Evaluation and application of methodology for omic imputation into genome-wide association studies of complex human traits to infer potential causal mechanisms

James John Fryett

Thesis submitted for the degree of Doctor of Philosophy in the Faculty of Medical Sciences, Newcastle University

Population Health Sciences Institute, Newcastle University

August 2020

Abstract

To date, genome-wide association studies (GWAS) have been successful at identifying associations between common genetic variants and complex traits. However, little is known about the mechanisms by which trait-associated variants identified through GWAS affect the traits. One method developed to address this problem is the transcriptome-wide association study (TWAS), in which known relationships between genotypes and gene expression are leveraged to impute gene expression levels into GWAS samples. These imputed gene expression levels are then tested for association with traits to identify potentially causal trait-associated genes. Here, I investigated a number of ways of improving TWAS to enable the detection of more trait-associated genes, before extending the TWAS approach to investigate other omics measurements. First, to identify the best software for conducting TWAS, a range of packages were compared through application to data from the Geuvadis and Wellcome Trust Case Control Consortium projects. Overall, the investigated packages predicted gene expression with similar accuracy and detected similar expression-trait associations, although some tested a broader set of genes, so were preferable. Following this, the accuracy with which gene expression could be predicted from genotype data was investigated by comparing different statistical modelling approaches using data from the Geuvadis project. Overall, the expression of most genes could not be predicted accurately using any approach, but the best estimates were achieved when using approaches that assumed sparsity. Furthermore, prediction accuracy was improved by increasing sample size and by carefully matching training and testing data in terms of ancestry and tissue. Finally, the TWAS approach was extended to investigate the prediction of other omics measurements from genotype data. By generating prediction models for these omics measurements and applying these models to publicly available GWAS data, many associations between omics measurements and complex traits were detected, improving understanding of the mechanisms underlying GWAS risk loci.

Acknowledgements

I would like to thank Heather Cordell for conducting some of the initial quality control of the WTCCC1 genotype data and of the PBC cases genotype data. Furthermore, I would like to thank Heather for conducting the PBC genome-wide meta-analysis, the summary data of which are used in this thesis. I would also like to thank John Todd for providing access to the type 1 diabetes GWAS summary statistics that are used in this thesis.

This thesis makes use of data generated by the Wellcome Trust Case Control Consortium. A full list of the investigators who contributed to the generation of the data is available from www.wtccc.org.uk. Funding for the Wellcome Trust Case Control Consortium project was provided by the Wellcome Trust under award 076113.

This thesis also makes use of data from the Genotype-Tissue Expression (GTEx) project. The GTEx project was supported by the Common Fund of the Office of the Director of the National Institutes of Health, and by NCI, NHGRI, NHLBI, NIDA, NIMH and NINDS. The data used for the analyses described in this manuscript were obtained from the GTEx Portal.

This thesis also makes use of data from ALSPAC. I am extremely grateful to all the families who took part in this study, the midwives for their help in recruiting them and the whole ALSPAC team, which includes interviewers, computer and laboratory technicians, clerical workers, research scientists, volunteers, managers, receptionists and nurses.

This thesis also makes use of data from Understanding Society. Understanding Society is an initiative funded by the Economic and Social Research Council and various Government Departments, with scientific leadership by the Institute for Social and Economic Research, University of Essex, and survey delivery by NatCen Social Research and Kantar Public. The research data are distributed by the UK Data Service.

Table of contents

List of tables and figures

Chapter 2:

Chapter 3:

Chapter 7:

Chapter 1. Introduction

1.1 Interpreting the results of genome-wide association studies

Over the last 15 years, the genome-wide association study (GWAS) has become the gold-standard for investigating the relationship between common genetic variation at single nucleotide polymorphisms (SNPs) and complex disease. In this approach, individuals are genotyped at a set of SNPs across the genome, and at each SNP, genotypes are tested for association with a measured phenotype, typically using either linear or logistic regression (depending on the phenotype), to identify genomic risk regions for disease. This approach has been highly successful in identifying risk regions for many complex traits, and has even led to therapeutic advances in a number of diseases, such as the repurposing of drugs targeting the IL-23 pathway to treat psoriasis (Hueber *et al.*, 2010; Visscher *et al.*, 2017).

In the early days of GWAS, just performing the analysis was a major challenge. For example, consider the seminal Wellcome Trust Case Control Consortium 1 (WTCCC1) study (Wellcome Trust Case Control Consortium, 2007), which conducted case-control GWAS for 7 traits using 2000 cases for each disease and 3000 shared controls. As one of the first large-scale GWAS to be conducted, this study faced a number of challenges, including the design and development of custom genotyping arrays, the development of suitable quality control procedures and data analysis software to handle the large amounts of genotype data, and the gathering of enough samples to obtain sufficient statistical power to detect small genotypic effects on disease. However, in recent years, a number of developments have made GWAS easier to conduct. First, the availability of large population-based resources with matched genotype and deep phenotype data, such as UK Biobank (Bycroft *et al.*, 2018), has allowed recent GWAS to utilise extremely large sample sizes, providing sufficient statistical power to detect even the smallest SNP effects on disease. Second, the development of resources with high quality, high-density genotype data such as the 1000 Genomes Project (Auton *et al.*, 2015) and the Haplotype Reference Consortium (McCarthy *et al.*, 2016) has allowed for imputation of genotypes at millions of variants across the genome, making it easier to combine multiple GWAS together in a genome-wide meta-analysis, further increasing sample sizes. Third,

data analysis methods and software, as well as the computational power available for the data analysis, have rapidly improved in recent years, allowing complex analyses to be performed quickly. Together, these factors have allowed for well-powered GWAS to be rapidly performed for a wide variety of traits, resulting in the identification of many associations between genotype and phenotype. This can be seen in the EBI GWAS catalogue (which acts as a repository for GWAS results), where over 175,000 associations between genetic variants and complex traits are currently listed (Buniello *et al.*, 2019). However, despite this rapid improvement in the ability to detect associations through GWAS, comparatively little functional follow-up analysis of GWAS findings has been performed (Gallagher and Chen-Plotkin, 2018). This means that the mechanism by which the majority of genetic variants at GWAS risk loci affect their associated phenotype still remains unknown.

There are two main mechanisms by which genetic variation can act on a phenotype. The first, and perhaps most intuitive manner, is through a direct change to a protein. This could be the result of a genetic variant changing the DNA code in a coding sequence of a gene, introducing an amino acid substitution, insertion, deletion or frameshift, changing the amino acids that form the protein. Alternatively, a variant may introduce or remove a DNA splice site, resulting in an exon being included or skipped where it should not be, altering the protein produced. Regardless of the specific mechanism, if the change to the protein occurs in a region important in its function or structure, this could cause a gain or loss of protein function, resulting in a phenotypic change. This is the mechanism by which variants act in many Mendelian diseases (Amberger *et al.*, 2019), and is also thought to explain some associations between common genetic variation and complex diseases, especially those with large effect sizes, such as the effects of variants in HLA genes on autoimmune diseases (Jorgenson *et al.*, 2016; Darlay *et al.*, 2018).

The second mechanism by which genetic variation can affect a phenotype is through altering the regulation of gene expression, leading to an increase or decrease of the level of a gene's mRNA, and subsequently the levels of the corresponding protein, leading to a phenotypic change. Common genetic variation at SNPs has repeatedly been shown to affect expression of genes, especially those genes in close proximity to the SNPs in question (Lappalainen *et al.*, 2013; GTEx Consortium, 2015). These genetic loci at which SNP genotypes are associated with gene expression levels are termed expression quantitative trait loci (eQTLs). Most genetic variants that have

been identified as associated with a complex disease through GWAS are located in non-coding regions of the genome (Maurano *et al.*, 2012). Furthermore, it has also been shown that eQTLs, as well as other markers of the regulation of gene expression such as DNAse I hypersensitivity sites, are enriched near complex disease risk loci identified through GWAS (Nicolae *et al.*, 2010; Maurano *et al.*, 2012). Thus, it is thought that most common genetic variation that acts on complex diseases does so through this second mechanism, and so the integration of GWAS results with gene expression data has become a popular post-GWAS approach to improve the understanding of the mechanisms underlying GWAS findings.

A simplistic approach to the integration of GWAS data with gene expression data that was adopted in early GWAS was to search for overlaps between regions of the genome containing eQTLs, and regions of the genome containing SNPs associated with the phenotype of interest (identified through GWAS). While this approach can be a simple way of easily identifying some potential disease genes, it has a number of drawbacks. First, this approach lacks a formal statistical framework and does not provide information on the effect size and significance of gene expression on the phenotype of interest. Second, the overlaps between eQTL and GWAS signals that are identified with this method are not necessarily indicative of either a causal or pleiotropic relationship between the gene expression and phenotype, but may be induced by linkage disequilibrium (LD) between two separate, independent SNPs, one of which regulates gene expression and the other of which affects the phenotype. These seemingly LD-induced relationships between gene expression and phenotype that this approach is vulnerable to detecting are not suitable for therapeutic intervention, and so are of much less biological interest than true causal relationships between genotype, gene expression and disease. Third, the approach of searching for overlaps between eQTL and GWAS signals only makes use of one SNP at a time within any given genomic locus. However, studies have found that the expression of many genes is regulated by multiple independent SNPs (Battle *et al.*, 2017), and so by using only one SNP at a time, the effects of the other regulatory SNPs on the phenotype are missed. This means that this approach is likely to have less ability to detect potential disease genes than an approach that could make use of multiple SNPs simultaneously.

