Chapter 2

Additive models and common quality-control steps

2.1 Chapter 2 outline

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

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

2.2 Datasets

2.2.1 Overview of the phenotypes considered

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

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

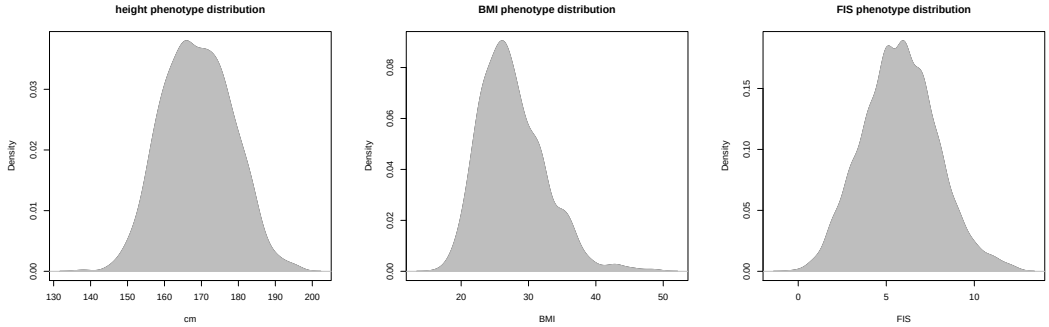
2.2.1.1 UK Biobank traits: height, BMI, fluid intelligence and asthma

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

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

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

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



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

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

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

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

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

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

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

2.2.1.2 IBD and its subphenotypes

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

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

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

2.2.2 UK Biobank genotype and phenotype data diagnostics

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

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

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

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

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

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

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

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

2.2.3 IBD datasets

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

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

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

2.3 Quality Control

2.3.1 Common quality control steps

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

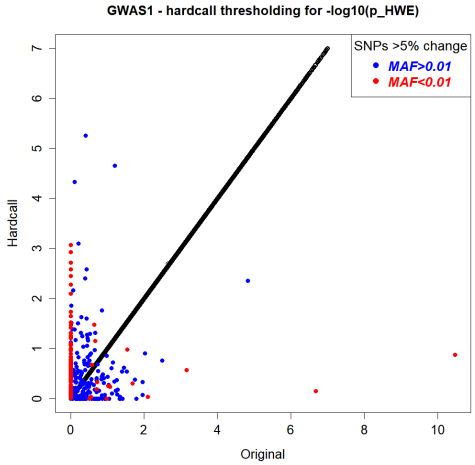
2.3.1.1 Converting genotype probabilities to hard calls

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

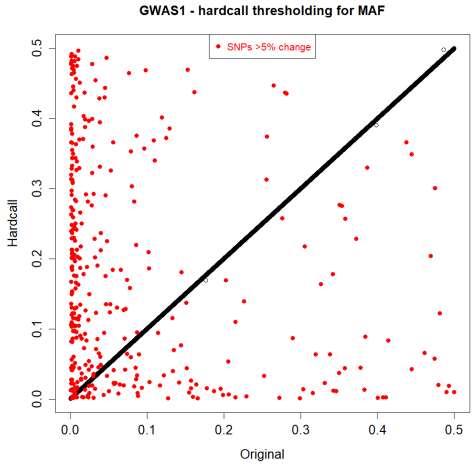
Converting genotype probabilities to hard calls is a lossy process that may result in substantial changes in allele frequencies for variants where allele dosages are uncertain. One possible option would have been to randomise the hard calling process to preserve the same allele frequencies that were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

