2015.0211.

Guosheng Su, Ole F. Christensen, Tage Ostersen, Mark Henryon, and Mogens S. Lund. Estimating Additive and Non-Additive Genetic Variances and Predicting Genetic Merits Using Genome-Wide Dense Single Nucleotide Polymorphism Markers. *PLoS ONE*, 7 (9):e45293, September 2012. ISSN 1932-6203. doi: 10.1371/journal.pone.0045293. URL http://dx.plos.org/10.1371/journal.pone.0045293.

Cathie Sudlow, John Gallacher, Naomi Allen, Valerie Beral, Paul Burton, John Danesh, Paul Downey, Paul Elliott, Jane Green, Martin Landray, Bette Liu, Paul Matthews, Giok Ong, Jill Pell, Alan Silman, Alan Young, Tim Sprosen, Tim Peakman, and Rory Collins. UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLOS Medicine*, 12(3): e1001779, March 2015. ISSN 1549-1676. doi: 10.1371/journal.pmed.1001779. URL http://dx.plos.org/10.1371/journal.pmed.1001779.

Vivian Tam, Nikunj Patel, Michelle Turcotte, Yohan Bossé, Guillaume Paré, and David Meyre. Benefits and limitations of genome-wide association studies. *Nature Reviews Genetics*, 20(8):467–484, August 2019. ISSN 1471-0056, 1471-0064. doi: 10.1038/s41576-019-0127-1. URL http://www.nature.com/articles/s41576-019-0127-1.

The All of Us Research Program Investigators. The "All of Us" Research Program. *New England Journal of Medicine*, 381(7):668–676, August 2019. ISSN 0028-4793, 1533-4406. doi: 10.1056/NEJMsr1809937. URL http://www.nejm.org/doi/10.1056/NEJMsr1809937.

The Australo-Anglo-American Spondyloarthritis Consortium (TASC), the Wellcome Trust Case Control Consortium 2 (WTCCC2), Spondyloarthritis Research Consortium of Canada (SPARCC), David M Evans, Chris C A Spencer, Jennifer J Pointon, Zhan Su, David Harvey, Grazyna Kochan, Udo Oppermann, Alexander Dilthey, Matti Pirinen, Millicent A Stone, Louise Appleton, Loukas Moutsianas, Stephen Leslie, Tom Wordsworth, Tony J Kenna, Tugce Karaderi, Gethin P Thomas, Michael M Ward, Michael H Weisman, Claire Farrar, Linda A Bradbury, Patrick Danoy, Robert D Inman, Walter Maksymowych, Dafna Gladman, Proton Rahman, Ann Morgan, Helena Marzo-Ortega, Paul Bowness, Karl Gaffney, J S Hill Gaston, Malcolm Smith, Jacome Bruges-Armas, Ana-Rita Couto, Rosa Sorrentino, Fabiana Paladini, Manuel A Ferreira, Huji Xu, Yu Liu, Lei Jiang, Carlos Lopez-Larrea, Roberto Díaz-Peña, Antonio López-Vázquez, Tetyana Zayats, Gavin Band, Céline Bellenguez, Hannah Blackburn, Jenefer M Blackwell, Elvira Bramon, Suzannah J Bumpstead, Juan P Casas, Aiden Corvin, Nicholas Craddock, Panos Deloukas, Serge Dronov, Audrey Duncanson, Sarah Edkins, Colin Freeman, Matthew Gillman, Emma Gray, Rhian Gwilliam, Naomi Hammond, Sarah E Hunt, Janusz Jankowski, Alagurevathi Jayakumar, Cordelia Langford, Jennifer Liddle, Hugh S Markus, Christopher G Mathew, Owen T McCann, Mark I McCarthy, Colin N A Palmer, Leena Peltonen, Robert Plomin, Simon C Potter, Anna Rautanen, Radhi Ravindrarajah, Michelle Ricketts, Nilesh Samani, Stephen J Sawcer, Amy Strange, Richard C Trembath, Ananth C Viswanathan, Matthew Waller, Paul Weston, Pamela Whittaker, Sara Widaa, Nicholas W Wood, Gilean McVean, John D Reveille, B Paul Wordsworth, Matthew A Brown, and Peter Donnelly. Interaction between ERAP1 and HLA-B27 in ankylosing spondylitis implicates peptide handling in the mechanism for HLA-B27 in disease susceptibility. *Nature Genetics*, 43(8): 761–767, August 2011. ISSN 1061-4036, 1546-1718. doi: 10.1038/ng.873. URL http://www.nature.com/articles/ng.873.

The International IBD Genetics Consortium (IIBDGC), Luke Jostins, Stephan Ripke, Rinse K. Weersma, Richard H. Duerr, Dermot P. McGovern, Ken Y. Hui, James C. Lee, L. Philip Schumm, Yashoda Sharma, Carl A. Anderson, Jonah Essers, Mitja Mitrovic, Kaida Ning, Isabelle Cleynen, Emilie Theatre, Sarah L. Spain, Soumya Raychaudhuri, Philippe Goyette, Zhi Wei, Clara Abraham, Jean-Paul Achkar, Tariq Ahmad, Leila Amininejad, Ashwin N. Ananthakrishnan, Vibeke Andersen, Jane M. Andrews, Leonard Baidoo, Tobias Balschun, Peter A. Bampton, Alain Bitton, Gabrielle Boucher, Stephan Brand, Carsten Büning, Ariella Cohain, Sven Cichon, Mauro D'Amato, Dirk De Jong, Kathy L. Devaney, Marla Dubinsky, Cathryn Edwards, David Ellinghaus, Lynnette R. Ferguson, Denis Franchimont, Karin Fransen, Richard Gearry, Michel Georges, Christian Gieger, Jürgen Glas, Talin Haritunians, Ailsa Hart, Chris Hawkey, Matija Hedl, Xinli Hu, Tom H. Karlsen, Limas Kupcinskas, Subra Kugathasan, Anna Latiano, Debby Laukens, Ian C. Lawrance, Charlie W. Lees, Edouard Louis, Gillian Mahy, John Mansfield, Angharad R. Morgan, Craig Mowat, William Newman, Orazio Palmieri, Cyriel Y. Ponsioen, Uros Potocnik, Natalie J. Prescott, Miguel Regueiro, Jerome I. Rotter, Richard K. Russell, Jeremy D. Sanderson, Miquel Sans, Jack Satsangi, Stefan Schreiber, Lisa A. Simms, Jurgita Sventoraityte, Stephan R. Targan,

Kent D. Taylor, Mark Tremelling, Hein W. Verspaget, Martine De Vos, Cisca Wijmenga, David C. Wilson, Juliane Winkelmann, Ramnik J. Xavier, Sebastian Zeissig, Bin Zhang, Clarence K. Zhang, Hongyu Zhao, Mark S. Silverberg, Vito Annese, Hakon Hakonarson, Steven R. Brant, Graham Radford-Smith, Christopher G. Mathew, John D. Rioux, Eric E. Schadt, Mark J. Daly, Andre Franke, Miles Parkes, Severine Vermeire, Jeffrey C. Barrett, and Judy H Cho. Host–microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature*, 491(7422):119–124, November 2012. ISSN 0028-0836, 1476-4687. doi: 10.1038/nature11582. URL http://www.nature.com/articles/nature11582.

The LifeLines Cohort Study, Jian Yang, Andrew Bakshi, Zhihong Zhu, Gibran Hemani, Anna A E Vinkhuyzen, Sang Hong Lee, Matthew R Robinson, John R B Perry, Ilja M Nolte, Jana V van Vliet-Ostaptchouk, Harold Snieder, Tonu Esko, Lili Milani, Reedik Mägi, Andres Metspalu, Anders Hamsten, Patrik K E Magnusson, Nancy L Pedersen, Erik Ingelsson, Nicole Soranzo, Matthew C Keller, Naomi R Wray, Michael E Goddard, and Peter M Visscher. Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. *Nature Genetics*, 47(10): 1114–1120, October 2015. ISSN 1061-4036, 1546-1718. doi: 10.1038/ng.3390. URL http://www.nature.com/articles/ng.3390.

The UK-PSC Consortium, The International IBD Genetics Consortium, The International PSC Study Group, Sun-Gou Ji, Brian D Juran, Sören Mucha, Trine Folseraas, Luke Jostins, Espen Melum, Natsuhiko Kumasaka, Elizabeth J Atkinson, Erik M Schlicht, Jimmy Z Liu, Tejas Shah, Javier Gutierrez-Achury, Kirsten M Boberg, Annika Bergquist, Severine Vermeire, Bertus Eksteen, Peter R Durie, Martti Farkkila, Tobias Müller, Christoph Schramm, Martina Sterneck, Tobias J Weismüller, Daniel N Gotthardt, David Ellinghaus, Felix Braun, Andreas Teufel, Mattias Laudes, Wolfgang Lieb, Gunnar Jacobs, Ulrich Beuers, Rinse K Weersma, Cisca Wijmenga, Hanns-Ulrich Marschall, Piotr Milkiewicz, Albert Pares, Kimmo Kontula, Olivier Chazouillères, Pietro Invernizzi, Elizabeth Goode, Kelly Spiess, Carmel Moore, Jennifer Sambrook, Willem H Ouwehand, David J Roberts, John Danesh, Annarosa Floreani, Aliya F Gulamhusein, John E Eaton, Stefan Schreiber, Catalina Coltescu, Christopher L Bowlus, Velimir A Luketic, Joseph A Odin, Kapil B Chopra, Kris V Kowdley, Naga Chalasani, Michael P Manns, Brijesh Srivastava, George Mells, Richard N Sandford, Graeme Alexander, Daniel J Gaffney, Roger W Chapman, Gideon M Hirschfield, Mariza de Andrade, Simon M Rushbrook, Andre Franke, Tom H Karlsen, Konstantinos N Lazaridis, and Carl A Anderson. Genome-wide association study of primary sclerosing cholangitis identifies new risk loci and quantifies the genetic relationship with inflammatory bowel disease. *Nature Genetics*, 49(2):269–273, February 2017. ISSN 1061-4036, 1546-1718. doi: 10.1038/ng.3745. URL http://www.nature.com/articles/ng.3745.