In an attempt to solve some of these problems, a number of more sophisticated methods of integrating gene expression data with GWAS data were developed. One

such family of methods approached the problem of integrating GWAS and gene expression data as a problem of colocalisation, attempting to test whether gene expression and the phenotype of interest are regulated by the same causal SNP. This family of methods includes the Bayesian colocalisation (Giambartolomei *et al.*, 2014), eCAVIAR (Hormozdiari *et al.*, 2016) and SHERLOCK (He *et al.*, 2013) software packages. These methods can identify instances where gene expression and the phenotype of interest are regulated by separate causal SNPs, and so should not be vulnerable to detecting LD-induced relationships between gene expression and phenotype, solving one of the problems of the approach discussed previously. However, these methods are still mostly limited by assumptions that there is only one causal SNP within a given locus, meaning that they will also have less power than a method that could make use of multiple SNPs at the same time. In addition, the relationships detected could represent pleiotropic effects (whereby a SNP acts independently on gene expression and on phenotype) rather than the arguably more interesting mechanism of a SNP acting on phenotype through altering gene expression.

Another family of methods developed to solve some of these problems was the transcriptome imputation, or transcriptome-wide association study (TWAS) method. These methods approach the problem of integration of gene expression data with GWAS data using a two-stage regression framework. Briefly, using genotype and expression data measured in the same set of individuals, gene expression is regressed on genotypes at multiple SNPs proximal to genes simultaneously to develop gene expression prediction models. These prediction models are then used to impute expression into GWAS samples, for which genotype and phenotype have been measured, but gene expression has not. Finally, the phenotype is regressed on the imputed gene expression values to test the association between imputed gene expression and phenotype. This family of methods includes the PrediXcan (Gamazon *et al.*, 2015), MetaXcan (Barbeira *et al.*, 2018), and FUSION (Gusev *et al.*, 2016) software packages.

These methods have a number of advantages over the simplistic approach of searching for overlaps between GWAS and eQTL signals. Firstly, these methods use a familiar regression framework that is well understood and produces estimates of the effect size and significance of the association of gene expression on phenotype. Second, by creating prediction models by regressing gene expression on genotypes

at multiple SNPs simultaneously, these methods allow for multiple SNPs to affect gene expression, potentially giving these approaches more power than the single-SNP approaches (in scenarios where multiple SNPs do indeed affect the expression of a gene). However, it should be noted that these approaches are not perfect and have their own drawbacks. For example, these approaches are still vulnerable to detecting LD-induced relationships between gene expression and phenotype, and so are not suitable to detect a sequence of causal relationships between genotype, gene expression and phenotype. However, connections between these approaches and two-stage least squares linear regression approaches for Mendelian Randomization (Burgess *et al.*, 2017) mean that, under certain assumptions, one could infer the existence of a causal relationship between gene expression and phenotype.

Given the relative advantages and the ease with which this method can be applied, and despite the drawbacks of the approach, it has become very popular. Since the approach was first proposed (Gamazon *et al.*, 2015), it has received over 150 citations in Pubmed, and has been widely used as a post-GWAS approach in many recent GWAS. This TWAS approach will be discussed in further detail throughout this thesis.

1.2 Transcriptome-wide association studies

The first step of TWAS is to fit a set of gene expression prediction models, which can later be used to impute gene expression levels into samples for which SNP genotype data and phenotype data have been measured (usually as part of a GWAS), but for which gene expression has not been measured.

However, prior to fitting gene expression prediction models, the set of genetic variants to use for predicting gene expression must first be determined. Attempting to fit a gene expression prediction model using all variants across the genome would not only be computationally intractable (especially when done for each of the tens of thousands of genes in the genome), but would also not reflect the current understanding of the mechanisms driving the genetic regulation of gene expression, as only a small proportion of genetic variants are thought to affect the expression of each gene (Wheeler *et al.*, 2016). As a result, gene expression prediction models are

not fitted using all variants across the genome but are instead fitted using only a subset of genetic variants. These variants are selected based on two main factors – their proximity to genes, and their type.

First, considering proximity – eQTL studies have consistently found that the SNPs nearest to the transcription start site (TSS) of a gene tend to have the strongest effects on its expression (Lappalainen *et al.*, 2013). For this reason, a proximity window is imposed around each gene, with only the SNPs within this window used to fit the gene expression prediction model. The exact size of the window used varies between different software packages that perform TWAS analysis. For example, prediction models from the PrediXcan package were fitted using SNPs within 1 megabase (Mb) of gene transcription start and end sites, while the models from the FUSION package were fitted using SNPs within 500 kilobases (Kb). This enables only the genetic variants with the strongest effects on gene expression to be used to fit the prediction models.

The second aspect of variant selection is selection based on variant type. Ultimately, the gene expression prediction models will be applied to GWAS data to impute the gene expression of GWAS samples. Typically, only common, biallelic SNPs are analysed in GWAS, with rare SNPs (those with a low minor allele frequency) and more complex variants such as insertions, deletions and structural variants excluded from analysis. Thus, to ensure compatibility between prediction models and GWAS data, prediction models are typically fitted using only common biallelic SNPs.

Having determined which variants to use for modelling, the next step is to fit the gene expression prediction models. In all current implementations of TWAS, it is assumed that the expected gene expression is a linear additive combination of weighted SNP genotypes (with the weights corresponding to the effect of a given SNP on gene expression). Thus, the gene expression prediction models take the form:

$$y_{ig} \sim \sum_{l=1}^{p} x_{il}\beta_{lg} + \varepsilon_{ig}$$

where y_{ig} is the expression of gene *g* in individual i, x_{il} is the effect allele count (0, 1 or 2) of SNP *l* in individual *i*, β_{lg} is the weight of SNP l on gene g, *p* is the total number of SNPs in the prediction model and ε_{ig} is an error term that includes all non-genetic effects on expression. This simple model is easily interpretable, and only requires

SNP genotypes to predict gene expression. However, it should be noted that a model such as this that only includes SNP genotypes as predictors of gene expression is unlikely to predict expression accurately, as it is known that a range of other non-genetic factors such as age (Yang *et al.*, 2015) and sex (Jansen *et al.*, 2014) influence gene expression levels. The issue of how accurately models of this form can predict gene expression will be discussed later.

These linear prediction models are fitted by regressing gene expression on the genotypes at the chosen set of SNPs. However, even after excluding distal, rare and complex genetics variants, thousands of genetic variants may remain. As most data sets used for prediction model fitting have only hundreds of samples, this results in a classic “large p, small n” problem, for which a standard ordinary least squares regression would be inappropriate (Iain and Titterington, 2009). Furthermore, both eQTL studies (GTEx Consortium, 2015) and more complex multivariate modelling efforts (Wheeler *et al.*, 2016) have consistently shown that gene expression is regulated by a few SNPs, each with a large effect size. This indicates that the genetic architecture of gene expression is likely to be sparse, rather than polygenic. As an ordinary least squares regression would use all chosen SNPs for prediction model fitting (so would be polygenic), it would also be inappropriate for this reason.

Instead, other methods that can overcome the problems described above are used to fit models. The specific method used to fit the gene expression prediction models differs between software packages. For example, prediction models from the PrediXcan software package were fitted using the elastic net (Zou and Hastie, 2005), a form of penalised regression that assumes that only a small proportion of SNPs actually affect gene expression. Similarly, some of the prediction models from the FUSION package were also fitted using elastic net, while others were fitted using a variety of methods including Least Absolute Shrinkage and Selection Operator (LASSO) (Tibshirani, 1996), another form of penalised regression similar to the elastic net, the Bayesian Sparse Linear Mixed Model (BSLMM) (Zhou *et al.*, 2013), another method that assumes sparsity, and the Best Linear Unbiased Predictor (BLUP) (de Los Campos *et al.*, 2013). While each of these methods make slightly different assumptions, they all fit linear models of the form shown above.

Although not strictly a necessary step in TWAS, the next step typically taken following prediction model fitting is to evaluate how accurately the prediction models

can predict gene expression from SNP genotypes. The prediction models that show poor predictive ability are typically discarded at this stage, while the remaining prediction models that can predict gene expression with sufficient accuracy (usually measured by correlating the predictions with measured expression measures) are then taken forward and applied to SNP genotype data from GWAS samples to impute their expression.

Finally, having imputed expression into the GWAS samples, the last step is to test for association between predicted expression values and the measured phenotype of interest. Depending on the structure of this phenotype, this is typically done using either linear or logistic regression. From this regression, the genes whose predicted expression is associated with the trait can be identified.

Many different software packages for TWAS have been developed, each broadly following the procedure described above but with slight methodological differences. The first publicly available TWAS software package was PrediXcan (Gamazon *et al.*, 2015). For this package, gene expression prediction models were fitted using the elastic net method, using data from the GTEx project (Battle *et al.*, 2017). Using this package, these prediction models are then applied to individual level genotype data from GWAS to impute expression levels, which are tested for association with the phenotype. This package is one of the most popular packages for TWAS, with over 100 PubMed citations since its release.