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

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



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



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

2.3.1.2 Post-imputation quality-control for the UKBB genotypes

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

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

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

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

2.3.1.3 Post-imputation quality-control for the IBD genotypes

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

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

2.3.1.4 Phenotype quality control

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

2.4 Experimental setup for later analyses

2.4.0.1 Cohort organisation in the UK Biobank

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

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

2.4.0.2 Dataset organisation for the IBD datasets

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

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

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

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

2.5 Additive association tests

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

2.5.1 GWAS

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

2.5.1.1 Post-association QC

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

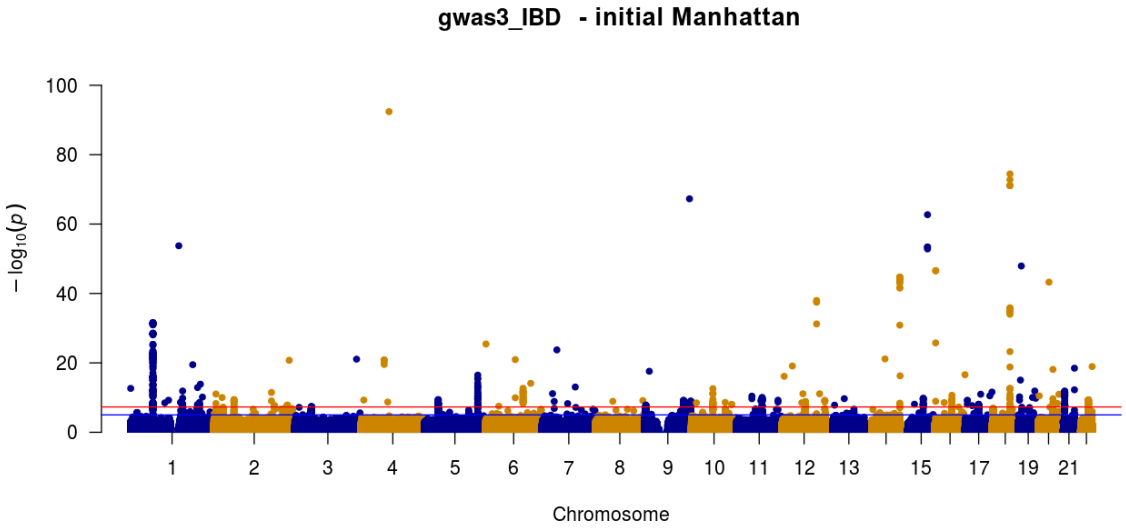


Fig. 2.3 **Manhattan plot visualising the GWAS1 study without applying post-association QC to consider LD patterns.** There are many associations above the genome-wide significance level with no LD structure to support them, a property that marks them out as potential false positives.

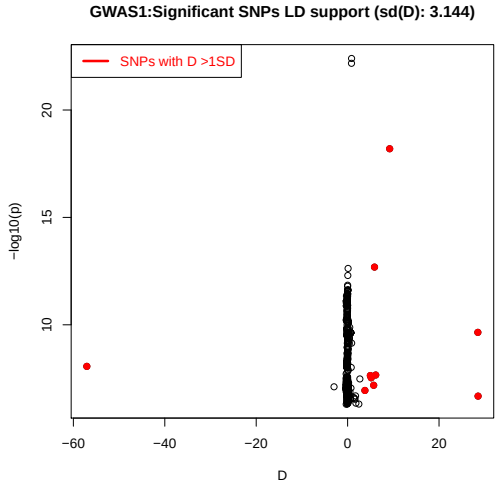
Traditionally, the quantitative allelic signals (intensity plots) of SNP associations suspected of being false positives are individually inspected for unexpected clustering patterns. In case of an imputed variant, several directly genotyped markers may be examined in the region. To make filtering for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

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

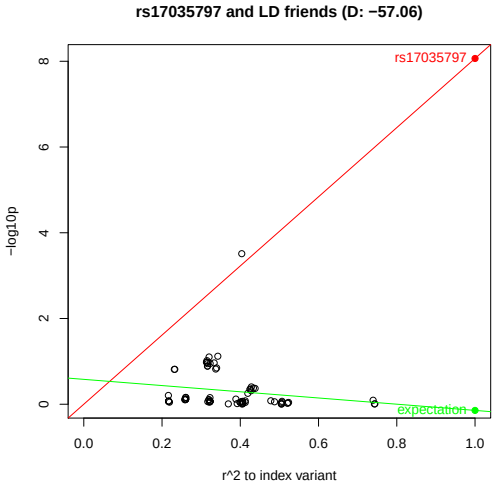
where r^2, β_{r^2} and e denote the LD to the target variant, its coefficient and the noise term, respectively. Next, using this model, I predicted the $-log10(p)$ of the target association using an r^2 of one (the target association's correlation squared with itself). Finally, I defined the value D,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