Simon F. Thomsen. Genetics of asthma: an introduction for the clinician. *European Clinical Respiratory Journal*, 2(1):24643, January 2015. ISSN 2001-8525. doi: 10.3402/ecrj.v2.24643. URL https://www.tandfonline.com/doi/full/10.3402/ecrj.v2.24643.

Ali Torkamani, Nathan E. Wineinger, and Eric J. Topol. The personal and clinical utility of polygenic risk scores. *Nature Reviews Genetics*, 19(9):581–590, September 2018. ISSN 1471-0056, 1471-0064. doi: 10.1038/s41576-018-0018-x. URL http://www.nature.com/articles/s41576-018-0018-x.

Michael Tsang, Dehua Cheng, and Yan Liu. Detecting statistical interactions from neural network weights. *arXiv preprint arXiv:1705.04977*, 2017.

K. Van Steen. Travelling the world of gene-gene interactions. *Briefings in Bioinformatics*, 13 (1):1–19, January 2012a. ISSN 1467-5463, 1477-4054. doi: 10.1093/bib/bbr012. URL https://academic.oup.com/bib/article-lookup/doi/10.1093/bib/bbr012.

K. Van Steen. Travelling the world of gene-gene interactions. *Briefings in Bioinformatics*, 13 (1):1–19, January 2012b. ISSN 1467-5463, 1477-4054. doi: 10.1093/bib/bbr012. URL https://academic.oup.com/bib/article-lookup/doi/10.1093/bib/bbr012.

Aravind Vasudevan, Andrew Anderson, and David Gregg. Parallel multi channel convolution using general matrix multiplication. In *2017 IEEE 28th International Conference on Application-specific Systems, Architectures and Processors (ASAP)*, pages 19–24. IEEE, 2017.

Bjarni J. Vilhjálmsson, Jian Yang, Hilary K. Finucane, Alexander Gusev, Sara Lindström, Stephan Ripke, Giulio Genovese, Po-Ru Loh, Gaurav Bhatia, Ron Do, Tristan Hayeck, Hong-Hee Won, Sekar Kathiresan, Michele Pato, Carlos Pato, Rulla Tamimi, Eli Stahl, Noah Zaitlen, Bogdan Pasaniuc, Gillian Belbin, Eimear E. Kenny, Mikkel H. Schierup, Philip De Jager, Nikolaos A. Patsopoulos, Steve McCarroll, Mark Daly, Shaun Purcell, Daniel Chasman, Benjamin Neale, Michael Goddard, Peter M. Visscher, Peter Kraft, Nick Patterson, Alkes L. Price, Stephan Ripke, Benjamin M. Neale, Aiden Corvin, James T.R. Walters, Kai-How Farh, Peter A. Holmans, Phil Lee, Brendan Bulik-Sullivan, David A. Collier, Hailiang Huang, Tune H. Pers, Ingrid Agartz, Esben Agerbo, Margot Albus, Madeline Alexander, Farooq Amin, Silviu A. Bacanu, Martin Begemann, Richard A. Belliveau, Judit Bene, Sarah E. Bergen, Elizabeth Bevilacqua, Tim B. Bigdeli, Donald W. Black, Richard Bruggeman, Nancy G. Buccola, Randy L. Buckner, William Byerley, Wiepke Cahn, Guiqing Cai, Dominique Campion, Rita M. Cantor, Vaughan J. Carr, Noa Carrera, Stanley V. Catts, Kimberly D. Chambert, Raymond C.K. Chan, Ronald Y.L. Chen, Eric Y.H. Chen, Wei Cheng, Eric F.C. Cheung, Siow Ann Chong, C. Robert Cloninger, David Cohen, Nadine Cohen, Paul Cormican, Nick Craddock, James J. Crowley, David Curtis, Michael Davidson, Kenneth L. Davis, Franziska Degenhardt, Jurgen Del Favero, Lynn E. DeLisi, Ditte Demontis, Dimitris Dikeos, Timothy Dinan, Srdjan Djurovic, Gary Donohoe, Elodie Drapeau, Jubao Duan, Frank Dudbridge, Naser Durmishi, Peter Eichhammer, Johan Eriksson, Valentina Escott-Price, Laurent Essioux, Ayman H. Fanous, Martilias S. Farrell, Josef Frank, Lude Franke, Robert Freedman, Nelson B. Freimer, Marion Friedl, Joseph I. Friedman, Menachem Fromer, Giulio Genovese, Lyudmila Georgieva, Elliot S. Gershon, Ina Giegling, Paola Giusti-Rodrguez, Stephanie Godard, Jacqueline I. Goldstein, Vera Golimbet, Srihari Gopal, Jacob Gratten, Jakob Grove, Lieuwe de Haan, Christian Hammer, Marian L. Hamshere, Mark Hansen, Thomas Hansen, Vahram Haroutunian, Annette M. Hartmann, Frans A. Henskens, Stefan Herms, Joel N. Hirschhorn, Per Hoffmann, Andrea Hofman, Mads V. Hollegaard, David M. Hougaard, Masashi Ikeda, Inge Joa, Antonio Julia, Rene S. Kahn, Luba Kalaydjieva, Sena Karachanak-Yankova, Juha Karjalainen, David Kavanagh, Matthew C. Keller, Brian J. Kelly, James L. Kennedy, Andrey Khrunin, Yunjung Kim, Janis Klovins, James A. Knowles, Bettina Konte, Vaidutis Kucinskas, Zita Ausrele Kucinskiene, Hana Kuzelova-Ptackova, Anna K. Kahler, Claudine Laurent, Jimmy Lee Chee Keong, S. Hong Lee, Sophie E. Legge, Bernard Lerer, Miaoxin Li, Tao

Li, Kung-Yee Liang, Jeffrey Lieberman, Svetlana Limborska, Carmel M. Loughland, Jan Lubinski, Jouko Lnnqvist, Milan Macek, Patrik K.E. Magnusson, Brion S. Maher, Wolfgang Maier, Jacques Mallet, Sara Marsal, Manuel Mattheisen, Morten Mattingsdal, Robert W. McCarley, Colm McDonald, Andrew M. McIntosh, Sandra Meier, Carin J. Meijer, Bela Melegh, Ingrid Melle, Raquelle I. Mesholam-Gately, Andres Metspalu, Patricia T. Michie, Lili Milani, Vihra Milanova, Younes Mokrab, Derek W. Morris, Ole Mors, Preben B. Mortensen, Kieran C. Murphy, Robin M. Murray, Inez Myin-Germeys, Bertram Mller-Myhsok, Mari Nelis, Igor Nenadic, Deborah A. Nertney, Gerald Nestadt, Kristin K. Nicodemus, Liene Nikitina-Zake, Laura Nisenbaum, Annelie Nordin, Eadbhard O'Callaghan, Colm O'Dushlaine, F. Anthony O'Neill, Sang-Yun Oh, Ann Olincy, Line Olsen, Jim Van Os, Christos Pantelis, George N. Papadimitriou, Sergi Papiol, Elena Parkhomenko, Michele T. Pato, Tiina Paunio, Milica Pejovic-Milovancevic, Diana O. Perkins, Olli Pietilinen, Jonathan Pimm, Andrew J. Pocklington, John Powell, Alkes Price, Ann E. Pulver, Shaun M. Purcell, Digby Quested, Henrik B. Rasmussen, Abraham Reichenberg, Mark A. Reimers, Alexander L. Richards, Joshua L. Roffman, Panos Roussos, Douglas M. Ruderfer, Veikko Salomaa, Alan R. Sanders, Ulrich Schall, Christian R. Schubert, Thomas G. Schulze, Sibylle G. Schwab, Edward M. Scolnick, Rodney J. Scott, Larry J. Seidman, Jianxin Shi, Engilbert Sigurdsson, Teimuraz Silagadze, Jeremy M. Silverman, Kang Sim, Petr Slominsky, Jordan W. Smoller, Hon-Cheong So, Chris C.A. Spencer, Eli A. Stahl, Hreinn Stefansson, Stacy Steinberg, Elisabeth Stogmann, Richard E. Straub, Eric Strengman, Jana Strohmaier, T. Scott Stroup, Mythily Subramaniam, Jaana Suvisaari, Dragan M. Svrakic, Jin P. Szatkiewicz, Erik Sderman, Srinivas Thirumalai, Draga Toncheva, Paul A. Tooney, Sarah Tosato, Juha Veijola, John Waddington, Dermot Walsh, Dai Wang, Qiang Wang, Bradley T. Webb, Mark Weiser, Dieter B. Wildenauer, Nigel M. Williams, Stephanie Williams, Stephanie H. Witt, Aaron R. Wolen, Emily H.M. Wong, Brandon K. Wormley, Jing Qin Wu, Hualin Simon Xi, Clement C. Zai, Xuebin Zheng, Fritz Zimprich, Naomi R. Wray, Kari Stefansson, Peter M. Visscher, Rolf Adolfsson, Ole A. Andreassen, Douglas H.R. Blackwood, Elvira Bramon, Joseph D. Buxbaum, Anders D. Børglum, Sven Cichon, Ariel Darvasi, Enrico Domenici, Hannelore Ehrenreich, Tonu Esko, Pablo V. Gejman, Michael Gill, Hugh Gurling, Christina M. Hultman, Nakao Iwata, Assen V. Jablensky, Erik G. Jonsson, Kenneth S. Kendler, George Kirov, Jo Knight, Todd Lencz, Douglas F. Levinson, Qingqin S. Li, Jianjun Liu, Anil K. Malhotra, Steven A. McCarroll, Andrew McQuillin, Jennifer L. Moran, Preben B. Mortensen, Bryan J. Mowry, Markus M. Nthen, Roel A. Ophoff, Michael J. Owen, Aarno Palotie, Carlos N. Pato, Tracey L. Petryshen, Danielle Posthuma, Marcella Rietschel, Brien P. Riley, Dan Rujescu, Pak C. Sham, Pamela Sklar, David St. Clair, Daniel R. Weinberger, Jens R. Wendland, Thomas Werge, Mark J. Daly, Patrick F. Sullivan, Michael C. O'Donovan, Peter Kraft, David J. Hunter, Muriel Adank, Habibul Ahsan, Kristiina Aittomäki, Laura Baglietto, Sonja Berndt, Carl Blomquist, Federico Canzian, Jenny Chang-Claude, Stephen J. Chanock, Laura Crisponi, Kamila Czene, Norbert Dahmen, Isabel dos Santos Silva, Douglas Easton, A. Heather Eliassen, Jonine Figueroa, Olivia Fletcher, Montserrat Garcia-Closas, Mia M. Gaudet, Lorna Gibson, Christopher A. Haiman, Per Hall, Aditi Hazra, Rebecca Hein, Brian E. Henderson, Albert Hofman, John L. Hopper, Astrid Irwanto, Mattias Johansson, Rudolf Kaaks, Muhammad G. Kibriya, Peter Lichtner, Sara Lindström, Jianjun Liu, Eiliv Lund, Enes Makalic, Alfons Meindl, Hanne Meijers-Heijboer, Bertram Müller-Myhsok, Taru A. Muranen, Heli Nevanlinna, Petra H. Peeters, Julian Peto, Ross L. Prentice, Nazneen Rahman, María José Sánchez, Daniel F. Schmidt,

Rita K. Schmutzler, Melissa C. Southey, Rulla Tamimi, Ruth Travis, Clare Turnbull, Andre G. Uitterlinden, Rob B. van der Luijt, Quinten Waisfisz, Zhaoming Wang, Alice S. Whittemore, Rose Yang, and Wei Zheng. Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores. *The American Journal of Human Genetics*, 97 (4):576–592, October 2015. ISSN 00029297. doi: 10.1016/j.ajhg.2015.09.001. URL https://linkinghub.elsevier.com/retrieve/pii/S0002929715003651.