Shortly after PrediXcan, the FUSION package (Gusev *et al.*, 2016) was released. Similar to PrediXcan, this package also performs TWAS using prediction models trained with GTEx data. However, there are a number of methodological differences between FUSION and PrediXcan. The major difference is that FUSION uses GWAS summary statistics (estimates of effect sizes of SNP genotypes on phenotype, and the associated standard error) alongside gene expression prediction models to directly impute results of the test of association between predicted expression and phenotype, rather than imputing the gene expression and then separately testing its association with phenotype. Additionally, this package uses prediction models trained with different methods to elastic net, uses only SNPs within 500 Kbs of genes when fitting gene expression prediction models, and deals with missing data in a different way to PrediXcan, all of which could result in the FUSION package achieving

different results to PrediXcan. FUSION is also one of the most popular software packages for TWAS, with over 100 PubMed citations since its release.

Soon after FUSION was released, the developers of PrediXcan released a new version of their software named MetaXcan (later renamed S-PrediXcan) that, like FUSION, could conduct the test of association between predicted gene expression and phenotype by using GWAS summary statistics (instead of individual-level genotype and phenotype data) and pre-calculated gene expression prediction models (Barbeira *et al.*, 2018). This software uses the exact same prediction models as PrediXcan, although a number of assumptions were made to allow summary statistics to be used, which could lead to differences between PrediXcan and MetaXcan. While the manuscript describing the software was not published until 2018, an early version of the MetaXcan software was available online in early 2016 (along with an associated manuscript on BioRXiv). Like PrediXcan and FUSION, it is also a popular software package with over 75 citations on PubMed since its release.

Finally, another package that conducts TWAS analysis is the Summary Mendelian Randomisation (SMR) package (Zhu *et al.*, 2016). This method approaches TWAS analysis using Mendelian Randomisation principles. By using Mendelian Randomisation to combine the effect size of a SNP on gene expression (determined through eQTL studies) and the effect size of a SNP on phenotype (determined through GWAS), the effect size of gene expression on the phenotype of interest can be determined. This approach is analogous to a TWAS in which the gene expression prediction models each consist of only a single SNP. By using only a single SNP for analysis, this method is different to those methods described above, and so could achieve different results. Like the packages described above, this approach has also been popular, with over 100 citations on Pubmed since its release.

While each of these packages perform TWAS analysis, there are a number of slight methodological differences in the way the TWAS analysis is conducted. However, prior to the work conducted here, there had been no direct comparison on the software packages, so it was unclear which package would be better. In Chapter 3, I aim to address this problem by comparing PrediXcan, MetaXcan, FUSION and SMR by applying each of them to GWAS data from WTCCC1 studies of type 1 diabetes and Crohn’s disease, and comparing the results.

1.3 Prediction accuracy of gene expression prediction in TWAS

An important consideration in TWAS is the accuracy with which gene expression can be predicted from SNP genotypes. The power to detect association between predicted gene expression and a phenotype of interest in TWAS relies at least in part on the accuracy with which the gene expression can be predicted. This is analogous to the relationship between genotype imputation accuracy and statistical power in a GWAS of imputed genotypes, where lower genotype imputation accuracy is correlated with reduced power in the GWAS (Das *et al.*, 2018). Given this, gene expression prediction accuracy should be maximised to allow the detection of as many associations between gene expression and complex traits as possible.

Despite the importance of this issue, it has received relatively little attention in the literature. Prior to the development of the TWAS framework and the PrediXcan software package, (Manor and Segal, 2013) first investigated the prediction of gene expression from SNP genotypes using a number of methods, including elastic net and a K-nearest neighbours approach. Overall, a small number of genes with high prediction accuracy were identified, although the average prediction accuracy over all genes examined was poor, suggesting that the prediction accuracy for the expression of most genes was low. However, these estimates were achieved using a sample size of only 210 individuals, and the gene expression prediction models were fitted using only SNPs within 100 Kbs of genes, which is far smaller than the typical proximity window of 1 Mb used by the PrediXcan package. There may be a number of SNPs located further than 100 Kbs but closer than 1 Mb from genes that could affect gene expression, the inclusion of which in the gene expression prediction models would likely improve prediction accuracy. Thus, the estimates achieved by Manor and Segal may not be fully reflective of those achievable in current TWAS using larger sample sizes and proximity window sizes. Further investigation using more samples and a larger proximity window may give a better estimate of the levels of gene expression prediction accuracy achievable in TWAS.

The issue of prediction accuracy was later explored in the manuscript detailing the PrediXcan software package, in which the authors used a cross-validation procedure to test prediction accuracy. Again, while highlighting a number of genes with

seemingly high prediction accuracy estimates, the reported average prediction accuracy estimate achieved was poor, providing further evidence that the expression of most genes cannot be accurately predicted from only SNP genotypes. Moreover, these reported prediction accuracy estimates may have been inflated due to a methodological error. In (Gamazon *et al.*, 2015), the same cross-validation was used to tune the model parameters of the elastic net prediction model as was used to evaluate the prediction accuracy of the models, likely leading to an overestimation of model prediction accuracy and suggesting that the true prediction accuracy of gene expression may be lower than reported in (Gamazon *et al.*, 2015). Further investigation using a nested cross-validation approach (in which the prediction model tuning procedure is separated from the evaluation of prediction model accuracy) is required to obtain a true, unbiased reflection of the prediction accuracy achievable in TWAS.

Although the prediction accuracy estimates reported in the literature are poor, it is possible that some methodological changes to the TWAS procedure may help to improve these estimates. First, prediction accuracy may be improved by changing how parameter tuning is performed when fitting the prediction models. In (Gamazon *et al.*, 2015), elastic net was used to fit gene expression prediction models. The elastic net uses two tuning parameters – λ (a regularisation parameter) and α (which determines the sparsity of the fitted model). While cross-validation was performed to determine an appropriate value for λ for each prediction model, the value of α was set to 0.5 for all prediction models. The reason for this was unclear, as the authors did not demonstrate that the best prediction accuracy was achieved by setting α to 0.5. Instead, cross-validation could have been performed to select the value of α at which the maximum prediction accuracy was achieved, which would likely lead to an overall improvement in prediction accuracy estimates.

Another factor that could affect accuracy is the method used to train models. A number of different methods have been used for fitting gene expression prediction models, including elastic net (Zou and Hastie, 2005), LASSO (Tibshirani, 1996), BSLMM (Zhou *et al.*, 2013) and BLUP (de Los Campos *et al.*, 2013). To date, there has only been one systematic comparison of methods for prediction of gene expression. This study compared the ability of four approaches (elastic net, LASSO, BSLMM and linear mixed model) to predict gene expression from SNP genotypes, finding that the elastic net, LASSO and BSLMM all performed similarly, while all

outperforming the linear mixed model approach (Zeng *et al.*, 2017). This study did not include a number of other methods that have been shown to be successful in the prediction of other complex traits from SNP genotypes, such as ridge regression (Hoerl and Kennard, 2000) and Random Forests (Breiman, 2001; Sarkar *et al.*, 2015), which may also be successful at predicting gene expression. A more comprehensive comparison that includes these other methods may identify a better method to use when fitting gene expression prediction models.

An additional factor that could affect prediction accuracy estimates is the ancestry/population of the samples used to train and evaluate prediction models. The first set of gene expression prediction models from the PrediXcan package were trained using data from version 6 of GTEx, which contained samples with a range of ancestries, including European, African-American and Asian-American. All samples, regardless of ancestry, were used to generate the gene expression prediction models, and no attempt was made to account for population stratification, which is known to strongly affect genetic associations with complex traits (Serre *et al.*, 2008). However, the effect that this has specifically on the prediction of gene expression is unknown. Insight into the effect of ancestry on TWAS can be gained from the field of polygenic scores. Similarly to TWAS, in polygenic scores one data set is used to estimate the effect sizes of SNP genotypes on a phenotype of interest. These effect sizes are then combined into a score, which is then used to predict phenotypic values (typically disease risk) for a second data set. A number of studies have found that polygenic scores generated using effect sizes estimated using data with samples of one ancestry often perform more poorly when applied to samples of an alternative ancestry (Martin *et al.*, 2017; Duncan *et al.*, 2019; Martin *et al.*, 2019). Thus, it may be expected that gene expression prediction accuracy may suffer when the ancestry of the gene expression prediction model training and testing data sets are different.

Another factor that is likely to have an impact on the accuracy with which gene expression can be predicted from SNP genotypes is the sample size of the data set used to fit the gene expression prediction models. Studies examining the prediction of other complex traits from SNP genotypes have consistently shown the accuracy with which the phenotype can be predicted is directly related to the sample size of the data set used to estimate the SNP effect sizes on the phenotype (Dudbridge, 2013; Wei *et al.*, 2013; Wray *et al.*, 2013). Thus, it may be expected that increasing the sample size of the data set used to fit gene expression prediction models for

TWAS would lead to improved prediction accuracy. However, no formal investigation of this has been conducted to date, and so further study is required to understand this.