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

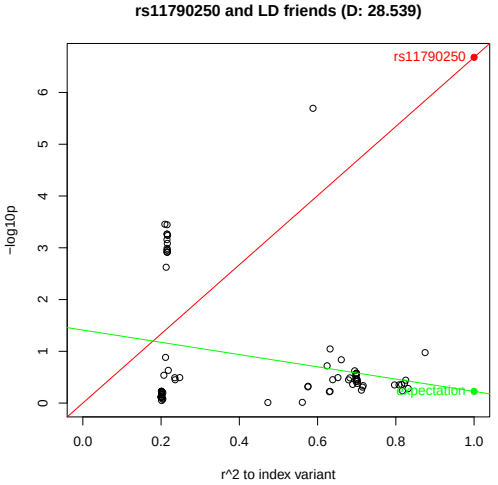
I note that D may be either negative or positive, depending on if the target SNP has higher or lower significance level than what would be expected by considering nearby variants. Upon examining the distribution of D and how it related to significance (Fig 2.4a), I set the exclusion criteria for markers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



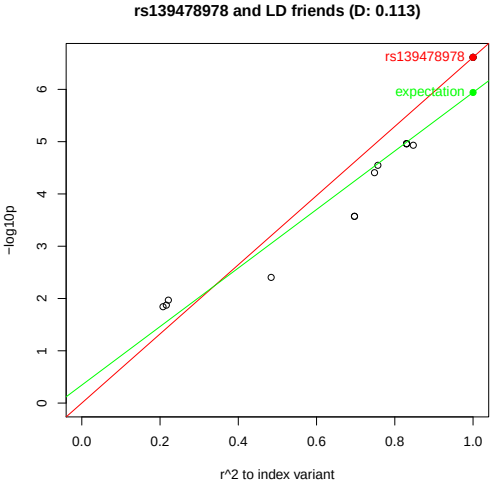
(a) **Demonstration of how the GWAS signal of SNPs in a study depends on D.** Outliers with an $SD(D) > 1$, highlighted in red, are the variants that were filtered out.



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



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

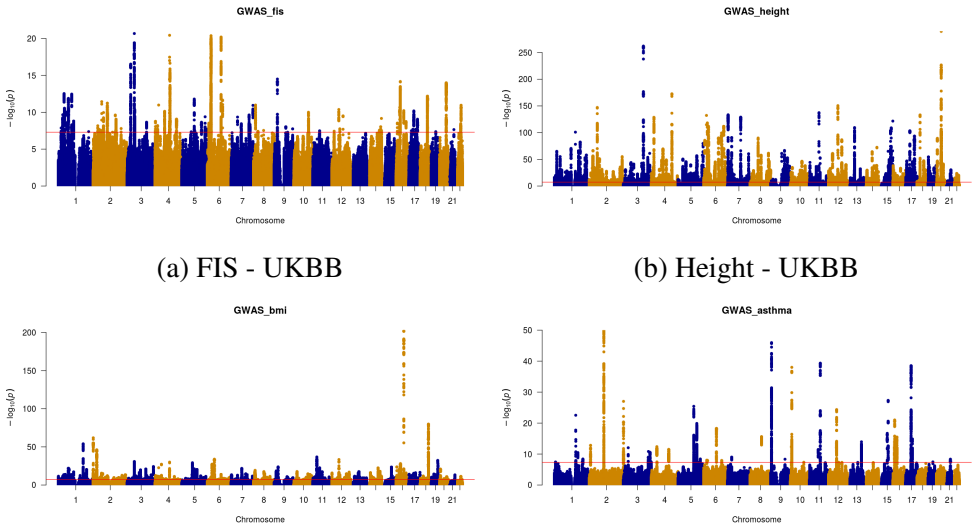


(d) **Example where the LD structure supports the variant as a genuine association.** The D value here is small, as the variant's predicted significance level is very close to the actual $-log10(p)$.