Peter M. Visscher, Naomi R. Wray, Qian Zhang, Pamela Sklar, Mark I. McCarthy, Matthew A. Brown, and Jian Yang. 10 Years of GWAS Discovery: Biology, Function, and Translation. *The American Journal of Human Genetics*, 101(1):5–22, July 2017. ISSN 00029297. doi: 10.1016/j.ajhg.2017.06.005. URL https://linkinghub.elsevier.com/retrieve/pii/S0002929717302409.

Peter M. Visscher, Matthew A. Brown, Mark I. McCarthy, and Jian Yang. Five Years of GWAS Discovery. *The American Journal of Human Genetics*, 90(1):7–24, January 2012. ISSN 00029297. doi: 10.1016/j.ajhg.2011.11.029. URL https://linkinghub.elsevier.com/retrieve/pii/S0002929711005337.

E Vittinghoff and DC Bauer. Case-only analysis of treatment–covariate interactions in clinical trials. *Biometrics*, 62(3):769–776, 2006.

Theo Vos, Amanuel Alemu Abajobir, Kalkidan Hassen Abate, Cristiana Abbafati, Kaja M Abbas, Foad Abd-Allah, Rizwan Suliankatchi Abdulkader, Abdishakur M Abdulle, Teshome Abuka Abebo, Semaw Ferede Abera, Victor Aboyans, Laith J Abu-Raddad, Ilana N Ackerman, Abdu Abdullahi Adamu, Olatunji Adetokunboh, Mohsen Afarideh, Ashkan Afshin, Sanjay Kumar Agarwal, Rakesh Aggarwal, Anurag Agrawal, Sutapa Agrawal, Hamid Ahmadieh, Muktar Beshir Ahmed, Miloud Taki Eddine Aichour, Amani Nidhal Aichour, Ibtihel Aichour, Sneha Aiyar, Rufus Olusola Akinyemi, Nadia Akseer, Faris Hasan Al Lami, Fares Alahdab, Ziyad Al-Aly, Khurshid Alam, Noore Alam, Tahiya Alam, Deena Alasfoor, Kefyalew Addis Alene, Raghib Ali, Reza Alizadeh-Navaei, Ala'a Alkerwi, François Alla, Peter Allebeck, Christine Allen, Fatma Al-Maskari, Rajaa Al-Raddadi, Ubai Alsharif, Shirina Alsowaidi, Khalid A Altirkawi, Azmeraw T Amare, Erfan Amini, Walid Ammar, Yaw Ampem Amoako, Hjalte H Andersen, Carl Abelardo T Antonio, Palwasha Anwari, Johan Ärnlöv, Al Artaman, Krishna Kumar Aryal, Hamid Asayesh, Solomon W Asgedom, Reza Assadi, Tesfay Mehari Atey, Niguse Tadele Atnafu, Sachin R Atre, Leticia Avila-Burgos, Euripide Frinel G Arthur Avokphako, Ashish Awasthi, Umar Bacha, Alaa Badawi, Kalpana Balakrishnan, Amitava Banerjee, Marlena S Bannick, Aleksandra Barac, Ryan M Barber, Suzanne L Barker-Collo, Till Bärnighausen, Simon Barquera, Lars Barregard, Lope H Barrero, Sanjay Basu, Bob Battista, Katherine E Battle, Bernhard T Baune, Shahrzad Bazargan-Hejazi, Justin Beardsley, Neeraj Bedi, Ettore Beghi, Yannick Béjot, Bayu Begashaw Bekele, Michelle L Bell, Derrick A Bennett, Isabela M Bensenor, Jennifer Benson, Adugnaw Berhane, Derbew Fikadu Berhe, Eduardo Bernabé, Balem Demtsu Betsu, Mircea Beuran, Addisu Shunu Beyene, Neeraj Bhala, Anil Bhansali, Samir Bhatt, Zulfiqar A Bhutta, Sibhatu Biadgilign, Burcu Kucuk Bicer, Kelly Bienhoff, Boris Bikbov, Charles Birungi, Stan Biryukov, Donal Bisanzio, Habtamu Mellie Bizuayehu, Dube Jara Boneya, Soufiane Boufous, Rupert R A Bourne, Alexandra Brazinova, Traolach S Brugha, Rachelle Buchbinder, Lemma Negesa Bulto Bulto, Blair R Bumgarner, Zahid A Butt, Lucero Cahuana-Hurtado, Ewan Cameron, Mate Car, Hélène Carabin, Jonathan R

Carapetis, Rosario Cárdenas, David O Carpenter, Juan Jesus Carrero, Austin Carter, Felix Carvalho, Daniel C Casey, Valeria Caso, Carlos A Castañeda-Orjuela, Chris D Castle, Ferrán Catalá-López, Hsing-Yi Chang, Jung-Chen Chang, Fiona J Charlson, Honglei Chen, Mirriam Chibalabala, Chioma Ezinne Chibueze, Vesper Hichilombwe Chisumpa, Abdulaal A Chitheer, Devasahayam Jesudas Christopher, Liliana G Ciobanu, Massimo Cirillo, Danny Colombara, Cyrus Cooper, Paolo Angelo Cortesi, Michael H Criqui, John A Crump, Abel Fekadu Dadi, Koustuv Dalal, Lalit Dandona, Rakhi Dandona, José das Neves, Dragos V Davitoiu, Barbora de Courten, Diego De De Leo, Barthelemy Kuate Defo, Louisa Degenhardt, Selina Deiparine, Robert P Dellavalle, Kebede Deribe, Don C Des Jarlais, Subhojit Dey, Samath D Dharmaratne, Preet Kaur Dhillon, Daniel Dicker, Eric L Ding, Shirin Djalalinia, Huyen Phuc Do, E Ray Dorsey, Kadine Priscila Bender dos Santos, Dirk Douwes-Schultz, Kerrie E Doyle, Tim R Driscoll, Manisha Dubey, Bruce Bartholow Duncan, Ziad Ziad El-Khatib, Jerisha Ellerstrand, Ahmadali Enayati, Aman Yesuf Endries, Sergey Petrovich Ermakov, Holly E Erskine, Babak Eshrati, Sharareh Eskandarieh, Alireza Esteghamati, Kara Estep, Fanuel Belayneh Bekele Fanuel, Carla Sofia E Sa Farinha, André Faro, Farshad Farzadfar, Mir Sohail Fazeli, Valery L Feigin, Seyed-Mohammad Fereshtehnejad, João C Fernandes, Alize J Ferrari, Tesfaye Regassa Feyissa, Irina Filip, Florian Fischer, Christina Fitzmaurice, Abraham D Flaxman, Luisa Sorio Flor, Nataliya Foigt, Kyle J Foreman, Richard C Franklin, Nancy Fullman, Thomas Fürst, Joao M Furtado, Neal D Futran, Emmanuela Gakidou, Morsaleh Ganji, Alberto L Garcia-Basteiro, Teshome Gebre, Tsegaye Tewelde Gebrehiwot, Ayele Geleto, Bikila Lencha Gemechu, Hailay Abrha Gesesew, Peter W Gething, Alireza Ghajar, Katherine B Gibney, Paramjit Singh Gill, Richard F Gillum, Ibrahim Abdelmageem Mohamed Ginawi, Ababi Zergaw Giref, Melkamu Dedefo Gishu, Giorgia Giussani, William W Godwin, Audra L Gold, Ellen M Goldberg, Philimon N Gona, Amador Goodridge, Sameer Vali Gopalani, Atsushi Goto, Alessandra Carvalho Goulart, Max Griswold, Harish Chander Gugnani, Rahul Gupta, Rajeev Gupta, Tanush Gupta, Vipin Gupta, Nima Hafezi-Nejad, Gessessew Bugssa Hailu, Alemayehu Desalegne Hailu, Randah Ribhi Hamadeh, Samer Hamidi, Alexis J Handal, Graeme J Hankey, Sarah Wulf Hanson, Yuantao Hao, Hilda L Harb, Habtamu Abera Hareri, Josep Maria Haro, James Harvey, Mohammad Sadegh Hassanvand, Rasmus Havmoeller, Caitlin Hawley, Simon I Hay, Roderick J Hay, Nathaniel J Henry, Ileana Beatriz Heredia-Pi, Julio Montañez Hernandez, Pouria Heydarpour, Hans W Hoek, Howard J Hoffman, Nobuyuki Horita, H Dean Hosgood, Sorin Hostiuc, Peter J Hotez, Damian G Hoy, Aung Soe Htet, Guoqing Hu, Hsiang Huang, Chantal Huynh, Kim Moesgaard Iburg, Ehimario Uche Igumbor, Chad Ikeda, Caleb Mackay Salpeter Irvine, Kathryn H Jacobsen, Nader Jahanmehr, Mihajlo B Jakovljevic, Simerjot K Jassal, Mehdi Javanbakht, Sudha P Jayaraman, Panniyammakal Jeemon, Paul N Jensen, Vivekanand Jha, Guohong Jiang, Denny John, Sarah Charlotte Johnson, Catherine O Johnson, Jost B Jonas, Mikk Jürisson, Zubair Kabir, Rajendra Kadel, Amaha Kahsay, Ritul Kamal, Haidong Kan, Nadim E Karam, André Karch, Corine Kakizi Karema, Amir Kasaeian, Getachew Mullu Kassa, Nigussie Assefa Kassaw, Nicholas J Kassebaum, Anshul Kastor, Srinivasa Vittal Katikireddi, Anil Kaul, Norito Kawakami, Peter Njenga Keiyoro, Andre Pascal Kengne, Andre Keren, Yousef Saleh Khader, Ibrahim A Khalil, Ejaz Ahmad Khan, Young-Ho Khang, Ardeshir Khosravi, Jagdish Khubchandani, Aliasghar Ahmad Kiadaliri, Christian Kieling, Yun Jin Kim, Daniel Kim, Pauline Kim, Ruth W Kimokoti, Yohannes Kinfu, Adnan Kisa, Katarzyna A Kissimova-Skarbek, Mika Kivimaki, Ann Kristin Knudsen, Yoshihiro Kokubo, Dhaval Kolte, Jacek A