Finally, an issue more specific to the prediction of gene expression is prediction across tissues. It may not always be possible to measure gene expression in the tissue of interest as the tissue may be inaccessible. In these instances, gene expression prediction models for the tissue of interest may not be available. Instead, prediction models fitted using gene expression data gathered an alternative tissue may be used instead. The relevance of the results from these "proxy tissue" prediction models to the true tissue of interest depends on how accurately the prediction models fitted using the proxy tissue data can predict gene expression in the true tissue of interest. To date, there has been little study of how well prediction models fitted using data from one tissue are able to predict gene expression of another tissue. In the original manuscript describing PrediXcan, prediction models trained using gene expression data from whole blood measured in the DGN cohort were tested through application to data from nine GTEx tissues, with the best prediction accuracy achieved for GTEx whole blood, and poorer accuracy achieved for the other eight GTEx tissues (Gamazon *et al.*, 2015). However, this study only considered a limited number of tissues, and further insight may be gained by examining the broader range of tissues that are currently available.

In Chapter 4, I investigate some of the factors affecting the accuracy of gene expression prediction using data from Geuvadis. I perform a 10-fold nested cross-validation using a range of different statistical methods to identify which method is able to predict gene expression with the greatest accuracy, before focussing on how sample size, ancestry and tissue all affect how accurately gene expression can be predicted from SNP genotypes.

1.4 Predicting omics other than gene expression

To date, the TWAS methodology has mainly been used to investigate the role of gene expression in complex traits, although the approach could theoretically be used to investigate the role that any intermediate trait on the causal pathway from genotype to phenotype plays, as long as that intermediate trait is under genetic

regulation. Indeed, the TWAS approach has been adapted to investigate how DNA splicing (Gusev *et al.*, 2019) and endophenotypes derived from brain imaging (Xu *et al.*, 2017) act on complex phenotypes, showing the flexibility of the TWAS approach.

One intermediate trait of particular interest is DNA methylation, in which a methyl (-CH3) group is added or removed from cytosine residues at cytosine-guanine (CpG) dinucleotides. CpG methylation is especially interesting for a number of reasons. First, aberrant methylation at CpG sites has been implicated as a potential disease mechanism in a number of complex diseases (Dhana *et al.*, 2018; Story Jovanova *et al.*, 2018; Xu *et al.*, 2018), and so may act as a potential mechanism in other complex traits. Second, CpG methylation is known to regulate the expression of genes proximal to CpG sites. The most well-known mechanism by which DNA methylation regulates gene expression is through promoter methylation, with increased methylation of CpG sites within promoter regions typically associated with a decrease in gene expression. However, there are additional mechanisms by which CpG methylation acts on gene expression, although these are more complex, and are often time- and context- dependent (Luo *et al.*, 2018).

Similarly to gene expression, DNA methylation is known to be under genetic regulation. Twin and family-based studies have identified a significant heritable component of CpG methylation, with estimates of heritability ranging from 16% to 20% (Bell *et al.*, 2012; Grundberg *et al.*, 2013; van Dongen *et al.*, 2016; Hannon *et al.*, 2018). Studies estimating CpG methylation heritability using SNPs have also found a significant heritable component, although estimates vary depending on which SNPs are used. For example, a large study using SNPs across the whole genome found an estimate of 19% (van Dongen *et al.*, 2016), similar to estimates from twin studies, whereas studies focussing on heritability attributable to SNPs proximal to CpG sites generated smaller estimates (Quon *et al.*, 2013; Rowlatt *et al.*, 2016). In addition, studies have consistently identified relationships between CpG methylation and genotypes at individual SNPs, termed methylation quantitative trait loci (mQTLs) (Gaunt *et al.*, 2016; Richardson *et al.*, 2016; Volkov *et al.*, 2016). The presence of these mQTLs and the non-zero heritability estimates of CpG methylation indicate that it may be possible to predict methylation from SNP genotypes. Thus, training models that can predict CpG methylation from SNP genotypes and applying these prediction models to GWAS data in a TWAS-like framework may be a powerful method for identifying associations between CpG methylation and complex traits, which could

help improve understanding of the role that CpG methylation may play in complex traits.

Prior to the investigation conducted in this thesis, only one study had examined how accurately CpG methylation could be predicted from SNP genotypes. This study used a linear mixed modelling approach to generate CpG methylation prediction models, before using these prediction models in a TWAS-like framework to identify associations between brain methylation and Parkinson's disease (Rawlik *et al.*, 2016). At the CpG sites examined, the authors found that methylation could be predicted accurately, with an average prediction accuracy of 0.27. However, the prediction accuracy was examined only at the set of ~1800 CpG sites which showed a heritability estimate that was significantly different from zero (at $p<0.05$), and so it would be expected that these CpG sites would show good prediction accuracy. Thus, this estimate of the average prediction accuracy is unlikely to be reflective of the true average of CpG methylation prediction accuracy that includes CpG sites with low heritability estimates. Further study using all CpG sites (regardless of heritability) would be required to determine this.

There may also be opportunities to improve on the prediction seen by Rawlik et al. First, the authors used a linear mixed modelling method to generate their CpG methylation prediction models. To date, there has been no comparison of different methods for the prediction of CpG methylation from proximal SNP genotypes. However, a comparison of methods for prediction of gene expression from proximal SNP genotypes showed that the linear mixed model performed less well than methods that made assumptions of sparsity, such as LASSO and elastic net (Zeng *et al.*, 2017). In addition to this, mQTL studies have often found that CpG sites are each regulated by genotypes at a small number of SNPs, each of which has a large effect size on CpG methylation, which is indicative of a sparse genetic architecture of CpG methylation (Gaunt *et al.*, 2016). A comparison of a number of methods for the prediction of CpG methylation from SNP genotypes would be useful, as it would identify the method at which maximum prediction accuracy could be achieved. Second, the data used to examine prediction accuracy in this study were small, both in terms of sample size and the number of CpG sites measured. This study used data from 150 samples to train prediction models. More recent data sets such as those from the Accessible Resource for Integrated Epigenomics Studies (ARIES) (Relton *et al.*, 2015) have measured methylation data in far more people, and so

using these larger sample sizes for prediction model training may improve the prediction accuracy estimates achieved. Additionally, this study used methylation data measured on the 27K chip, which was an early chip that measured methylation at ~27,000 CpG sites. More recently gathered data sets have used the 450K and EPIC chips to measure methylation at ~450,000 and ~850,000 CpG sites respectively. Using these data sets, the prediction accuracy at more CpG sites across the genome may be examined, which could allow for the testing of more CpG sites in downstream analysis.

I attempt to address these issues in Chapter 5 by training and testing CpG methylation prediction models using a number of different methods. After identifying an optimal method for the prediction of CpG methylation from SNP genotypes, I then train and validate a set of CpG methylation prediction models using data from the ARIES and Understanding Society projects, which have much larger sample sizes and methylation measurements at many more CpG sites than data used in previous studies. I then follow up this analysis in Chapter 6, in which I conduct a methylome wide association study (MWAS) by applying these validated CpG methylation prediction models to publicly available GWAS summary statistics for 30 complex traits. In addition to the MWAS, I also conduct a TWAS, and use colocalisation and Mendelian Randomisation methods to integrate the results of the MWAS and TWAS to identify potentially causal relationships between CpG methylation, gene expression and complex traits.

In addition to DNA methylation, there are a number of other intermediate traits on the causal pathway from genotype to phenotype. One such trait is protein levels. Following transcription of DNA to mRNA, mRNA is then translated to proteins, which then carry out the vast majority of key functions in the human body. As proteins are so vital in human biology, it is likely that a change in protein levels could directly lead to phenotypic change and could play a causal role in complex disease biology.

Like gene expression and DNA methylation, protein levels are known to be under genetic regulation. To date, a number of studies have identified SNPs that regulate the expression levels of proteins, both proximally and distally (Battle *et al.*, 2015; Sun *et al.*, 2018). Thus, it may be possible to use SNP genotypes to predict protein levels. However, to date there has been no previous study of the prediction of protein levels using SNP genotypes. Likewise, there has also been no prior application of protein

level prediction models to GWAS data in a TWAS-like framework. Thus, there is an opportunity to use the TWAS methodology to explore the role of protein levels in complex traits.

I address this in Chapter 7, in which I use a nested cross-validation approach to investigate how accurately serum protein levels can be predicted from SNP genotypes using data from two sources – the INTERVAL study, and matched genotype and proteomics data from a group of primary biliary cholangitis (PBC) patients. Following this, I then conduct a proteome-wide association study (PWAS) (using protein level prediction models fitted using the INTERVAL and PBC patients data), an MWAS (using CpG methylation prediction models fitted using data from ARIES) and a TWAS (using gene expression prediction models fitted using data from GTEx), before using multi-trait colocalisation to integrate the results of these three approaches and to identify potentially causal relationships between PBC, gene expression, CpG methylation and serum protein levels.

Chapter 2. Methods and materials

This chapter will describe the data used throughout the thesis, and the quality control procedures applied to these data. Additionally, the methodology of transcriptome imputation and other methods used throughout this thesis will be described here. Methodology that is more specific to one specific chapter will be described in the relevant chapter.