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

2.5.1.2 UKBB association test results

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



(a) FIS - UKBB

(b) Height - UKBB

(c) BMI - UKBB

(d) Asthma - UKBB

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

2.5.1.3 IBD association test results

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

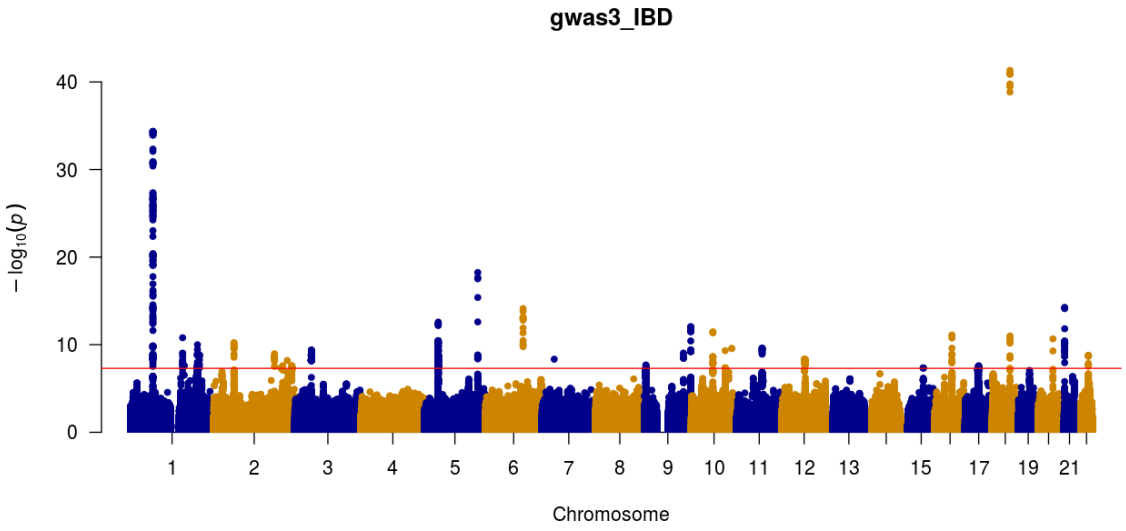
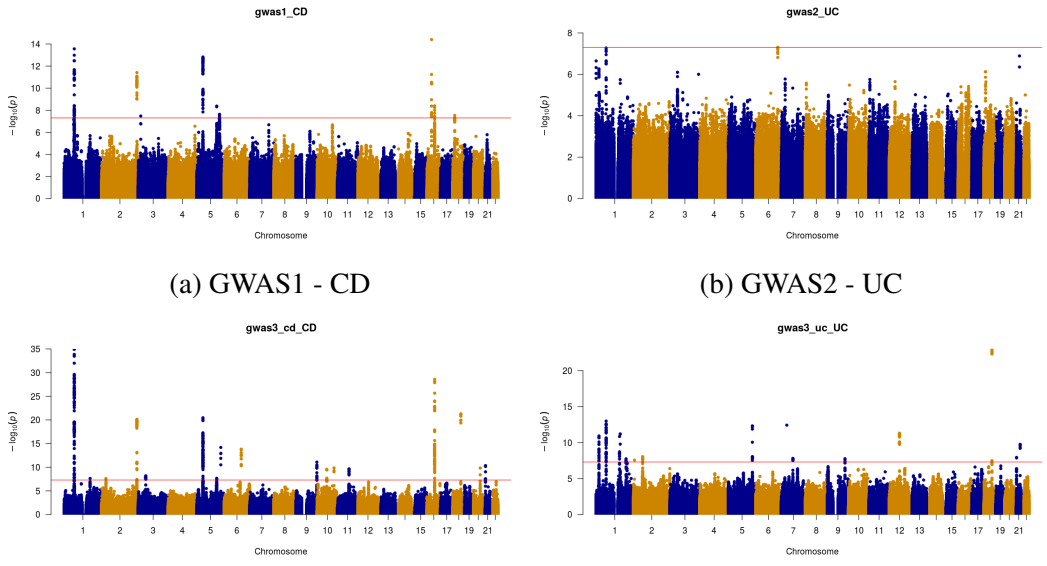