Kopec, Soewarta Kosen, Parvaiz A Koul, Ai Koyanagi, Michael Kravchenko, Sanjay Krishnaswami, Kristopher J Krohn, G Anil Kumar, Pushpendra Kumar, Sanjiv Kumar, Hmwe H Kyu, Dharmesh Kumar Lal, Ratilal Lalloo, Nkurunziza Lambert, Qing Lan, Anders Larsson, Pablo M Lavados, Janet L Leasher, Paul H Lee, Jong-Tae Lee, James Leigh, Cheru Tesema Leshargie, Janni Leung, Ricky Leung, Miriam Levi, Yichong Li, Yongmei Li, Darya Li Kappe, Xiaofeng Liang, Misgan Legesse Liben, Stephen S Lim, Shai Linn, Patrick Y Liu, Angela Liu, Shiwei Liu, Yang Liu, Rakesh Lodha, Giancarlo Logroscino, Stephanie J London, Katharine J Looker, Alan D Lopez, Stefan Lorkowski, Paulo A Lotufo, Nicola Low, Rafael Lozano, Timothy C D Lucas, Erlyn Rachelle King Macarayan, Hassan Magdy Abd El Razek, Mohammed Magdy Abd El Razek, Mahdi Mahdavi, Marek Majdan, Reza Majdzadeh, Azeem Majeed, Reza Malekzadeh, Rajesh Malhotra, Deborah Carvalho Malta, Abdullah A Mamun, Helena Manguerra, Treh Manhertz, Ana Mantilla, Lorenzo G Mantovani, Chabila C Mapoma, Laurie B Marczak, Jose Martinez-Raga, Francisco Rogerlândio Martins-Melo, Ira Martopullo, Winfried März, Manu Raj Mathur, Mohsen Mazidi, Colm McAlinden, Madeline McGaughey, John J McGrath, Martin McKee, Claire McNellan, Suresh Mehata, Man Mohan Mehndiratta, Tefera Chane Mekonnen, Peter Memiah, Ziad A Memish, Walter Mendoza, Mubarek Abera Mengistie, Desalegn Tadese Mengistu, George A Mensah, Tuomo J Meretoja, Atte Meretoja, Haftay Berhane Mezgebe, Renata Micha, Anoushka Millear, Ted R Miller, Edward J Mills, Mojde Mirarefin, Erkin M Mirrakhimov, Awoke Misganaw, Shiva Raj Mishra, Philip B Mitchell, Karzan Abdulmuhsin Mohammad, Alireza Mohammadi, Kedir Endris Mohammed, Shafiu Mohammed, Sanjay K Mohanty, Ali H Mokdad, Sarah K Mollenkopf, Lorenzo Monasta, Marcella Montico, Maziar Moradi-Lakeh, Paula Moraga, Rintaro Mori, Chloe Morozoff, Shane D Morrison, Mark Moses, Cliff Mountjoy-Venning, Kalayu Birhane Mruts, Ulrich O Mueller, Kate Muller, Michele E Murdoch, Gudlavalleti Venkata Satyanarayana Murthy, Kamarul Imran Musa, Jean B Nachega, Gabriele Nagel, Mohsen Naghavi, Aliya Naheed, Kovin S Naidoo, Luigi Naldi, Vinay Nangia, Gopalakrishnan Natarajan, Dumessa Edessa Negasa, Ruxandra Irina Negoi, Ionut Negoi, Charles R Newton, Josephine Wanjiku Ngunjiri, Trang Huyen Nguyen, Quyen Le Nguyen, Cuong Tat Nguyen, Grant Nguyen, Minh Nguyen, Emma Nichols, Dina Nur Anggraini Ningrum, Sandra Nolte, Vuong Minh Nong, Bo Norrving, Jean Jacques N Noubiap, Martin J O'Donnell, Felix Akpojene Ogbo, In-Hwan Oh, Anselm Okoro, Olanrewaju Oladimeji, Tinuke Oluwasefunmi Olagunju, Andrew Toyin Olagunju, Helen E Olsen, Bolajoko Olubukunola Olusanya, Jacob Olusegun Olusanya, Kanyin Ong, John Nelson Opio, Eyal Oren, Alberto Ortiz, Aaron Osgood-Zimmerman, Majdi Osman, Mayowa O Owolabi, Mahesh Pa, Rosana E Pacella, Adrian Pana, Basant Kumar Panda, Christina Papachristou, Eun-Kee Park, Charles D Parry, Mahboubeh Parsaeian, Scott B Patten, George C Patton, Katherine Paulson, Neil Pearce, David M Pereira, Norberto Perico, Konrad Pesudovs, Carrie Beth Peterson, Max Petzold, Michael Robert Phillips, David M Pigott, Julian David Pillay, Christine Pinho, Dietrich Plass, Martin A Pletcher, Svetlana Popova, Richie G Poulton, Farshad Pourmalek, Dorairaj Prabhakaran, Noela M Prasad, Narayan Prasad, Carrie Purcell, Mostafa Qorbani, Reginald Quansah, Beatriz Paulina Ayala Quintanilla, Rynaz H S Rabiee, Amir Radfar, Anwar Rafay, Kazem Rahimi, Afarin Rahimi-Movaghar, Vafa Rahimi-Movaghar, Mohammad Hifz Ur Rahman, Mahfuzar Rahman, Rajesh Kumar Rai, Sasa Rajsic, Usha Ram, Chhabi Lal Ranabhat, Zane Rankin, Puja C Rao, Paturi Vishnupriya Rao, Salman Rawaf, Sarah E Ray, Robert C Reiner, Nikolas Reinig, Marissa B Reitsma, Giuseppe Remuzzi, Andre M N Renzaho, Serge Resnikoff, Satar

Rezaei, Antonio L Ribeiro, Luca Ronfani, Gholamreza Roshandel, Gregory A Roth, Ambuj Roy, Enrico Rubagotti, George Mugambage Ruhago, Soheil Saadat, Nafis Sadat, Mahdi Safdarian, Sare Safi, Saeid Safiri, Rajesh Sagar, Ramesh Sahathevan, Joseph Salama, Huda Omer Ba Saleem, Joshua A Salomon, Sundeep Santosh Salvi, Abdallah M Samy, Juan R Sanabria, Damian Santomauro, Itamar S Santos, João Vasco Santos, Milena M Santric Milicevic, Benn Sartorius, Maheswar Satpathy, Monika Sawhney, Sonia Saxena, Maria Inês Schmidt, Ione J C Schneider, Ben Schöttker, David C Schwebel, Falk Schwendicke, Soraya Seedat, Sadaf G Sepanlou, Edson E Servan-Mori, Tesfaye Setegn, Katya Anne Shackelford, Amira Shaheen, Masood Ali Shaikh, Mansour Shamsipour, Sheikh Mohammed Shariful Islam, Jayendra Sharma, Rajesh Sharma, Jun She, Peilin Shi, Chloe Shields, Girma Temam Shifa, Mika Shigematsu, Yukito Shinohara, Rahman Shiri, Reza Shirkoohi, Shreya Shirude, Kawkab Shishani, Mark G Shrime, Abla Mehio Sibai, Inga Dora Sigfusdottir, Diego Augusto Santos Silva, João Pedro Silva, Dayane Gabriele Alves Silveira, Jasvinder A Singh, Narinder Pal Singh, Dhirendra Narain Sinha, Eirini Skiadaresi, Vegard Skirbekk, Erica Leigh Slepak, Amber Sligar, David L Smith, Mari Smith, Badr H A Sobaih, Eugene Sobngwi, Reed J D Sorensen, Tatiane Cristina Moraes Sousa, Luciano A Sposato, Chandrashekhar T Sreeramareddy, Vinay Srinivasan, Jeffrey D Stanaway, Vasiliki Stathopoulou, Nicholas Steel, Murray B Stein, Dan J Stein, Timothy J Steiner, Caitlyn Steiner, Sabine Steinke, Mark Andrew Stokes, Lars Jacob Stovner, Bryan Strub, Michelle Subart, Muawiyyah Babale Sufiyan, Bruno F Sunguya, Patrick J Sur, Soumya Swaminathan, Bryan L Sykes, Dillon O Sylte, Rafael Tabarés-Seisdedos, Getachew Redae Taffere, Jukka S Takala, Nikhil Tandon, Mohammad Tavakkoli, Nuno Taveira, Hugh R Taylor, Arash Tehrani-Banihashemi, Tesfalidet Tekelab, Abdullah Sulieman Terkawi, Dawit Jember Tesfaye, Belay Tesssema, Ornwipa Thamsuwan, Katie E Thomas, Amanda G Thrift, Tenaw Yimer Tiruye, Ruoyan Tobe-Gai, Mette C Tollanes, Marcello Tonelli, Roman Topor-Madry, Miguel Tortajada, Mathilde Touvier, Bach Xuan Tran, Suryakant Tripathi, Christopher Troeger, Thomas Truelsen, Derrick Tsoi, Kald Beshir Tuem, Emin Murat Tuzcu, Stefanos Tyrovolas, Kingsley N Ukwaja, Eduardo A Undurraga, Chigozie Jesse Uneke, Rachel Updike, Olalekan A Uthman, Benjamin S Chudi Uzochukwu, Job F M van Boven, Santosh Varughese, Tommi Vasankari, S Venkatesh, Narayanaswamy Venketasubramanian, Ramesh Vidavalur, Francesco S Violante, Sergey K Vladimirov, Vasiliy Victorovich Vlassov, Stein Emil Vollset, Fiseha Wadilo, Tolassa Wakayo, Yuan-Pang Wang, Marcia Weaver, Scott Weichenthal, Elisabete Weiderpass, Robert G Weintraub, Andrea Werdecker, Ronny Westerman, Harvey A Whiteford, Tissa Wijeratne, Charles Shey Wiysonge, Charles D A Wolfe, Rachel Woodbrook, Anthony D Woolf, Abdulhalik Workicho, Denis Xavier, Gelin Xu, Simon Yadgir, Mohsen Yaghoubi, Bereket Yakob, Lijing L Yan, Yuichiro Yano, Pengpeng Ye, Hassen Hamid Yimam, Paul Yip, Naohiro Yonemoto, Seok-Jun Yoon, Marcel Yotebieng, Mustafa Z Younis, Zoubida Zaidi, Maysaa El Sayed Zaki, Elias Asfaw Zegeye, Zerihun Menlkalew Zenebe, Xueying Zhang, Maigeng Zhou, Ben Zipkin, Sanjay Zodpey, Liesl Joanna Zuhlke, and Christopher J L Murray. Global, regional, and national incidence, prevalence, and years lived with disability for 328 diseases and injuries for 195 countries, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016. *The Lancet*, 390(10100):1211–1259, September 2017. ISSN 01406736. doi: 10.1016/S0140-6736(17)32154-2. URL https://linkinghub.elsevier.com/retrieve/pii/S0140673617321542.