2.1 Data used in the thesis

2.1.1 Wellcome Trust Case Control Consortium 1 (WTCCC1) GWAS data

The WTCCC1 data set is used in Chapter 3 for the comparison of transcriptome imputation software packages, as well as in Chapter 4 for a comparison of gene expression prediction models trained using different statistical methods. These data were used in (Gamazon *et al.*, 2015), so using these data here allowed comparison of the results with those in (Gamazon *et al.*, 2015). These data were well characterised, having been studied many times in the last 10 years, and had undergone considerable quality control, making them an ideal data set to use early in the project.

These data have been described in much detail in (Wellcome Trust Case Control Consortium, 2007). To summarise briefly, the WTCCC1 performed GWAS for seven common and complex conditions (bipolar disorder, Crohn's disease, coronary artery disease, hypertension, type 1 diabetes, type 2 diabetes and rheumatoid arthritis) prevalent in the UK. For each disease, approximately 2000 cases were recruited, in addition to 3000 samples from the UK Blood Service and the 1958 British Birth Cohort to be used as controls in the GWAS for each disease. All samples were genotyped at approximately 500,000 variants using the Affymetrix GeneChip 500k Mapping Array Set. Genotypes were called using CHIAMO and underwent standard genotype quality control procedures.

Only data for Crohn's disease (CD) and type 1 diabetes (T1D) were chosen for study in this thesis, as multiple significant associations were identified for each of these phenotypes in the original PrediXcan paper. Data for the remaining 5 complex

diseases were not used in this thesis. Initial quality control and removal of SNPs failing the WTCCC automated quality control procedures was done by Heather Cordell using PLINK, before PLINK genotype files for the 1,748 Crohn's disease cases, the 1963 type 1 diabetes cases and the 2938 shared controls were sent over to me for subsequent analysis.

Extensive quality control had been carried out a part of the original WTCCC1 study and by Heather Cordell (prior to the data being sent to me), so little extra quality control was required. To that end, SNPs with minor allele frequency (MAF)<0.01 and SNPs with abnormal genotyping cluster plots were removed. All remaining SNPs and samples were taken forward for imputation. For each disease, cases and controls were combined into a single VCF file and were imputed together, as was done in (Gamazon *et al.*, 2015). Imputation was done using the Michigan Imputation Server (Das *et al.*, 2016) using the 1000 Genomes Phase 1 reference panel (all ancestries) with ShapeIT used to phase genotypes. After downloading the imputed genotypes, insertions, deletions, SNPs with imputation quality $R^2 < 0.8$ and SNPs with MAF < 0.01 were removed.

2.1.2 Geuvadis data

Data from the Geuvadis project (Lappalainen *et al.*, 2013) are briefly used in Chapter 3 for a comparison of gene expression predicted using popular transcriptome imputation packages with measured gene expression. The data are used heavily in Chapter 4, where they are used for a comparison of gene expression prediction models trained using seven different statistical methods.

The Geuvadis project was designed to conduct early transcriptome sequencing via RNA-seq in individuals from a range of populations across Europe and Africa, most of whom had already been genotyped as part of the 1000 Genomes project. These data were used alongside genotype data to characterise the relationship between genotype and expression through identification of eQTLs across the genome. The relatively large samples size (compared to other reference panels available at the time such as GTEx), multi-ethnic nature of the data and the public availability of these data made them ideal for investigating factors affecting the prediction of gene expression from SNP genotypes.

The sample collection, genotyping and imputation procedures are described in detail in (Lappalainen *et al.*, 2013). Briefly, the project used data from 465 samples of Northern and Western European (CEU), Finnish (FIN), British (GBR), Tuscan (TSI) or Yoruba (YRI) ancestry. These samples had all been genotyped as part of either Phase 1 or Phase 2 of the 1000 Genomes project. 421 of the 465 Geuvadis samples were genotyped at approximately 38 million variants across the genome using whole genome sequencing, exome sequencing and SNP genotyping (through SNP chips) as part of Phase 1, while the remaining Geuvadis samples were genotyped on an Omni 2.5M array as part of Phase 2. The genotype data for the samples measured as part of Phase 2 of the 1000 Genomes project were then imputed using the entire 1000 Genomes Phase 1 data set as a reference panel. Measured genotypes for the 421 samples from Phase 1 and imputed genotypes for the samples from Phase 2 were then combined into a single set of 22 VCF files (one per chromosome).

After downloading these VCF files, some genotype quality control was conducted by myself. Three samples with genotype data but without gene expression data were removed, leaving a total of 462 samples. These samples were split into two groups - one containing the 89 YRI samples, and the other group containing the 373 samples of CEU, FIN, GBR and TSI ancestries. Within each group, genotype quality control was conducted on a per-variant level. Insertions, deletions, SNPs with a MAF<0.01, SNPs with imputation quality < 0.8 and SNPs with missing data in any of the samples were removed from each group. Data at the remaining variants and samples were taken forward for analysis.

In addition to details of the genotype data, full details of the gene expression data from Geuvadis are also provided in (Lappalainen *et al.*, 2013). Briefly, RNA sequencing was conducted on 462 individuals that had been genotyped in 1000 Genomes. RNA was extracted from Epstein Barr Virus (EBV) transformed lymphoblastoid cell lines (LCL) generated from 462 samples. Paired-end sequencing was then conducted using the Illumina HiSeq 2000. Reads were mapped using the GEM pipeline, and RPKM quantifications were calculated. Read count quantifications were normalised to the median number of well-mapped reads, and PEER normalisation was used to account for technical variation. These processed data were downloaded from the Geuvadis website and were used as-is for the downstream data analysis, with no additional processing.

2.1.3 ARIES data

Data from the Accessible Resource for Integrated Epigenomics Studies (ARIES) are used in Chapter 5 for investigating the prediction of CpG methylation data using SNP genotype data. The CpG methylation prediction models derived from these data are then used in Chapter 6 to investigate the role of genetically-regulated CpG methylation in complex traits.

The ARIES study measured methylation in approximately 1000 pairs of mothers and children for which genotype data had already been measured as part of the Avon Longitudinal Study of Parents and Children (ALSPAC) project. At the time of study, this was one of the largest data sets with matched measures of genotype and CpG methylation, making it ideal for this investigation.

All samples were genotyped as part of the larger ALSPAC project. Children were genotyped on the Illumina Human Hap 550-quad array, and mothers were genotyped separately on the Illumina human660W-quad array. For both groups, genotypes were called using Illumina Studio. Quality control and imputation were done at the ALSPAC cohort-wide level by the central ALSPAC team. Mothers and children underwent quality control separately. In each group, samples showing sex mismatch, high missingness, abnormal heterozygosity, non-European ancestry (determined by multidimensional scaling analysis) or a high degree of relatedness were excluded. At the variant level, SNPs showing a high degree of missingness, low MAF and low HWE p-value were removed. Genotype data from mothers and children were then combined at SNPs that passed QC in both groups and underwent imputation together. Samples were phased together using ShapeIT v2 and imputed to the 1000 Genomes phase 1 version 3 reference panel using Impute v2.2.2. Post-imputation genotypes were then provided to us.

Post-imputation quality control was then conducted by myself. Insertions, deletions, SNPs with imputation quality (determined via the INFO score) < 0.8, SNPs with MAF < 0.01, and A/T and C/G SNPs which could show strand ambiguity were removed.

Following this, PLINK files were then generated for the mothers for whom methylation was measured at the “antenatal” time point. Then, a genetic relationship matrix (GRM) was generated using HapMap3 SNPs. Pairs of samples showing

genetic relatedness > 0.05 in the GRM were identified, and one of each pair was removed, until no such pairs remained within the data for each time point. Genotype data for the remaining samples were taken forward.

Methylation for ARIES mothers and children was measured on the Illumina Infinium HumanMethylation450 BeadChip using bisulphite-converted DNA extracted from whole blood. For children, methylation was measured at 3 time points – birth, childhood and adolescence. For mothers, methylation was measured at 2 time points – at an antenatal clinic, and a follow-up clinic approximately 15 years after childbirth.

Processing, normalisation and initial quality control of the methylation data was done by the ARIES team. Raw IDAT files were read into R using the “meffil” package. Quality control was then performed at the per-sample and per-CpG level. At the sample level, samples showing genotype mismatches, sex mismatches, abnormal methylation intensities, dye bias, low bead number or more than 10% of probes being undetected were excluded. At the per-CpG level, CpG sites showing low bead number and CpG sites undetected in more than 10% of samples were excluded. For the remaining CpGs, meffil was used to conduct technical normalisation, which aims to account for batch variables, as well as performing dye bias and background signal corrections and correcting for the top 10 PCs in the methylation data.

After receiving the processed, normalised data, some additional quality control was performed by myself. CpG sites mapping to multiple regions in the genome, CpG sites at SNPs, and those with a SNP within the probe-binding sequence of the CpG site were all excluded.

Following this, methylation data for the mothers at the “antenatal” time point were extracted and were chosen to take forward. Any samples with missing covariate data were excluded, and samples that were earlier excluded during the genotype QC due to high relatedness with other samples were also excluded. A quality control report sent over by the ARIES team showed that top principal components of the normalised, post-processed methylation data were still associated with the batch variable “BCD_plate”, a categorical variable detailing the number of the plate used for DNA bisulfite conversion for each sample. To account for this, and other covariates, linear regression was used. At each CpG site, normalised methylation values were regressed on age, the “BCD_plate” variable, and the six estimated white blood cell

proportions. Residuals from these regressions were taken forward and used as methylation values in downstream analysis.

