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:

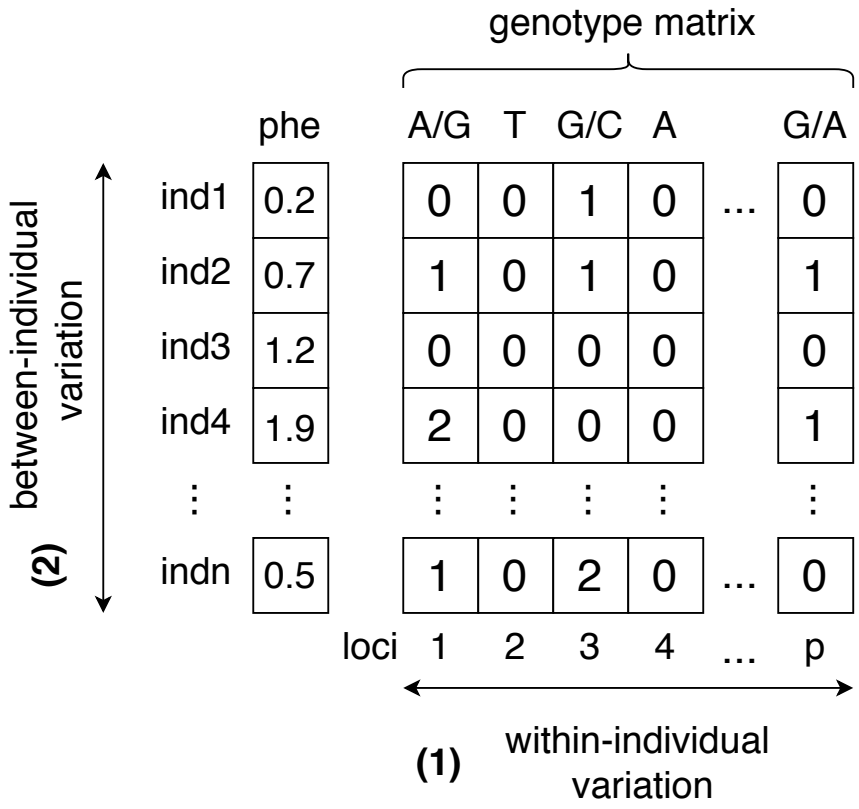

Fig. 1.1 **Hypothetical genotype matrix of n individuals at p loci.** Functional epistasis can take place between any of the p loci. However, statistical epistasis can only take place between loci 1, 3 and p, which are SNPs. The two loci, 2 and 4, do not vary in a population, therefore interactions between 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:

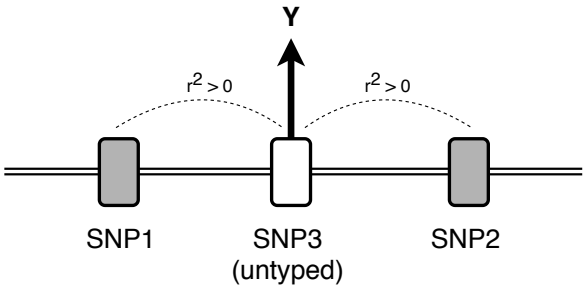

Fig. 1.2 **Illustration of the haplotype effect as an artefact generator for statistical epistasis.** The two SNPs imperfectly tag ($r^2 > 0$) the untyped SNP3, which is the causal variant affecting the phenotype Y. This pattern can arise even if the r^2 between SNP1 and SNP2 is zero. Statistical epistasis 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.