Damjan Vukcevic, Eliana Hechter, Chris Spencer, and Peter Donnelly. Disease model distortion in association studies. *Genetic Epidemiology*, 35(4):278–290, May 2011.

ISSN 0741-0395, 1098-2272. doi: 10.1002/gepi.20576. URL https://onlinelibrary.wiley.com/doi/10.1002/gepi.20576.

Urmo Võsa, Annique Claringbould, Harm-Jan Westra, Marc Jan Bonder, Patrick Deelen, Biao Zeng, Holger Kirsten, Ashis Saha, Roman Kreuzhuber, Silva Kasela, Natalia Pervjakova, Isabel Alvaes, Marie-Julie Fave, Mawusse Agbessi, Mark Christiansen, Rick Jansen, Ilkka Seppälä, Lin Tong, Alexander Teumer, Katharina Schramm, Gibran Hemani, Joost Verlouw, Hanieh Yaghootkar, Reyhan Sönmez, Andrew Brown, Viktorija Kukushkina, Anette Kalnapenkis, Sina Rüeger, Eleonora Porcu, Jaanika Kronberg-Guzman, Johannes Kettunen, Joseph Powell, Bernett Lee, Futao Zhang, Wibowo Arindrarto, Frank Beutner, BIOS Consortium, Harm Brugge, i2QTL Consortium, Julia Dmitreva, Mahmoud Elansary, Benjamin P. Fairfax, Michel Georges, Bastiaan T. Heijmans, Mika Kähönen, Yungil Kim, Julian C. Knight, Peter Kovacs, Knut Krohn, Shuang Li, Markus Loeffler, Urko M. Marigorta, Hailang Mei, Yukihide Momozawa, Martina Müller-Nurasyid, Matthias Nauck, Michel Nivard, Brenda Penninx, Jonathan Pritchard, Olli Raitakari, Olaf Rotzchke, Eline P. Slagboom, Coen D.A. Stehouwer, Michael Stumvoll, Patrick Sullivan, Peter A.C. 't Hoen, Joachim Thiery, Anke Tönjes, Jenny van Dongen, Maarten van Iterson, Jan Veldink, Uwe Völker, Cisca Wijmenga, Morris Swertz, Anand Andiappan, Grant W. Montgomery, Samuli Ripatti, Markus Perola, Zoltan Kutalik, Emmanouil Dermitzakis, Sven Bergmann, Timothy Frayling, Joyce van Meurs, Holger Prokisch, Habibul Ahsan, Brandon Pierce, Terho Lehtimäki, Dorret Boomsma, Bruce M. Psaty, Sina A. Gharib, Philip Awadalla, Lili Milani, Willem Ouwehand, Kate Downes, Oliver Stegle, Alexis Battle, Jian Yang, Peter M. Visscher, Markus Scholz, Gregory Gibson, Tõnu Esko, and Lude Franke. Unraveling the polygenic architecture of complex traits using blood eQTL metaanalysis. preprint, Genomics, October 2018. URL http://biorxiv.org/lookup/doi/10.1101/447367.

Louise V Wain, Nick Shrine, Suzanne Miller, Victoria E Jackson, Ioanna Ntalla, María Soler Artigas, Charlotte K Billington, Abdul Kader Kheirallah, Richard Allen, James P Cook, Kelly Probert, Ma'en Obeidat, Yohan Bossé, Ke Hao, Dirkje S Postma, Peter D Paré, Adaikalavan Ramasamy, Reedik Mägi, Evelin Mihailov, Eva Reinmaa, Erik Melén, Jared O'Connell, Eleni Frangou, Olivier Delaneau, Colin Freeman, Desislava Petkova, Mark McCarthy, Ian Sayers, Panos Deloukas, Richard Hubbard, Ian Pavord, Anna L Hansell, Neil C Thomson, Eleftheria Zeggini, Andrew P Morris, Jonathan Marchini, David P Strachan, Martin D Tobin, and Ian P Hall. Novel insights into the genetics of smoking behaviour, lung function, and chronic obstructive pulmonary disease (UK BiLEVE): a genetic association study in UK Biobank. *The Lancet Respiratory Medicine*, 3(10):769–781, October 2015. ISSN 22132600. doi: 10.1016/S2213-2600(15)00283-0. URL https://linkinghub.elsevier.com/retrieve/pii/S2213260015002830.

Pierrick Wainschtein, Deepti P. Jain, Loic Yengo, Zhili Zheng, TOPMed Anthropometry Working Group, Trans-Omics for Precision Medicine Consortium, L. Adrienne Cupples, Aladdin H. Shadyab, Barbara McKnight, Benjamin M. Shoemaker, Braxton D. Mitchell, Bruce M. Psaty, Charles Kooperberg, Dan Roden, Dawood Darbar, Donna K. Arnett, Elizabeth A. Regan, Eric Boerwinkle, Jerome I. Rotter, Matthew A. Allison, Merry-Lynn N. McDonald, Mina K Chung, Nicholas L. Smith, Patrick T. Ellinor, Ramachandran S. Vasan, Rasika A. Mathias, Stephen S. Rich, Susan R. Heckbert, Susan Redline, Xiuqing Guo, Y.-D Ida Chen, Ching-Ti Liu, Mariza de Andrade, Lisa R. Yanek, Christine M. Albert, Ryan D. Hernandez, Stephen T. McGarvey, Kari E. North, Leslie A.

Lange, Bruce S. Weir, Cathy C. Laurie, Jian Yang, and Peter M. Visscher. Recovery of trait heritability from whole genome sequence data. preprint, Genetics, March 2019. URL http://biorxiv.org/lookup/doi/10.1101/588020.

Xiang Wan, Can Yang, Qiang Yang, Hong Xue, Xiaodan Fan, Nelson L.S. Tang, and Weichuan Yu. BOOST: A Fast Approach to Detecting Gene-Gene Interactions in Genome-wide Case-Control Studies. *The American Journal of Human Genetics*, 87(3): 325–340, September 2010. ISSN 00029297. doi: 10.1016/j.ajhg.2010.07.021. URL https://linkinghub.elsevier.com/retrieve/pii/S0002929710003782.

Gao Wang, Abhishek Sarkar, Peter Carbonetto, and Matthew Stephens. A simple new approach to variable selection in regression, with application to genetic fine mapping. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(5): 1273–1300, December 2020. ISSN 1369-7412, 1467-9868. doi: 10.1111/rssb.12388. URL https://onlinelibrary.wiley.com/doi/10.1111/rssb.12388.

Jianhua Wang, Dandan Huang, Yao Zhou, Hongcheng Yao, Huanhuan Liu, Sinan Zhai, Chengwei Wu, Zhanye Zheng, Ke Zhao, Zhao Wang, Xianfu Yi, Shijie Zhang, Xiaorong Liu, Zipeng Liu, Kexin Chen, Ying Yu, Pak Chung Sham, and Mulin Jun Li. CAUSALdb: a database for disease/trait causal variants identified using summary statistics of genome-wide association studies. *Nucleic Acids Research*, page gkz1026, November 2019. ISSN 0305-1048, 1362-4962. doi: 10.1093/nar/gkz1026. URL https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkz1026/5613679.

Xuefeng Wang, Robert C. Elston, and Xiaofeng Zhu. The Meaning of Interaction. *Human Heredity*, 70(4):269–277, 2010. ISSN 1423-0062, 0001-5652. doi: 10.1159/000321967. URL https://www.karger.com/Article/FullText/321967.

Xuefeng Wang, Robert C. Elston, and Xiaofeng Zhu. Statistical interaction in human genetics: how should we model it if we are looking for biological interaction? *Nature Reviews Genetics*, 12(1):74–74, January 2011. ISSN 1471-0056, 1471-0064. doi: 10.1038/nrg2579-c2. URL http://www.nature.com/articles/nrg2579-c2.

Wen-Hua Wei, Yunfei Guo, Alida S.D. Kindt, Tony R. Merriman, Colin A. Semple, Kai Wang, and Chris S. Haley. Abundant local interactions in the 4p16.1 region suggest functional mechanisms underlying SLC2A9 associations with human serum uric acid. *Human Molecular Genetics*, 23(19):5061–5068, October 2014a. ISSN 0964-6906, 1460-2083. doi: 10.1093/hmg/ddu227. URL https://academic.oup.com/hmg/article-lookup/doi/10.1093/hmg/ddu227.