Overall, this left matched methylation and genotype data for 841 samples to be taken forward for analysis.

2.1.4 Understanding Society data

Matched genotype and CpG methylation data from the UK Household Longitudinal Study (also known as Understanding Society) are used in Chapter 5 to generate models that predict CpG methylation from SNP genotype data. These prediction models are used in Chapters 5 and 6 for the investigation of CpG methylation in a range of complex traits.

Approximately 10,000 samples from Understanding Society were genotyped using the Illumina Infinium HumanCoreExome BeadChip. The HumanCoreExome chip genotypes individuals at over 500,000 variants across the genome, with approximately 250,000 of those as genome-wide tagging SNPs and approximately 250,000 of these as rare, exonic variants. Genotypes were called using the gencall algorithm in Illumina GenomeStudio. Genotype data for the 1175 samples for which methylation was also measured were sent over by the Understanding Society team.

After I received the genotype data from the Understanding Society team, PLINK was used to identify and exclude SNPs with MAF<0.01, SNPs with per-SNP missingness ≥ 0.05, SNPs with a Hardy-Weinberg equilibrium test p value < 1e-05, SNPs located on sex chromosomes and to identify and exclude samples with missingness ≥ 0.05. To account for relatedness, a GRM was constructed using the SNPs that passed the quality control procedure. 51 pairs of samples with a genetic relatedness ≥ 0.05 were identified, and one from each pair was removed until no such pairs existed in the data set using PLINK with the *–rel-cutoff* function. Genotypes at the remaining SNPs were then taken forward for genotype imputation.

Genotype imputation was then performed using the Michigan Imputation Server. Genotypes were imputed to the 1000 Genomes Phase 1 reference panel, with ShapeIT used for phasing. Following imputation, indels, variants with imputation quality < 0.8, variants with MAF < 0.01 and A/T and C/G variants (which could show strand ambiguity) were removed. Additionally, samples with incomplete methylation

covariate data were removed. The remaining imputed genotypes were taken forward for analysis.

Genome-wide CpG methylation levels were measured in approximately 1,200 individuals (with matched genotype data) from the Understanding Society study. Measurement, data processing and quality control were all performed by the Understanding Society team. DNA from each sample was bisulfite-treated, before CpG methylation was using the Illumina Infinium HumanMethylationEPIC BeadChip. Generated IDAT files were processed in R using the *wateRmelon* and *bigmelon* package and converted to beta values. Outlier samples were identified and removed using the *outlyx* function in *wateRmelon*, and poor quality samples showing less than 85% of DNA bisulfite converted were identified and removed using the *bscon* function in *wateRmelon*. The *dasen* function in *wateRmelon* was then used to normalise betas, and the *qual* function was used to identify and remove samples showing a large difference in beta values between the pre-normalisation and post-normalisation betas. The *pfilter* function was then applied to remove probes with low bead count and poor detection p values. Overall, methylation data at 857,071 CpG sites in 1,175 samples passed this quality control. The original pre-normalised data for these 857,071 CpG sites and 1,175 samples were then extracted, and were normalised using the *dasen* function. These normalised data were then sent over to myself.

Alongside the normalised beta values, I also received a number of covariates for the methylation data. These included standard covariates, such as age and sex, as well as two batch variables and six estimates of white blood cell proportions per individual. Samples with incomplete covariate data were excluded at this stage. Following this, linear regression was used to adjust for covariates. For each CpG site, normalised beta values were regressed on age, sex (as a factor), the six estimated white blood cell proportions, and two batch variables. Residuals from linear regression were then taken forward for analysis.

Following this, problematic CpG sites were identified and removed. Supplementary tables downloaded from (McCartney *et al.*, 2016) were used to determine whether CpG sites mapped to multiple locations in the genome, and whether CpG sites were located at or near SNPs. CpG sites located on the sex chromosomes, CpG sites located at SNPs, CpG sites with SNPs with MAF>0.01 in the 1000 Genomes EUR population that were located in the probe-binding sequence, and CpG sites that mapped to multiple genomic regions were identified and removed. Additionally, the

samples removed from the genotype data due to high relatedness with other samples were also removed from the methylation data at this stage.

Overall, this left matched genotype and CpG methylation data at 787,334 CpG sites for 1,120 individuals to be taken forward for analysis.

2.1.5 INTERVAL data:

Matched genotype and proteomics data from the INTERVAL study are used in Chapter 7 to investigate prediction of protein levels from SNP genotypes. Protein level prediction models generated with these data are then used to investigate factors contributing to PBC in Chapter 7. This study had a large sample size, and analysis in (Sun *et al.*, 2018) implied that there were many effects of SNP genotypes on protein levels that could be modelled, making these data ideal for this section of work.

Genotyping, pre-imputation quality control and imputation were conducted as part of the INTERVAL study, and are described in detail in (Astle *et al.*, 2016). Approximately 50,000 INTERVAL participants were genotyped using the Affymetrix Axiom UK Biobank array. Following genotyping, sample-level QC was conducted to remove duplicates, samples showing low call rate, high heterozygosity, sex mismatch, highly related samples and samples of non-European ancestry, as determined by multidimensional scaling. Per-SNP QC was also conducted, removing SNPs on sex chromosomes, non biallelic SNPs, SNPs with a low call rate and SNPs not in Hardy-Weinberg equilibrium. Following quality control, imputation to a combined UK10K-1000 Genomes phase 3 reference panel was conducted using the Sanger Imputation Server.

Following imputation, some quality control was then performed by (Sun *et al.*, 2018), who also carried out the proteomics measurements. They removed imputed variants with imputation quality < 0.7, variants with minor allele count < 8, variants with Hardy-Weinberg equilibrium p value < 5e-6, and where duplicate variants had the same position and alleles, the variant with the lowest imputation quality was removed. This results in imputed genotypes for 10,572,788 variants in 3301 samples that also had proteomics measures. These genotype data were then sent to us.

To ensure a high quality data set was used for modelling, additional post-imputation variant filtering was conducted by myself. The *–summary_stats_only* function in SNPTEST was used to identify indels, A/T and C/G SNPs, SNPs showing imputation quality < 0.8 and SNPs with MAF < 0.01 and SNPs without an RS number. These variants were then removed using *bgenix*. No samples were removed at this stage.

Proteomics measurements and quality control were conducted by (Sun *et al.*, 2018). 150ul aliquots of plasma were taken from 3301 individuals, and plasma protein levels were measured using the microarray-based assay SOMAscan. Proteins were measured using 4,034 unique SOMAmers.

Control probes were used to calculate hybridisation scale factors which were used to normalise data for within-run variation, while control samples were used to calculate calibration scale factors, which were used to normalise data for between-run variation. Samples showing extreme hybridisation scale factors and SOMAmers showing extreme calibration scale factors (compared to the median) were excluded. In addition, SOMAmers binding to non-human targets and SOMAmers showing coefficient of variation below 20% were excluded, leaving 3283 SOMAmers to be taken forward. Protein levels were adjusted for age, sex, batch variables and the top 3 principal components (from multi-dimensional scaling) using linear regression, with residuals taken forward for analysis).

Post-QC protein levels were then sent over to myself. Uniprot IDs were mapped to Ensembl gene IDs using the Uniprot website, and chromosome and TSS of these genes were identified using Gencode v27. Proteins mapping to multiple genes, and proteins mapping to genes on sex chromosomes were removed, leaving 3,106 SOMAmers for downstream analysis.

2.1.6 PBC cases proteomics data

Genotyping and imputation of PBC cases was conducted as part of a larger ongoing PBC meta-analysis from which summary statistics are used. DNA extracted from blood samples from PBC cases of UK ancestry was genotyped. Following genotype calling, per-variant and per-sample QC were performed by Heather Cordell. The genotypes of the 418 samples with matched proteomics data were then extracted, and further QC was carried out on the post-imputation genotypes by myself. SNPs

with MAF < 0.01, A/T and C/G SNPs and SNPs showing missingness in any samples were excluded from further analysis.

The concentrations of 368 proteins in 630 blood serum samples taken from PBC cases were measured using the Olink assay (Lundberg *et al.*, 2011). Proteins were measured on 4 panels - Cardiovascular II, Cardiovascular III, Inflammation and Oncology II. In-house data processing, quality control and data normalisation was done by the Olink team, who provided protein measures as NPX (normalised protein expression).

Uniprot protein IDs of the 368 proteins (provided by Olink) were matched to Ensembl gene IDs using the Uniprot website, and the chromosome and transcription start site of these genes were identified using Gencode v27. One protein mapped to multiple gene IDs and 8 proteins mapped to genes located on the X chromosome, so were excluded. As recommended by Olink, protein measures for which a sample failed their internal QC thresholds and protein measures below the limit of detection were considered to be missing. 18 proteins for which 50% or more of samples showed missing data were excluded. This left 341 proteins to be taken forward for downstream analysis.

