Chapter 3

Regression based models of statistical epistasis

3.1 Chapter 3 outline

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

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

3.2 Dimensionality reduction in the UKBB

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

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

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

3.2.1 Transcriptome and protein score data-sets

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

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

3.2.1.1 FIRM protein scores

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

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

3.2.1.2 BLUEPRINT transcriptome data

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

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

3.2.2 TWAS for asthma in the UKBB

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

3.2.2.1 Imputing the transcriptome

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

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

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

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

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

3.2.2.2 Expression association to the phenotype

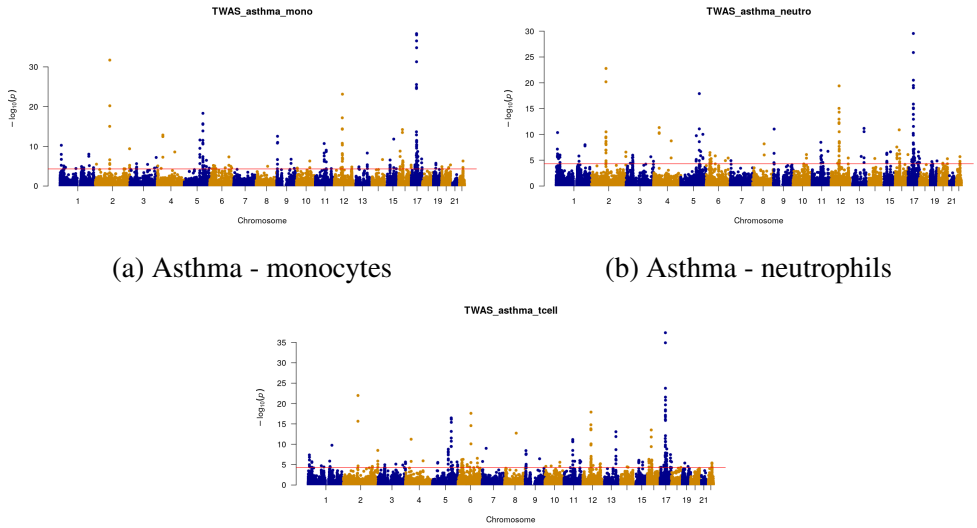
To perform the standard TWAS additive association test on the Main Set, I fit a simple univariate OLS linear model of the phenotype against each gene's predicted expression level as

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

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

3.2.2.3 UKBB asthma TWAS dimensionality reduction results

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



(a) Asthma - monocytes

(b) Asthma - neutrophils

(c) Asthma - T-cells

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

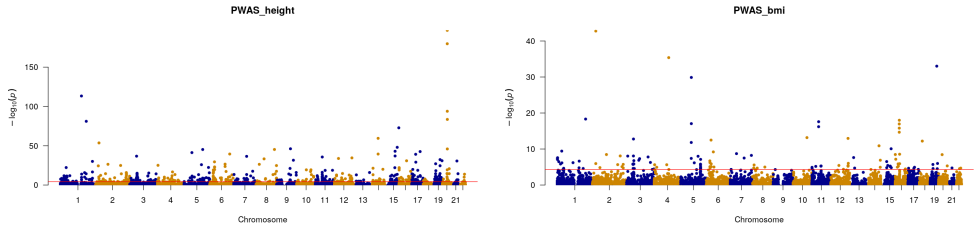
3.2.3 Protein burden score tests in the UKBB

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

3.2.3.1 Protein burden score dimensionality reduction results

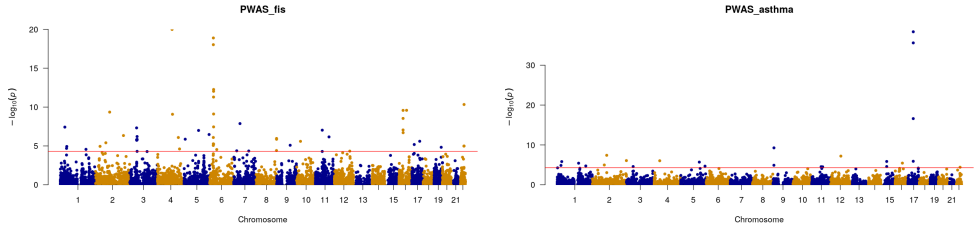
The protein burden test results for the four UKBB phenotypes are presented in Fig 3.2. The UKBB protein burden score test results also appear to be broadly congruent with their GWAS Manhattan counterparts, which reflect the fact that they were both derived from the same