Wen-Hua Wei, Gibran Hemani, and Chris S. Haley. Detecting epistasis in human complex traits. *Nature Reviews Genetics*, 15(11):722–733, November 2014b. ISSN 1471-0056, 1471-0064. doi: 10.1038/nrg3747. URL http://www.nature.com/articles/nrg3747.

Omer Weissbrod, Dan Geiger, and Saharon Rosset. Multikernel linear mixed models for complex phenotype prediction. *Genome Research*, 26(7):969–979, July 2016. ISSN 1088-9051, 1549-5469. doi: 10.1101/gr.201996.115. URL http://genome.cshlp.org/lookup/doi/10.1101/gr.201996.115.

Juyang Weng, Narendra Ahuja, and Thomas S Huang. Cresceptron: a self-organizing neural network which grows adaptively. In *[Proceedings 1992] IJCNN International Joint Conference on Neural Networks*, volume 1, pages 576–581. IEEE, 1992.

Paul Werbos and Paul John. Beyond regression : new tools for prediction and analysis in the behavioral sciences /. 01 1974.

Rick L. Williams. A Note on Robust Variance Estimation for Cluster-Correlated Data. *Biometrics*, 56(2):645–646, June 2000. ISSN 0006341X. doi: 10.1111/j.0006-341X.2000.00645.x. URL http://doi.wiley.com/10.1111/j.0006-341X.2000.00645.x.

Genevieve L. Wojcik, Mariaelisa Graff, Katherine K. Nishimura, Ran Tao, Jeffrey Haessler, Christopher R. Gignoux, Heather M. Highland, Yesha M. Patel, Elena P. Sorokin, Christy L. Avery, Gillian M. Belbin, Stephanie A. Bien, Iona Cheng, Sinead Cullina, Chani J. Hodonsky, Yao Hu, Laura M. Huckins, Janina Jeff, Anne E. Justice, Jonathan M. Kocarnik, Unhee Lim, Bridget M. Lin, Yingchang Lu, Sarah C. Nelson, Sung-Shim L. Park, Hannah Poisner, Michael H. Preuss, Melissa A. Richard, Claudia Schurmann, Veronica W. Setiawan, Alexandra Sockell, Karan Vahi, Marie Verbanck, Abhishek Vishnu, Ryan W. Walker, Kristin L. Young, Niha Zubair, Victor Acuña-Alonso, Jose Luis Ambite, Kathleen C. Barnes, Eric Boerwinkle, Erwin P. Bottinger, Carlos D. Bustamante, Christian Caberto, Samuel Canizales-Quinteros, Matthew P. Conomos, Ewa Deelman, Ron Do, Kimberly Doheny, Lindsay Fernández-Rhodes, Myriam Fornage, Benyam Hailu, Gerardo Heiss, Brenna M. Henn, Lucia A. Hindorff, Rebecca D. Jackson, Cecelia A. Laurie, Cathy C. Laurie, Yuqing Li, Dan-Yu Lin, Andres Moreno-Estrada, Girish Nadkarni, Paul J. Norman, Loreall C. Pooler, Alexander P. Reiner, Jane Romm, Chiara Sabatti, Karla Sandoval, Xin Sheng, Eli A. Stahl, Daniel O. Stram, Timothy A. Thornton, Christina L. Wassel, Lynne R. Wilkens, Cheryl A. Winkler, Sachi Yoneyama, Steven Buyske, Christopher A. Haiman, Charles Kooperberg, Loic Le Marchand, Ruth J. F. Loos, Tara C. Matise, Kari E. North, Ulrike Peters, Eimear E. Kenny, and Christopher S. Carlson. Genetic analyses of diverse populations improves discovery for complex traits. *Nature*, 570(7762):514–518, June 2019. ISSN 0028-0836, 1476-4687. doi: 10.1038/s41586-019-1310-4. URL http://www.nature.com/articles/s41586-019-1310-4.

Sharon L I Wong and Maria B Sukkar. The SPARC protein: an overview of its role in lung cancer and pulmonary fibrosis and its potential role in chronic airways disease: SPARC in lung inflammation, remodelling and malignancy. *British Journal of Pharmacology*, 174(1):3–14, January 2017. ISSN 00071188. doi: 10.1111/bph.13653. URL http://doi.wiley.com/10.1111/bph.13653.

Andrew R. Wood, Marcus A. Tuke, Mike A. Nalls, Dena G. Hernandez, Stefania Bandinelli, Andrew B. Singleton, David Melzer, Luigi Ferrucci, Timothy M. Frayling, and Michael N. Weedon. Another explanation for apparent epistasis. *Nature*, 514:E3, October 2014. URL https://doi.org/10.1038/nature13691.

Naomi R. Wray, Cisca Wijmenga, Patrick F. Sullivan, Jian Yang, and Peter M. Visscher. Common Disease Is More Complex Than Implied by the Core Gene Omnigenic Model. *Cell*, 173(7):1573–1580, June 2018. ISSN 00928674. doi: 10.1016/j.cell.2018.05.051. URL https://linkinghub.elsevier.com/retrieve/pii/S0092867418307141.

Sewall Wright. *The roles of mutation, inbreeding, crossbreeding, and selection in evolution*, volume 1. na, 1932.

Paul R. WTCCC, David G. Clayton, Lon R. Cardon, Nick Craddock, Panos Deloukas, Audrey Duncanson, Dominic P. Kwiatkowski, Mark I. McCarthy, Willem H. Ouwehand, Nilesh J. Samani, John A. Todd, Peter Donnelly, Jeffrey C. Barrett, Paul R. Burton, Dan Davison, Peter Donnelly, Doug Easton, David Evans, Hin-Tak Leung, Jonathan L. Marchini, Andrew P. Morris, Chris C. A. Spencer, Martin D. Tobin, Lon R. Cardon, David G. Clayton, Antony P. Attwood, James P. Boorman, Barbara Cant, Ursula Everson, Judith M. Hussey, Jennifer D. Jolley, Alexandra S. Knight, Kerstin Koch, Elizabeth Meech, Sarah Nutland, Christopher V. Prowse, Helen E. Stevens, Niall C. Taylor, Graham R. Walters, Neil M. Walker, Nicholas A. Watkins, Thilo Winzer, John A. Todd, Willem H. Ouwehand, Richard W. Jones, Wendy L. McArdle, Susan M. Ring, David P. Strachan, Marcus Pembrey, Gerome Breen, David St Clair, Sian Caesar, Katherine Gordon-Smith, Lisa Jones, Christine Fraser, Elaine K. Green, Detelina Grozeva, Marian L. Hamshere, Peter A. Holmans, Ian R. Jones, George Kirov, Valentina Moskvina, Ivan Nikolov, Michael C. O'Donovan, Michael J. Owen, Nick Craddock, David A. Collier, Amanda Elkin, Anne Farmer, Richard Williamson, Peter McGuffin, Allan H. Young, I. Nicol Ferrier, Stephen G. Ball, Anthony J. Balmforth, Jennifer H. Barrett, D. Timothy Bishop, Mark M. Iles, Azhar Maqbool, Nadira Yuldasheva, Alistair S. Hall, Peter S. Braund, Paul R. Burton, Richard J. Dixon, Massimo Mangino, Suzanne Stevens, Martin D. Tobin, John R. Thompson, Nilesh J. Samani, Francesca Bredin, Mark Tremelling, Miles Parkes, Hazel Drummond, Charles W. Lees, Elaine R. Nimmo, Jack Satsangi, Sheila A. Fisher, Alastair Forbes, Cathryn M. Lewis, Clive M. Onnie, Natalie J. Prescott, Jeremy Sanderson, Christopher G. Mathew, Jamie Barbour, M. Khalid Mohiuddin, Catherine E. Todhunter, John C. Mansfield, Tariq Ahmad, Fraser R. Cummings, Derek P. Jewell, John Webster, Morris J. Brown, David G. Clayton, G. Mark Lathrop, John Connell, Anna Dominiczak, Nilesh J. Samani, Carolina A. Braga Marcano, Beverley Burke, Richard Dobson, Johannie Gungadoo, Kate L. Lee, Patricia B. Munroe, Stephen J. Newhouse, Abiodun Onipinla, Chris Wallace, Mingzhan Xue, Mark Caulfield, Martin Farrall, Anne Barton, , The Biologics in RA Genetics Genomics (BRAGGS), Ian N. Bruce, Hannah Donovan, Steve Eyre, Paul D. Gilbert, Samantha L. Hider, Anne M. Hinks, Sally L. John, Catherine Potter, Alan J. Silman, Deborah P. M. Symmons, Wendy Thomson, Jane Worthington, David G. Clayton, David B. Dunger, Sarah Nutland, Helen E. Stevens, Neil M. Walker, Barry Widmer, John A. Todd, Timothy M. Frayling, Rachel M. Freathy, Hana Lango, John R. B. Perry, Beverley M. Shields, Michael N. Weedon, Andrew T. Hattersley, Graham A. Hitman, Mark Walker, Kate S. Elliott, Christopher J. Groves, Cecilia M. Lindgren, Nigel W. Rayner, Nicholas J. Timpson, Eleftheria Zeggini, Mark I. McCarthy, Melanie Newport, Giorgio Sirugo, Emily Lyons, Fredrik Vannberg, Adrian V. S. Hill, Linda A. Bradbury, Claire Farrar, Jennifer J. Pointon, Paul Wordsworth, Matthew A. Brown, Jayne A. Franklyn, Joanne M. Heward, Matthew J. Simmonds, Stephen C. L. Gough, Sheila Seal, Breast Cancer Susceptibility Collaboration (UK), Michael R. Stratton, Nazneen Rahman, Maria Ban, An Goris, Stephen J. Sawcer, Alastair Compston, David Conway, Muminatou Jallow, Melanie Newport, Giorgio Sirugo, Kirk A. Rockett, Dominic P. Kwiatkowski, Suzannah J. Bumpstead, Amy Chaney, Kate Downes, Mohammed J. R. Ghori, Rhian Gwilliam, Sarah E. Hunt, Michael Inouye, Andrew Keniry, Emma King, Ralph McGinnis, Simon Potter, Rathi Ravindrarajah, Pamela Whittaker, Claire Widden, David Withers, Panos Deloukas, Hin-Tak Leung, Sarah Nutland, Helen E. Stevens, Neil M. Walker, John A. Todd, Doug

Easton, David G. Clayton, Paul R. Burton, Martin D. Tobin, Jeffrey C. Barrett, David Evans, Andrew P. Morris, Lon R. Cardon, Niall J. Cardin, Dan Davison, Teresa Ferreira, Joanne Pereira-Gale, Ingileif B. Hallgrimsdóttir, Bryan N. Howie, Jonathan L. Marchini, Chris C. A. Spencer, Zhan Su, Yik Ying Teo, Damjan Vukcevic, Peter Donnelly, David Bentley, Matthew A. Brown, Lon R. Cardon, Mark Caulfield, David G. Clayton, Alistair Compston, Nick Craddock, Panos Deloukas, Peter Donnelly, Martin Farrall, Stephen C. L. Gough, Alistair S. Hall, Andrew T. Hattersley, Adrian V. S. Hill, Dominic P. Kwiatkowski, Christopher G. Mathew, Mark I. McCarthy, Willem H. Ouwehand, Miles Parkes, Marcus Pembrey, Nazneen Rahman, Nilesh J. Samani, Michael R. Stratton, John A. Todd, and Jane Worthington. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145): 661–678, June 2007. ISSN 0028-0836, 1476-4687. doi: 10.1038/nature05911. URL http://www.nature.com/doifinder/10.1038/nature05911.