8 duplicate samples were identified among the 630 samples. The sample in each duplicate pair showing the most missing data was removed. In addition, samples that failed Olink QC on multiple panels were also excluded, leaving 403 samples to be taken forward for downstream analysis.

In total, there were matched genotype and protein level data for 403 samples, which was the maximum sample size available for modelling. However, as there were some missing protein data, a slightly reduced sample size was used for modelling most proteins.

2.1.7 GWAS summary data used in the thesis

Publicly available summary statistics from GWAS of a wide range of complex traits are used throughout this thesis.

CD data from (Liu *et al.*, 2015a) and T1D data from (Cooper *et al.*, 2017) are first used to conduct a TWAS of CD and T1D in Chapter 3 in an attempt to detect more

associations than detected just using the WTCCC CD and T1D data. These summary data were used as-is, with no additional processing.

30 sets of GWAS summary statistics are used to conduct the MWAS, TWAS, colocalisation and Mendelian Randomisation analyses in Chapter 6. The details of these summary statistics are given in Table 2.1. Note that the T1D data shown in Table 2.1 are the same as those used in Chapter 3, while the CD data are derived from a more recent GWAS.

Complex trait	Abbreviation	Reference
Age at menarche	AAM	http://www.nealelab.is/uk-biobank
ALS	ALS	van Rheenen *et al.* (2016)
Asthma	Asthma	http://www.nealelab.is/uk-biobank
Body mass index	BMI	http://www.nealelab.is/uk-biobank
Basal metabolic rate	BMR	http://www.nealelab.is/uk-biobank
Bipolar disorder	BP	http://www.nealelab.is/uk-biobank
Crohn's disease	CD	de Lange *et al.* (2017)
Chronic heart disease	CHD	http://www.nealelab.is/uk-biobank
Diastolic blood pressure	DBP	http://www.nealelab.is/uk-biobank
Forced vital capacity	FVC	http://www.nealelab.is/uk-biobank
Glaucoma	Glaucoma	http://www.nealelab.is/uk-biobank
Hayfever, allergic rhinitis or eczema diagnosed by a doctor	HAE	http://www.nealelab.is/uk-biobank
Hand grip strength	HGS	http://www.nealelab.is/uk-biobank
Heel bone mineral density	HBMD	http://www.nealelab.is/uk-biobank
High-density lipoprotein	HDL	http://www.nealelab.is/uk-biobank
Height	Height	http://www.nealelab.is/uk-biobank
Inflammatory bowel disease	IBD	de Lange *et al.* (2017)
Low-density lipoprotein	LDL	http://www.nealelab.is/uk-biobank
Pulse rate	Pulse	http://www.nealelab.is/uk-biobank
Red blood cell count	RBCCount	http://www.nealelab.is/uk-biobank
Recurrent depressive disorder	RDD	http://www.nealelab.is/uk-biobank
Systolic blood pressure	SBP	http://www.nealelab.is/uk-biobank
Schizophrenia	SCZ	http://www.nealelab.is/uk-biobank
Type 1 diabetes	T1D	(Cooper *et al.*, 2017)
Type 2 diabetes	T2D	(Scott *et al.*, 2017)
Total cholesterol	TC	http://www.nealelab.is/uk-biobank
Triglycerides	TG	http://www.nealelab.is/uk-biobank
Ulcerative colitis	UC	de Lange *et al.* (2017)
Weight	Weight	http://www.nealelab.is/uk-biobank
White blood cell count	WBCCount	http://www.nealelab.is/uk-biobank

Table 2.1. 30 sets of publicly available GWAS summary data used for MWAS and TWAS analyses in Chapter 6

For the summary statistics from the GWAS of UK Biobank data conducted by Ben Neale's group, a list of SNPs that were found to fail standard GWAS quality control procedures by Ben Neale's group were removed from the summary statistics prior to

any downstream analysis. For the other sets of GWAS summary data that were not conducted by Ben Neale's group, no additional quality control or data processing was performed, and the summary statistics were used as-is in downstream analyses.

Summary statistics from a recent meta-analysis of PBC are used in multiple analyses in Chapter 7. These summary statistics were generated by Heather Cordell and are as yet unpublished (manuscript in preparation).

2.1.8 QTL summary data used in the thesis

In addition to the GWAS summary data described above, QTL summary statistics are also used in this thesis. Summary statistics from cis-eQTL analysis are used for colocalisation analysis in Chapters 6 and 7. The eQTL analysis used to generate these summary statistics was performed by the GTEx team and is described in detail on the GTEx portal website. Briefly, using data from GTEx version 7, normalised whole blood gene expression levels were regressed on genotypes at SNPs with MAF > 0.01 within 1 Mb of the gene transcription start and end sites, with top genotype principal components, PEER factors, sex and genotyping platform included as covariates, using linear regression implemented in FastQTL. The summary statistics for all tests performed in this analysis were downloaded from the GTEx portal. No additional processing or quality control was done to these summary statistics, which were used as-is.

Self-generated cis-mQTL summary statistics are used for colocalisation analysis in Chapter 6. For each CpG site, CpG methylation was regressed on genotypes of SNPs within 3 Mbs of the CpG site. This analysis was carried out using PLINK.

Summary statistics from published cis-mQTL analysis are used for colocalisation analysis in Chapter 7. These mQTL summary data were taken from (Gaunt *et al.*, 2016). Briefly, using matched genotype data and DNA methylation data measured at the antenatal time point from the ARIES study, methylation at each CpG site was regressed on genotypes at SNPs across the genome, with age, sex, top genotype principal components, CpG methylation batch variables and blood cell proportion estimates used as covariates. Summary statistics for the tests with a p value < 1×10^{-7} were downloaded. No additional processing or quality control were performed after downloading these data, which were then used as-is.

Summary statistics from published cis-pQTL analysis are used for colocalisation analyses in Chapter 7. These pQTL summary data were taken from (Sun *et al.*, 2018). Briefly, using data from the INTERVAL study, protein levels were regressed on genotypes at SNPs using the SNPTEST program. No additional processing or quality control were performed after downloading these data, which were then used as-is in the colocalisation analysis.

2.2 Evaluating prediction accuracy of gene expression, CpG methylation and protein levels

2.2.1 Comparison of seven different statistical approaches for predicting gene expression from SNP genotypes

In Chapter 4.2 the ability of seven different statistical methods to predict gene expression from SNP genotypes is compared using a nested cross-validation approach. A nested cross-validation approach was chosen for this analysis because some of the statistical approaches being compared required their model parameters to be tuned, but using the same cross-validation to select appropriate values for model tuning parameters and to estimate prediction accuracy is thought to lead to overestimation of prediction accuracy. By using a nested cross-validation, the model tuning and prediction accuracy evaluation steps are performed separately, avoiding this inflation of prediction accuracy estimates. The details of the seven methods being compared are described in detail in Chapter 4.1.

A nested cross-validation consists of an outer cross-validation loop, which is used to evaluate prediction accuracy and inner cross-validation loops, which are used to tune model parameters. First, the samples are split into ten groups of near-equal size that are used for the outer cross-validation loop. For each fold of the outer cross-validation loop, one of the ten sample groups is assigned as the outer model testing group, and the other nine groups of samples are combined together and assigned as the outer model training group. After assigning the training and testing sets within the fold of the outer loop, the inner cross-validation loop is then conducted to allow for parameter tuning. A range of potential values is defined for the parameter of interest. The outer model training group is then itself split into ten groups of near-equal size. For each parameter value in the list, a prediction model is then trained using nine of these groups (the inner training set) and is applied to data from the remaining group

(the inner test set). This process is then repeated another nine times, each time using a different group of samples as the inner test set, and so that each sample that is used in the inner loop is part of the inner test set once and only once. This gives an estimate of prediction accuracy for each of the parameter values tested. The parameter value at which the maximum correlation between the predicted and observed values is then chosen as the optimal parameter value.

Following this, a prediction model is then trained using all samples from the outer training group, using the optimal parameter value chosen via the inner cross-validation loop. This prediction model is then applied to the outer testing group, and the correlation between predicted and observed values is estimated. This whole procedure is then performed another nine times, each time using a different group of samples as the outer testing group, and so that each sample is used in the outer testing group once and only once. The overall prediction accuracy estimate from the 10-fold nested cross-validation is taken as the mean of the ten correlations between predicted and observed values obtained from the outer loop.

This approach is used in Chapter 4.2 to evaluate the prediction accuracy achieved by seven different statistical approaches. These approaches were LASSO (Tibshirani, 1996), ridge regression (Hoerl and Kennard, 2000), two forms of elastic net (Zou and Hastie, 2005), BSLMM (Zhou *et al.*, 2013), BLUP (de Los Campos *et al.*, 2013) and Random Forests (Breiman, 2001). These approaches are described in further detail in Chapter 4.1. For the LASSO, ridge regression, and elastic net (with alpha = 0.5) approaches, the inner cross-validation was used to tune the lambda parameter (a tuning parameter). For the elastic net (with alpha determined by cross-validation), the inner cross-validation was used to tune both the lambda and alpha parameters (where lambda is the tuning parameter and alpha determines the sparsity of the model). For BSLMM, BLUP and Random Forests, the inner loop was not performed. The BSLMM and BLUP approaches use Markov chain Monte Carlo (MCMC) to estimate their hyperparameters, so the inner loop was not required. Although the Random Forests approach has parameters that could have been tuned using the inner loop, tuning these parameters is computationally demanding and time consuming, and so the default parameter values were used and the inner loop was not required.