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



(a) Height - PWAS

(b) BMI - PWAS



(c) FIS - PWAS

(d) Asthma - PWAS

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

3.2.4 Filtering the protein burden and gene expression scores

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

3.2.5 GWAS data

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

3.3 Interaction tests

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

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

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

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

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

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

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

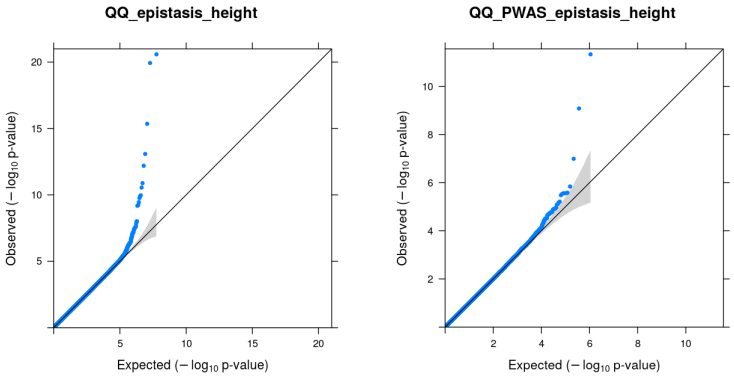
3.3.1 Post-association QC

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

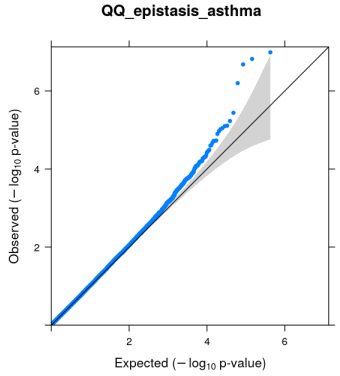
I examined the QQ-plots of my initial interaction tests (Fig 3.3) and I observed that only the height SNP/protein burden score tests and the asthma SNP tests appeared to deviate from the null. I decided to examine these two phenotypes in more detail to assess the potential for the aforementioned two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