Fig. 2.6 **Manhattan plot visualising the GWAS3 dataset IBD association result.** y-axis represents the $-log10$ of the additive association p-values and the x-axis displays the genomic coordinates. The red line represents the genome-wide significance level of $5 * 10^{-8}$.



(a) GWAS1 - CD

(b) GWAS2 - UC

(c) GWAS3 - CD

(d) GWAS3 UC

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

2.5.2 Summary of the additive association experiments

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

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

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

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

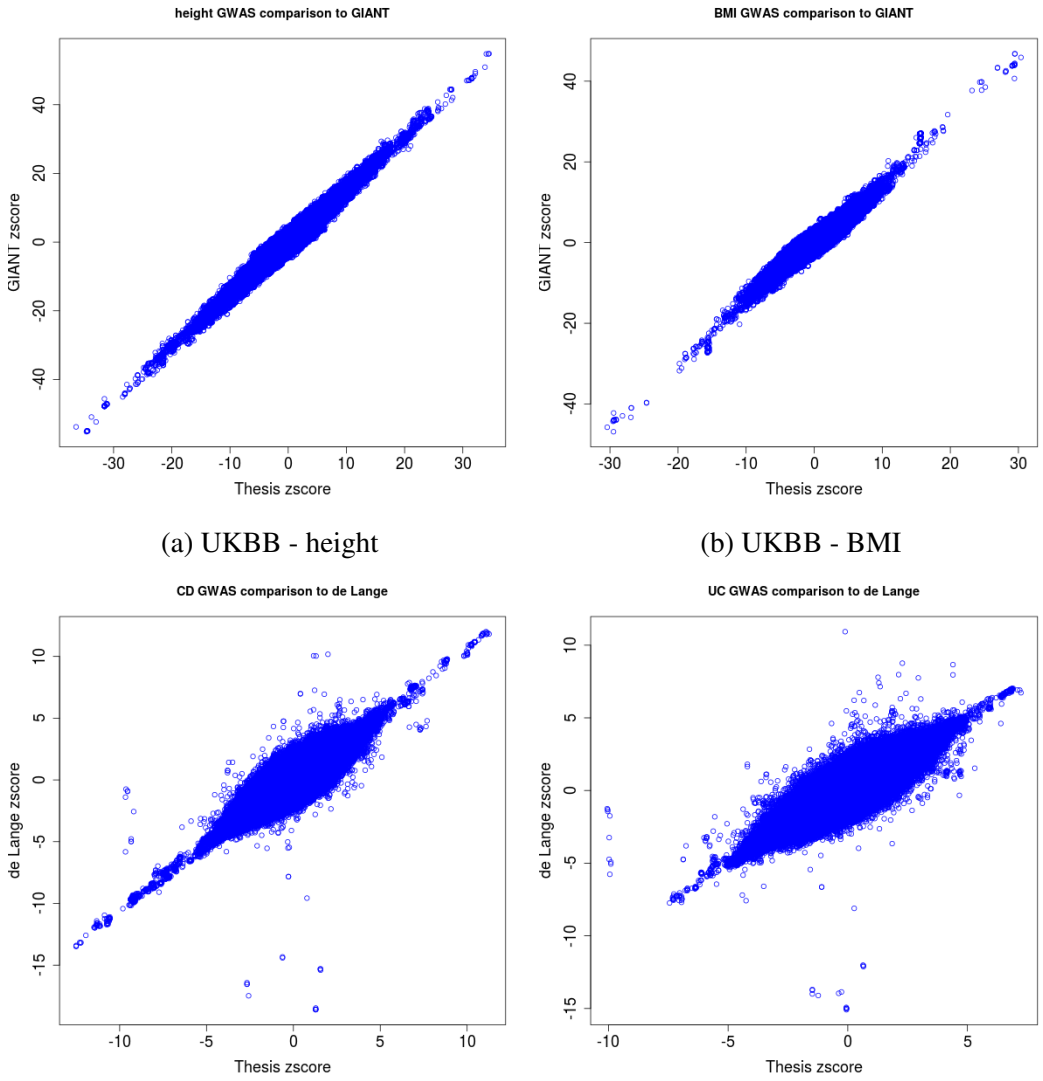
Table 2.6 **Landmark associations for my IBD analyses.** Comparisons of associations between the GWAS3 dataset and the study by de Lange et al. (2017). 'de Lange p' is the p-value from the de Lange et al. study, and 'chrom' indicates the chromosome.