Ting Xie, Samina Akbar, Maria G. Stathopoulou, Thierry Oster, Christine Masson, Frances T. Yen, and Sophie Visvikis-Siest. Epistatic interaction of apolipoprotein E and lipolysis-stimulated lipoprotein receptor genetic variants is associated with Alzheimer's disease. *Neurobiology of Aging*, 69:292.e1–292.e5, September 2018. ISSN 01974580. doi: 10.1016/j.neurobiolaging.2018.04.013. URL https://linkinghub.elsevier.com/retrieve/pii/S0197458018301477.

ChangJiang Xu, Ioanna Tachmazidou, Klaudia Walter, Antonio Ciampi, Eleftheria Zeggini, Celia M. T. Greenwood, and the UK10K Consortium. Estimating Genome-Wide Significance for Whole-Genome Sequencing Studies. *Genetic Epidemiology*, 38(4): 281–290, April 2014. ISSN 0741-0395, 1098-2272. doi: 10.1002/gepi.21797. URL https://onlinelibrary.wiley.com/doi/10.1002/gepi.21797.

Yu Xu, Dragana Vuckovic, Scott C Ritchie, Parsa Akbari, Tao Jiang, Jason Grealey, Adam S. Butterworth, Willem H Ouwehand, David J Roberts, Emanuele Di Angelantonio, John Danesh, Nicole Soranzo, and Michael Inouye. Learning polygenic scores for human blood cell traits. preprint, Genetics, February 2020. URL http://biorxiv.org/lookup/doi/10.1101/2020.02.17.952788.

Mohsen Yaghoubi, Amin Adibi, Abdollah Safari, J. Mark FitzGerald, and Mohsen Sadatsafavi. The Projected Economic and Health Burden of Uncontrolled Asthma in the United States. *American Journal of Respiratory and Critical Care Medicine*, 200 (9):1102–1112, November 2019. ISSN 1073-449X, 1535-4970. doi: 10.1164/rccm.201901-0016OC. URL https://www.atsjournals.org/doi/10.1164/rccm.201901-0016OC.

Jian Yang, S. Hong Lee, Michael E. Goddard, and Peter M. Visscher. GCTA: A Tool for Genome-wide Complex Trait Analysis. *The American Journal of Human Genetics*, 88(1):76–82, January 2011. ISSN 00029297. doi: 10.1016/j.ajhg.2010.11.011. URL https://linkinghub.elsevier.com/retrieve/pii/S0002929710005987.

Jian Yang, Noah A Zaitlen, Michael E Goddard, Peter M Visscher, and Alkes L Price. Advantages and pitfalls in the application of mixed-model association methods. *Nature genetics*, 46(2):100–106, 2014.

Loic Yengo, Julia Sidorenko, Kathryn E Kemper, Zhili Zheng, Andrew R Wood, Michael N Weedon, Timothy M Frayling, Joel Hirschhorn, Jian Yang, Peter M Visscher, and the

GIANT Consortium. Meta-analysis of genome-wide association studies for height and body mass index in ~700000 individuals of European ancestry. *Human Molecular Genetics*, 27(20):3641–3649, October 2018. ISSN 0964-6906, 1460-2083. doi: 10.1093/hmg/ddy271. URL https://academic.oup.com/hmg/article/27/20/3641/5067845.

Michael K. Yu, Jianzhu Ma, Jasmin Fisher, Jason F. Kreisberg, Benjamin J. Raphael, and Trey Ideker. Visible Machine Learning for Biomedicine. *Cell*, 173(7):1562–1565, June 2018. ISSN 00928674. doi: 10.1016/j.cell.2018.05.056. URL https://linkinghub.elsevier.com/retrieve/pii/S0092867418307190.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.

Shuai Zhang, Lina Yao, Aixin Sun, and Yi Tay. Deep learning based recommender system: A survey and new perspectives. *ACM Computing Surveys (CSUR)*, 52(1):1–38, 2019.

Zhong-Qiu Zhao, Peng Zheng, Shou-tao Xu, and Xindong Wu. Object detection with deep learning: A review. *IEEE transactions on neural networks and learning systems*, 30 (11):3212–3232, 2019.

Grace X Y Zheng, Billy T Lau, Michael Schnall-Levin, Mirna Jarosz, John M Bell, Christopher M Hindson, Sofia Kyriazopoulou-Panagiotopoulou, Donald A Masquelier, Landon Merrill, Jessica M Terry, Patrice A Mudivarti, Paul W Wyatt, Rajiv Bharadwaj, Anthony J Makarewicz, Yuan Li, Phillip Belgrader, Andrew D Price, Adam J Lowe, Patrick Marks, Gerard M Vurens, Paul Hardenbol, Luz Montesclaros, Melissa Luo, Lawrence Greenfield, Alexander Wong, David E Birch, Steven W Short, Keith P Bjornson, Pranav Patel, Erik S Hopmans, Christina Wood, Sukhvinder Kaur, Glenn K Lockwood, David Stafford, Joshua P Delaney, Indira Wu, Heather S Ordonez, Susan M Grimes, Stephanie Greer, Josephine Y Lee, Kamila Belhocine, Kristina M Giorda, William H Heaton, Geoffrey P McDermott, Zachary W Bent, Francesca Meschi, Nikola O Kondov, Ryan Wilson, Jorge A Bernate, Shawn Gauby, Alex Kindwall, Clara Bermejo, Adrian N Fehr, Adrian Chan, Serge Saxonov, Kevin D Ness, Benjamin J Hindson, and Hanlee P Ji. Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. *Nature Biotechnology*, 34(3):303–311, March 2016. ISSN 1087-0156, 1546-1696. doi: 10.1038/nbt.3432. URL http://www.nature.com/articles/nbt.3432.

Jian Zhou and Olga G Troyanskaya. Predicting effects of noncoding variants with deep learning–based sequence model. *Nature Methods*, 12(10):931–934, October 2015. ISSN 1548-7091, 1548-7105. doi: 10.1038/nmeth.3547. URL http://www.nature.com/articles/nmeth.3547.

Jian Zhou, Chandra L. Theesfeld, Kevin Yao, Kathleen M. Chen, Aaron K. Wong, and Olga G. Troyanskaya. Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. *Nature Genetics*, 50(8):1171–1179, August 2018. ISSN 1061-4036, 1546-1718. doi: 10.1038/s41588-018-0160-6. URL http://www.nature.com/articles/s41588-018-0160-6.

Jian Zhou, Christopher Y Park, Chandra L Theesfeld, Aaron K Wong, Yuan Yuan, Claudia Scheckel, John J Fak, Julien Funk, Kevin Yao, Yoko Tajima, et al. Whole-genome deep-learning analysis identifies contribution of noncoding mutations to autism risk. *Nature genetics*, 51(6):973–980, 2019.

Zhaozhong Zhu, Xi Zhu, Cong-Lin Liu, Huwenbo Shi, Sipeng Shen, Yunqi Yang, Kohei Hasegawa, Carlos A. Camargo, and Liming Liang. Shared genetics of asthma and mental health disorders: a large-scale genome-wide cross-trait analysis. *European Respiratory Journal*, 54(6):1901507, December 2019. ISSN 0903-1936, 1399-3003. doi: 10.1183/13993003.01507-2019. URL http://erj.ersjournals.com/lookup/doi/10.1183/13993003.01507-2019.

Zhihong Zhu, Futao Zhang, Han Hu, Andrew Bakshi, Matthew R Robinson, Joseph E Powell, Grant W Montgomery, Michael E Goddard, Naomi R Wray, Peter M Visscher, and Jian Yang. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nature Genetics*, 48(5):481–487, May 2016. ISSN 1061-4036, 1546-1718. doi: 10.1038/ng.3538. URL http://www.nature.com/articles/ng.3538.

Appendix A

Simulation results supplementary

sample size	additive	2nd order	3rd order	4th order
10%	0.48	0.44	0.58	0.54
25%	0.24	0.23	0.30	0.46
50%	0.23	0.24	0.34	0.52
75%	0.23	0.21	0.32	0.61
100%	0.19	0.22	0.35	0.71

Table A.1 **Fractions of experiments where a non-linear solution was found by the NN out 100 simulations with a causal fraction of 0.5 of SNPs involved in statistical epistasis.** The values in the 'additive' column represent experiments where the ground truth genetic architecture was purely additive.