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

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



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



(c) Asthma - SNP

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

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

3.3.2 Interaction test results

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

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

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

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

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

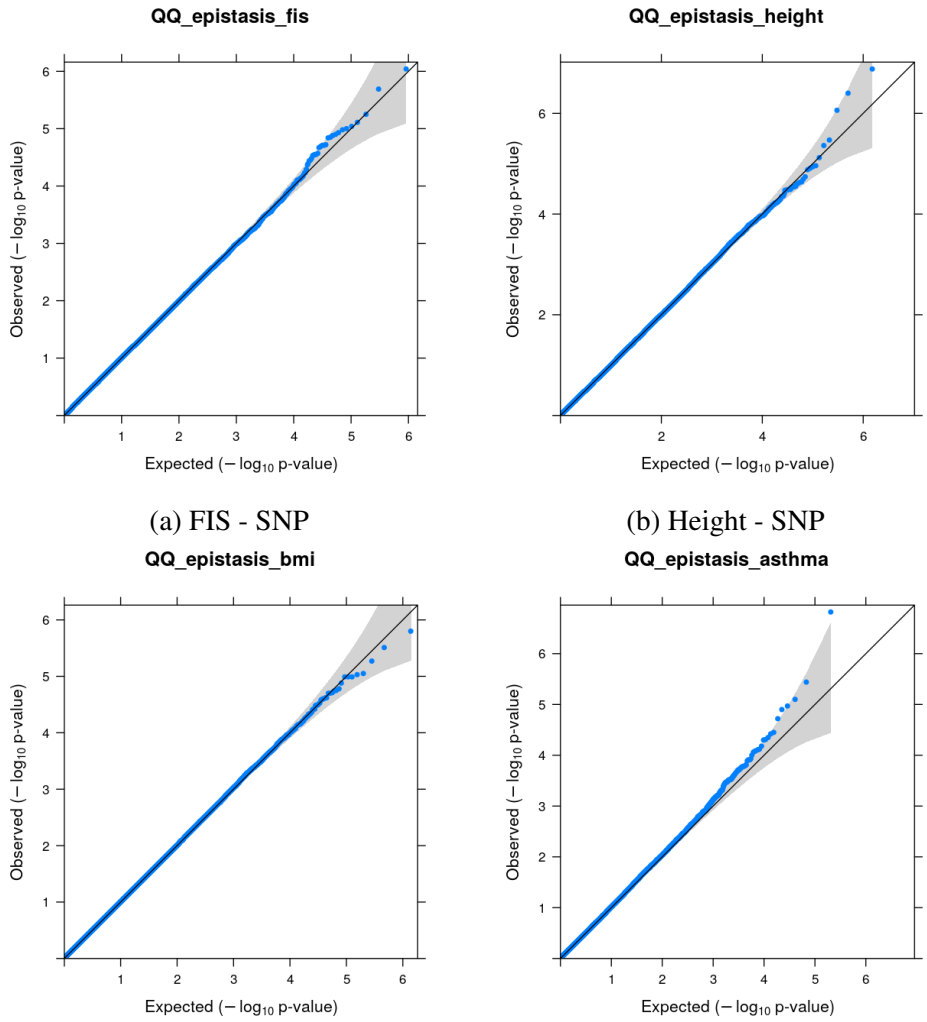
Tissue	**TWAS**	
	minimum FDR	number of tests
monocytes	0.734	34,716
neutrophils	0.422	23,653
T-cells	0.764	31,878

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

.