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

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

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

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



(a) UKBB - height

(b) UKBB - BMI

(c) GWAS3 - CD

(d) GWAS3 UC

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

.

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

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

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

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

2.6.1 Establishing baselines

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

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

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

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

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

2.6.2 Estimating SNP heterogeneity of effect in the IBD studies

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

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

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

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

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

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

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

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

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

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

2.6.3 Finding the balance between the subphenotypes and IBD

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

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

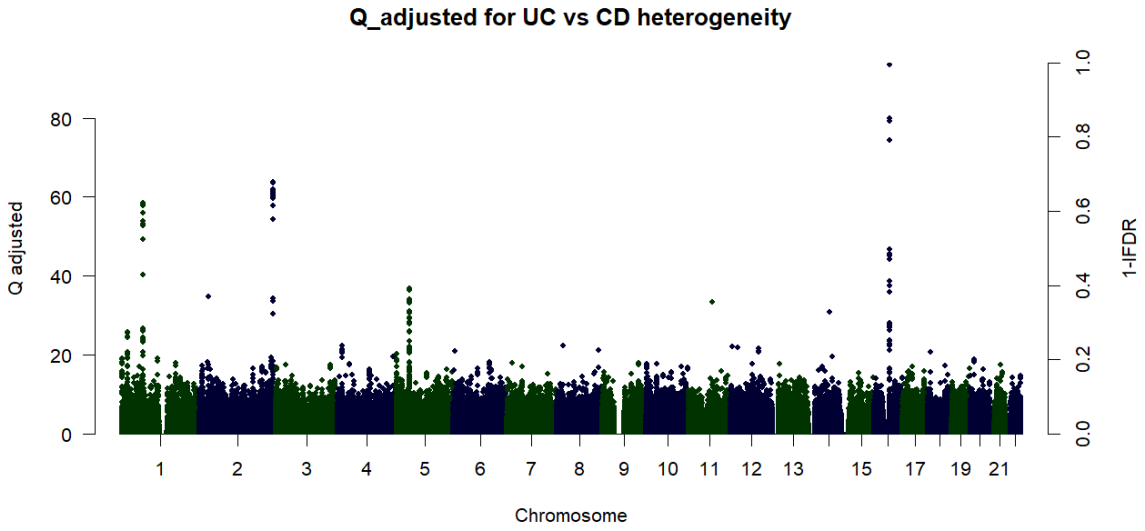


Fig. 2.9 **Manhattan visualising the adjusted Q values that measured SNP heterogeneity of effect between CD and UC.** Left y-axis shows the adjusted Q-values and right y-axis shows 1-lFDR. x-axis represents genomic coordinates.