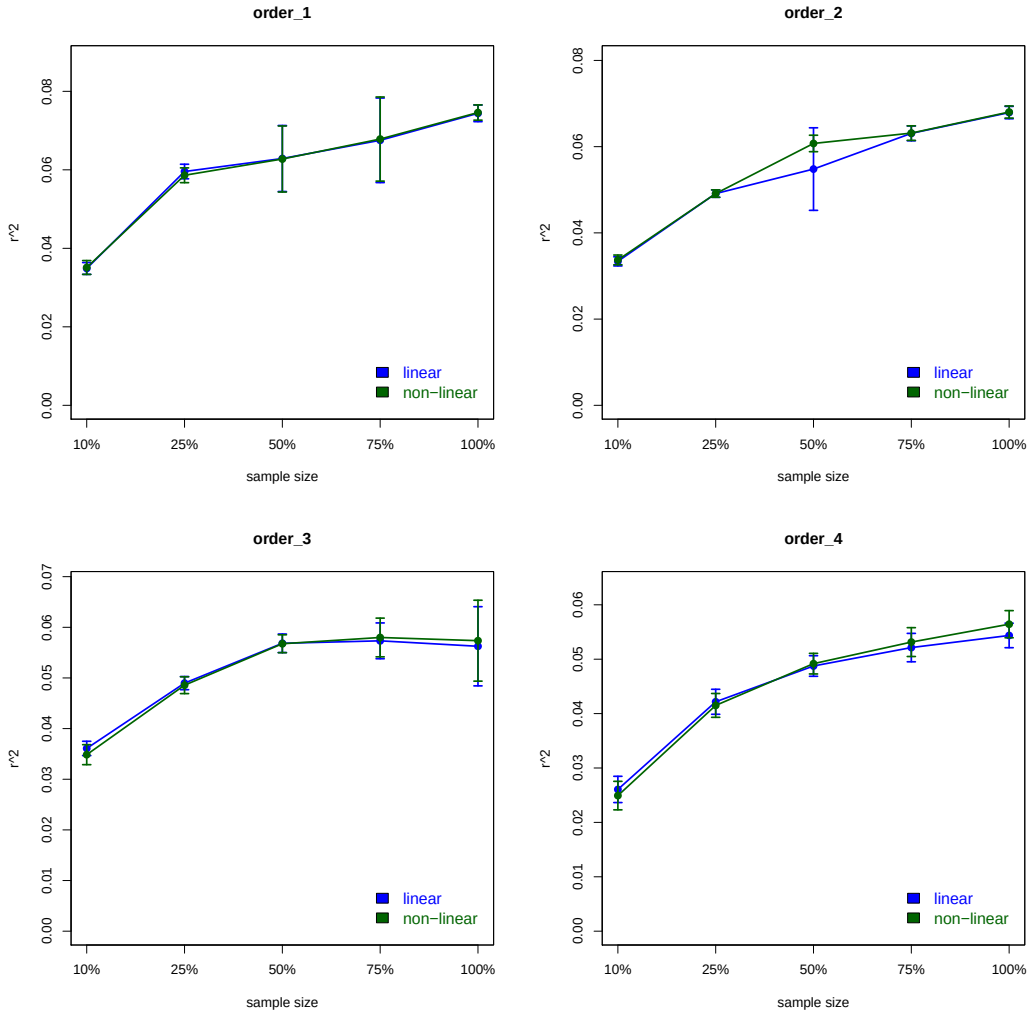


Fig. A.1 **NN performance on obtaining non-linear solutions under varying conditions for the experiments with a causal fraction of 0.5 of SNPs involved in statistical epistasis.** x-axis represents the % of sample size used and y-axis represents the r^2 of predicted vs observed phenotypes on the test set. Facets display experiments of genetic architectures that involve either additive, second, third and fourth-order interactions.

Appendix B

Neural-network supplementary

B.0.0.1 Convolutional neural-networks

The NN I described in Chapter one is what is known as a fully-connected NN, or FNN, which fit a hypothesis-free model that assumes that all input features are equally likely to interact with each other. However, for many data types, features that are spatially closer together in the input space are more likely to interact. Consider two intuitive examples: in most natural images the values of nearby pixels are more likely to form salient features like edges; similarly, nearby nucleotides are more likely to be part of the same regulatory element in a DNA sequence. Considering such local structures forms the basis of the convolutional neural-network (CNN) models.

The assumption of structure in the data is leveraged by CNNs via the introduction of a new layer type: the *convolution layer*. Here, neurons only consider a smaller local subset of the full input space which is termed the 'receptive field' of the neuron (a term which originated in the neuroscience literature (Hubel and Wiesel, 1962)). Instead of the hypothesis-free learning of FNN, in a CNN the neurons learn a vocabulary of features of a pre-defined size, known as 'kernels' or filters. As the term 'kernel' has many completely unrelated mathematical meanings, from here on, I will be using the term filter, to avoid confusion. Since these neurons no longer consider the entire input space; instead, they attempt to extract smaller reoccurring features, they need far fewer parameters to learn which results in greater power.

I will now move on to describe the details of the convolution operation itself. The convolution operation is a linear transformation where the filter is slid, or 'convolved', across the entire feature space which obtains a final output via element-wise multiplications. As my work involves one dimensional data (SNPs), I will illustrate this concept with 1D convolutions. Consider the following 1×2 size filter weight w and a 1×4 size $Input$:

$$w = \begin{bmatrix} w_1 & w_2 \end{bmatrix}, Input = \begin{bmatrix} I_1 & I_2 & I_3 & I_4 \end{bmatrix}. \tag{B.1}$$

Assuming a stride of one and no padding, the filter w may be applied to the $Input$ at three locations, or patches, to obtain the following $Output$:

$$Output = \begin{bmatrix} O_1 & O_2 & O_3 \end{bmatrix}, \tag{B.2}$$

where the entries in the $Output$ are defined by

$$O_1 = w_1 I_1 + w_2 I_2 \tag{B.3}$$

$$O_2 = w_1 I_2 + w_2 I_3 \tag{B.4}$$

$$O_3 = w_1 I_3 + w_2 I_4. \tag{B.5}$$

However, looping through the input space this way is inefficient, as high performance applications rely on massive parallelisation of computations via generalized matrix multiplications (Vasudevan et al., 2017). To facilitate this, the $Input$ is first transformed via an 'im2col' function that stretches the input out so that all possible patches are represented in a single matrix $\mathbf{L}$ as

$$\mathbf{L} = im2col(Input) = \begin{bmatrix} I_1 & I_2 & I_3 \\ I_2 & I_3 & I_4 \end{bmatrix}. \tag{B.6}$$

$\mathbf{L}$ may then be conveniently used in a single matrix multiplication to obtain a vector identical to B.2 by

$$Output = w\mathbf{L}. \tag{B.7}$$

To generate the entire output ($\mathbf{C}$) of a layer with d filters, w is replaced by a matrix representing all neuron weights ($\mathbf{W} \in \mathbb{R}^{d \times q}$) which modifies the above equation to

$$\mathbf{C} = \mathbf{W}\mathbf{L}. \tag{B.8}$$

It is notable, that in contrast to the fully-connected NN (eq 1.39), this weight matrix is now on the left hand side. This is because the layer has d neurons that are restricted to be able to only learn pre-defined filters of size q. The left multiplication by $\mathbf{W}$ also illustrates the parameter-saving attribute of the convolution layer, as the number of parameters to be learned ($d \times q$) no longer depends on the number of features in the $Input$. This allows CNNs to surmount high-dimensional data, such as high-resolution images or long DNA sequence reads, which would be beyond the reach of FNNs. Equation B.8 obtains the output $\mathbf{C} \in \mathbb{R}^{d \times (3*n)}$, where all individual observations are flattened to be stored along one dimension. To clarify, this

would mean that the transformed genotype observations for n individuals are concatenated into one dimension. Therefore, to connect the output of this layer to the flow of the rest of the NN function, the matrix $\mathbf{C}$ needs to be reshaped and transposed so that each of the n individuals stay on the rows as

$$\mathbf{C}' = Vec^{-1}(\mathbf{C})^T, \tag{B.9}$$

where Vec^{-1} denotes the reshaping operation.

In summary, the convolutional layer's function may be described as the extraction of smaller subsets from the input space. These reusable features are then passed forward as inputs from which subsequent layers learn higher-order representations, which result provides an explanation why it is a common CNN architectural trait that shallower convolution layers have fewer filters and deeper ones have more. Shallower layers' filters learn lower-order features (such as edges in an image or short motifs in a DNA sequence), and deeper layers' filters learn higher-order features made up from the shallower layers' representations. This is in contrast with fully-connected layers which tend start wide and each subsequent layer narrows towards the output.

As a side note, my description so far was a simplified explanation of how convolution layers generate an output, as in most practical applications there is an extra dimension to be considered. These would represent either the three colour channels for images, or one of the four nucleotides in the case of DNA sequence data. The equations would then change to involve tensors instead of 2D arrays, but otherwise would remain identical.

After the aforementioned convolution operation, a subsampling step is commonly used as the dimensionality of the output would increase by a factor of $d * Q/p$, where Q is the number of patches (three in the example) and p is the size of the input. To manage the dimensionality, and also to make the layer less sensitive to a small local changes, either another convolution layer is used with a larger stride (Springenberg et al., 2014), or a so called '*pooling layer*' is applied that summarises the output of the a convolution (Weng et al., 1992).

A popular method that accomplishes the subsampling operation is the '*Max Pooling*' function which is applied by taking the maximum of each image patch. In the example that I described so far, this would be equivalent to $MP = max(\sigma(\mathbf{C}))$, which would return the largest scalar value from the output after the activation by σ. In practice, pooling layers may have different sizes than the filters. The pooling size used most frequently is two which downsamples the output of each convolution layer by half. While 'Max Pooling' is primarily

used to down sample the activations, it is important to note that this also adds non-linearity as the $max()$ function depends on more than one value.

In conclusion, a single convolution layer may be added into the network I described in eq 1.39 by

$$Y = \sigma_k(\dots\sigma_2(\sigma_1(\mathbf{C'W_1})\mathbf{W_2})\dots W_k), \tag{B.10}$$

where $\mathbf{C'}$ is the output of the last convolution layer I derived in B.9.

To maintain clarity of the overall model, a short-hand notation may be used that emphasises the layer-by-layer sequential transformations from the input towards the output. In this notation the model I described so far can be expressed as

$$NN : [In, C_1, FC_1, FC_2, \dots, FC_k, Out], \tag{B.11}$$

where C_1 is the convolution layer, FC are k fully connected-layers and in and Out are the input and output layers, respectively. The element-wise activation functions are also not shown, but are assumed to take place after each layer with trainable weights. The advantage of this format is that adding j convolution layers may then simply be expressed by

$$NN : \left[In, C_1, C_2, \dots, C_j, FC_1, FC_2, \dots, FC_k, Out\right]. \tag{B.12}$$