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

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



(a) FIS - SNP

(b) Height - SNP

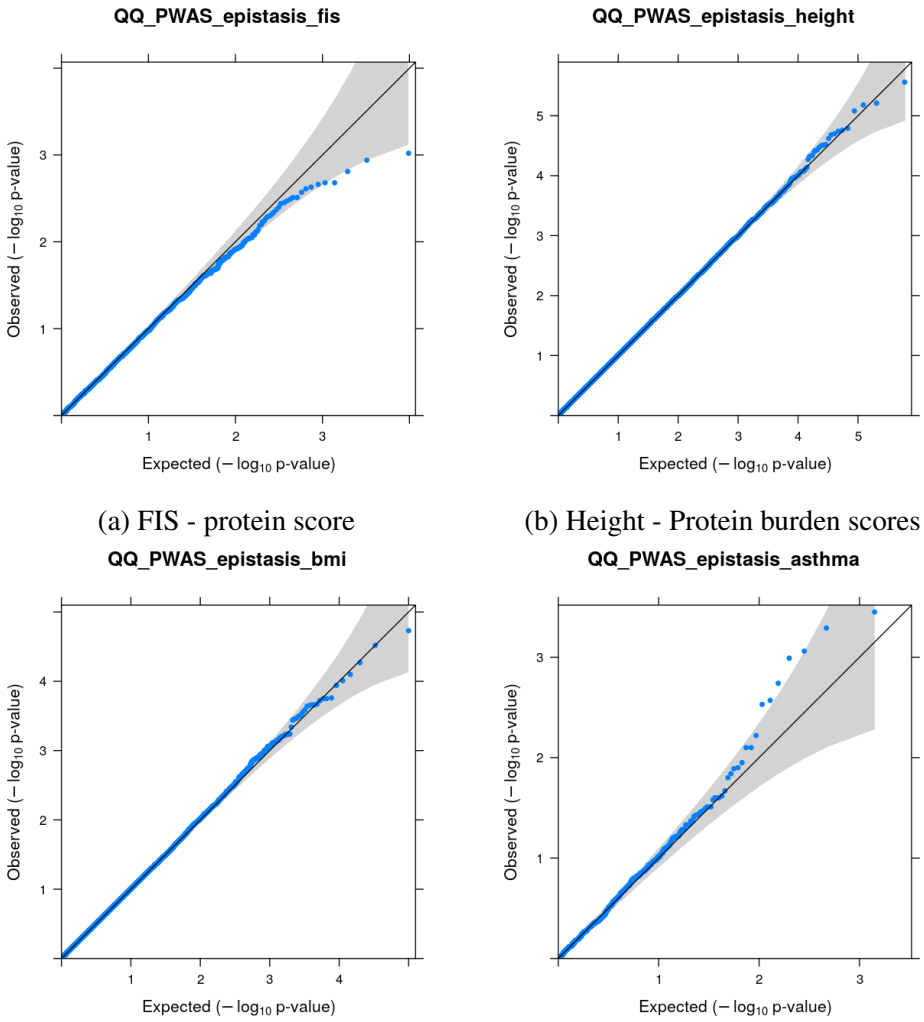
(c) BMI - SNP

(d) Asthma - SNP

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

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

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



(a) FIS - protein score

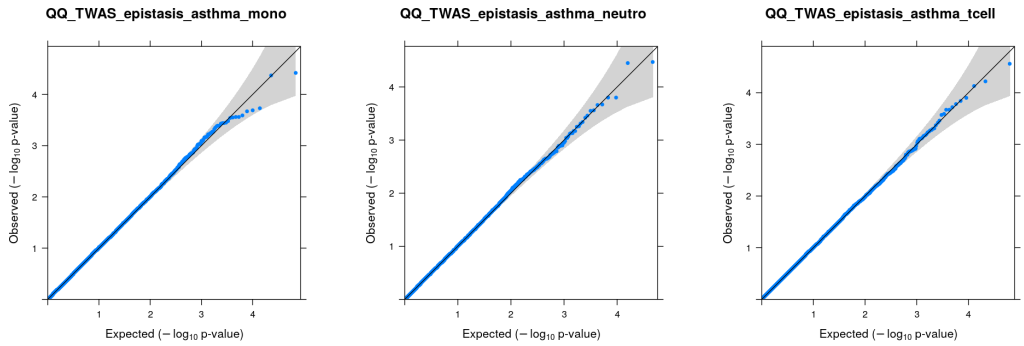
(b) Height - Protein burden scores

(c) BMI - protein score

(d) Asthma - protein score

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

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



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

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

Cases

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

Controls

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

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

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

3.4 Cross-domain interaction tests

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

3.4.1 Cross-domain filtering

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

3.4.1.1 Gene filter for asthma TWAS and protein burden scores

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

3.4.1.2 SNP-Gene cross-filtering

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

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

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

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

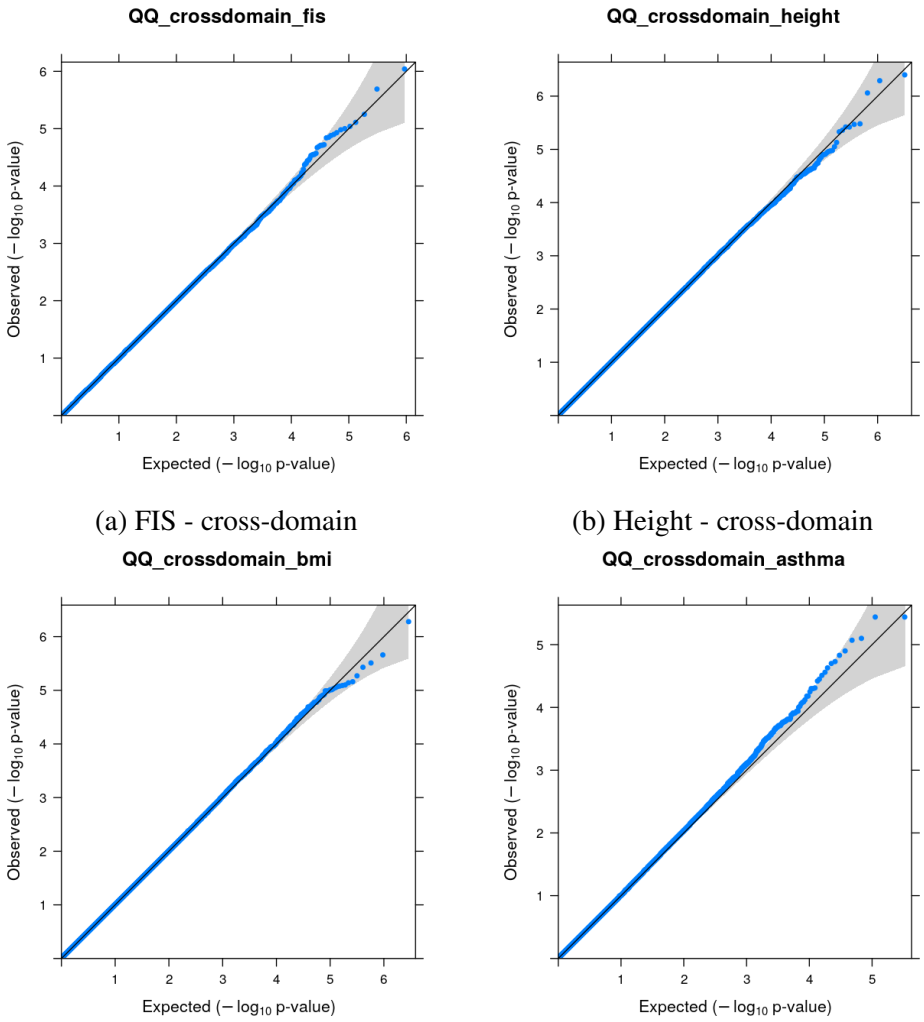
Table 3.7 **Summary of the cross-domain filtering process.**