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

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

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

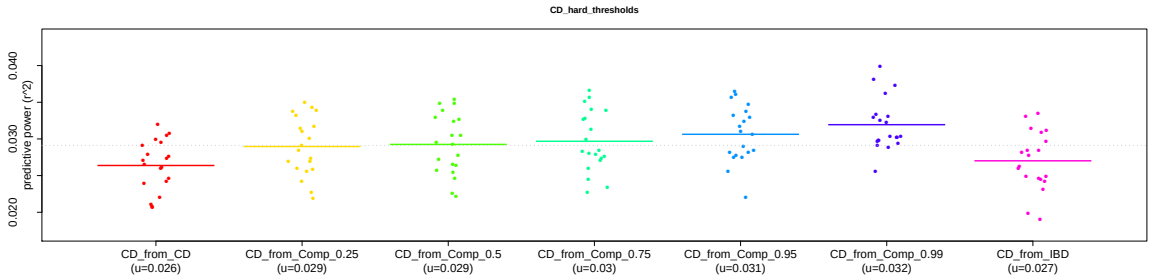
I implemented the composite PRS method by swapping the SNP summary statistics between the subphenotype and IBD as

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

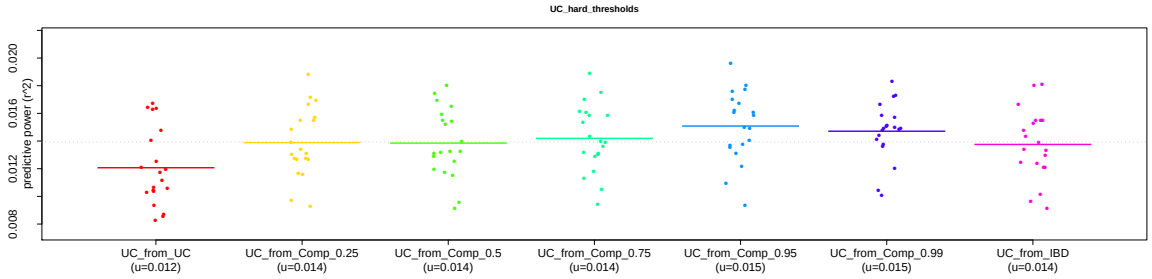
where $SS^i_{threshold}$ is a summary statistic associated with SNP_i for the current threshold that included β, SE, p and N (the number of individuals used to perform the association). I is an indicator function defined as

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

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



(a) CD composite PRS - hard thresholds



(b) UC composite PRS - hard thresholds

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

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

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

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

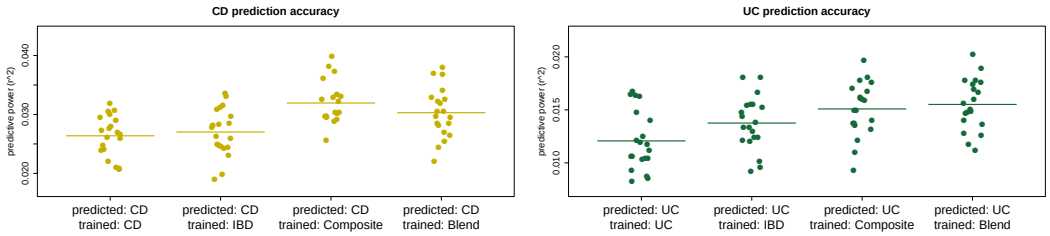
$$\begin{aligned} O &= (1 - lFDR)^2 SE^2_{subpheno} + SE^2_{IBD} lFDR^2 \\ SE_{blend} &= \sqrt{O + 2lFDR(1 - lFDR)SE_{subpheno}SE_{IBD} * cor(\beta_{subpheno}, \beta_{IBD})}. \end{aligned} \tag{2.12}$$

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

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

2.6.4 Results for predicting IBD subphenotypes

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



(a) CD final PRS plots

(b) UC final PRS plots

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

2.6.5 Discussion of the improved IBD subphenotype PRS

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

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

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

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

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

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