3.4.2 Cross-domain interaction results

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

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

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



(a) FIS - cross-domain

(b) Height - cross-domain

(c) BMI - cross-domain

(d) Asthma - cross-domain

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

3.4.3 Summary of the UKBB interaction test experiments

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

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

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

3.5 Interaction tests in the IBD datasets

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

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

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

3.5.1 Biological insight to reduce search-space

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

3.5.2 Statistical haplotype phasing

3.5.2.1 The definition and the utility of haplotype phase

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

3.5.2.2 Overview of phasing methods

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

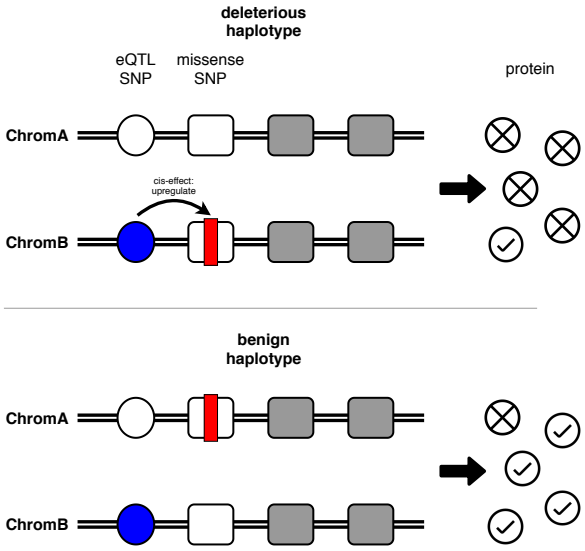


Fig. 3.8 **Missense-eQTL schematic diagram** Top: Illustration of the haplotype-specific interaction effect between a missense variant and a cis-regulatory SNP. In the deleterious haplotype configuration, the missense and the eQTL upregulatory alleles are on the same chromosome which results in an increase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

3.5.2.3 Statistical Methods

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

3.5.2.4 Trio and pedigree based phasing

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

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

3.5.2.5 Hidden markov model based phasing

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

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

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

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

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

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

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

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

3.5.2.6 Phasing summary

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

3.5.3 Genotype and summary data

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

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

3.5.3.1 Collating summary statistics for IBD

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

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

3.5.3.2 Obtaining haplotype configurations

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

3.5.4 Two statistical models to evaluate haplotype-specific interaction effects

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

3.5.4.1 '#Bad haplo' model

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

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

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

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

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

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

3.5.4.2 'haplo regression' model

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

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

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

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

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

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

3.5.4.3 Results for the haplotype-specific interaction tests

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

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

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

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

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

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

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

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

3.6 Concluding remarks

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

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

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

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

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