2.2.2 Examining the effect of sample size on gene expression prediction accuracy

In Chapter 4.4 the effect of the sample size of the reference panel used to train gene expression prediction models on gene expression prediction accuracy is investigated using a similar nested cross-validation approach. The 10-fold nested cross-validation procedure described above is repeated, but in each fold of the outer loop of cross-validation, only one of the ten groups was used as the prediction model training set, with the other nine groups combined and used as the testing set. The rest of the nested cross-validation procedure remained unchanged. These gene expression prediction models were trained using the elastic net method, with the alpha parameter set to 0.5 and the lambda parameter tuned via the inner 10-fold cross-validation, with the value of lambda chosen as that at which the maximum correlation between the predicted and observed gene expression was observed in the inner 10-fold cross-validation. The prediction accuracy was calculated as the mean of the ten correlations between predicted and observed expression obtained from the outer cross-validation loop. The prediction accuracy estimates obtained from this nested cross-validation approach using the reduced training set sample size were then compared with the estimates obtained in Chapter 4.2. Following this, the nested cross-validation was repeated a further seven times, each time using a different proportion of the samples to train the gene expression prediction models (20% of samples in the first repeat, 30% of samples in the second repeat, up to 80% of samples in the seventh repeat). For each repeat, the prediction accuracy was calculated as the mean of the ten correlations between predicted and observed expression obtained from the outer cross-validation loop. The prediction accuracy estimates obtained from these nested cross-validations were also compared with the estimates obtained in Chapter 4.2.

2.2.3 Examining the effect of ancestry on gene expression prediction accuracy

In Chapter 4.5 the effect of the ancestry of the gene expression prediction model training and testing data sets on gene expression prediction accuracy is examined. First, the ability of gene expression prediction models trained using data from European (EUR) samples to predict the gene expression of Yoruban (YRI) samples was investigated. To do this, gene expression prediction models were trained using

the whole set of 373 EUR samples. These gene expression prediction models were trained using the elastic net method, with the alpha parameter set to 0.5 and the lambda parameter tuned via 10-fold cross-validation, with the value of lambda chosen as that at which the maximum correlation between the predicted and observed gene expression was observed in the 10-fold cross-validation. These prediction models were then applied to the 89 YRI samples, and predictive performance was evaluated as the correlation between gene expression predicted by the EUR-trained prediction models and the measured gene expression of the YRI samples. These prediction accuracy estimates were then compared with those obtained from the 10-fold nested cross-validation conducted using only EUR samples in Chapter 4.2.

Then, to examine the ability of gene expression prediction models trained using data from YRI samples to predict the gene expression of EUR samples, gene expression prediction models were trained using data from the 89 YRI samples. These gene expression prediction models were trained using the elastic net method, with the alpha parameter set to 0.5 and the lambda parameter tuned via 10-fold cross-validation, with the value of lambda chosen as that at which the maximum correlation between the predicted and observed gene expression was observed in the 10-fold cross-validation. These prediction models were then applied to the 373 EUR samples, and the correlation between the gene expression predicted by the models trained using the YRI data and the measured expression of the EUR samples was calculated. In addition, a 10-fold nested cross-validation was performed on the 89 YRI samples using the same procedure as used for the 10-fold nested cross-validation on EUR samples. The prediction accuracy estimates from the application of the YRI-trained prediction models to the EUR samples were then compared with the prediction accuracy estimates obtained from the 10-fold nested cross-validation on the 89 YRI samples.

Following this, a combined analysis was performed. The 373 EUR and 89 YRI samples were combined into a single group of 462 samples, and down-sampled to 373 samples, keeping the relative proportion of EUR and YRI samples the same as in the larger group of 462 samples. A 10-fold nested cross-validation was performed on this mixed group of 373 samples, using the procedure as described above. These gene expression prediction models were trained using the elastic net method, with the alpha parameter set to 0.5 and the lambda parameter tuned via 10-fold cross-

validation, with the value of lambda chosen as that at which the maximum correlation between the predicted and observed gene expression was observed in the 10-fold cross-validation. The prediction accuracy estimates obtained from this were then compared with the prediction accuracy estimates obtained from the 10-fold nested cross-validation on the 373 EUR samples.

2.2.4 Examining the effect of tissue on gene expression prediction accuracy

In Chapter 4.6 the ability of prediction models to predict across tissues is examined. In this analysis, gene expression prediction models trained using data from version 6 of GTEx were used. These prediction models were trained by the developers of the PrediXcan software and were downloaded from predictdb.org. In total, 48 sets of gene expression prediction models were used, each of which was trained using gene expression data from a different tissue. Each set of prediction models was applied to genotype data from the EUR samples from the Geuvadis project to predict gene expression. The correlation between these predicted gene expression values and the measured Geuvadis gene expression values was then calculated. The prediction accuracy estimates achieved by the different sets of gene expression prediction models were then compared.

2.2.5 Comparison of three methods for predicting CpG methylation from SNP genotypes

In Chapter 5.1 a comparison of three methods (LASSO, elastic net (with the alpha tuning parameter set to 0.5) and ridge regression) for training CpG methylation prediction models is performed. Using each method, CpG methylation prediction models were trained using a training set that comprised 50% of the samples from ARIES at the antenatal time point. For each method, the lambda parameter was determined by a 10-fold cross-validation on the training set. Any values of lambda that produced a final prediction model that did not contain any SNPs were excluded. Of the remaining values of lambda, the value at which the minimum mean squared error between predicted and observed methylation was achieved in the cross-validation was then selected. A prediction model was then trained on the whole training set using this optimal value of lambda. This prediction model was then

applied to a testing set that consisted of 20% of ARIES samples, and the correlation between the predicted and observed methylation values in this test set was calculated.

2.2.6 Comparison of five window sizes for predicting CpG methylation from SNP genotypes

In Chapter 5.2 a comparison of five window sizes is performed. These five window sizes were 250Kb, 500Kb, 1Mb, 2Mb and 3Mb. For each window size, CpG methylation prediction models were trained using the training set consisting of 50% of ARIES samples. CpG methylation prediction models were trained using the elastic net method, with the alpha parameter set to 0.5, and the lambda parameter determined by the same cross-validation procedure as described above for Chapter 5.1. For each window size, the CpG methylation prediction models were trained using all SNPs within the specified distance from the CpG site. CpG methylation prediction models were then applied to the test set that consisted of 20% of ARIES samples, and the correlation between the predicted and observed methylation was calculated. The optimal window size was then determined as the window size at which the maximum correlation between predicted and measured methylation was observed.

2.2.7 Evaluating CpG methylation prediction accuracy at the optimal method and window size

In Chapter 5.3, having determined the optimal method and window size for the prediction of CpG methylation, I investigate the prediction accuracy of the procedure. Prediction models were trained using the optimal method and window size. Here, the training set consists of 70% of samples (the previous 50% training set and 20% testing set combined into a single group), while the testing set comprises the remaining 30% of samples that have not been used at any point prior to this. Again, the prediction models were trained using the elastic net method, with the alpha parameter set to 0.5, and the lambda parameter determined by the same cross-validation procedure as described above for Chapter 5.1. Prediction models were

trained using CpG-specific window sizes that had been determined as optimal in Chapter 5.2.

2.2.8 Evaluating protein level prediction accuracy

In Chapter 7, protein level prediction models are trained and tested to determine how accurately protein levels can be predicted from SNP genotypes. To do this, a 10-fold nested cross-validation approach was used. This approach was highly similar to the approach described in Chapter 2.2.1 that was used to examine prediction accuracy for gene expression, although there were some differences.

When examining the prediction accuracy of protein levels using INTERVAL data, the same procedure described in Chapter 2.2.1 was used, but with one difference. In the inner cross-validation, the value of lambda was not selected as that at which the maximum correlation between predicted and measured values was observed. Instead, the value of lambda was selected by excluding any values of lambda at which the prediction model did not contain any SNPs, and then of the remaining values of lambda, selecting the value at which the minimum mean squared error was observed in the inner cross-validation.

When examining the prediction accuracy of protein levels using the PBC cases data, the same procedure described in Chapter 2.2.1 was used, but with two key differences. First, as there was some missing proteomics data in the PBC cases data, the data were split into the 10 groups used for the outer cross-validation on a protein-by-protein basis. For each protein, any samples with missing proteomics data were excluded, and the remaining samples were randomly split into 10 groups of near equal size. Second, lambda was tuned not by selecting the value at which the maximum correlation between predicted and measured values was observed in the inner cross-validation. Instead, lambda was chosen by first excluding any values of lambda at which the prediction model did not contain any SNPs, and then of the remaining values of lambda, selecting the value at which the minimum mean squares error was observed in the inner cross-validation.

In both of these 10-fold nested cross-validation analyses, protein level prediction models were fitted by regressing protein levels on genotypes of SNPs within 1 Mb of the transcription start site of the protein's corresponding gene. As the protein level

prediction models were eventually intended to be applied only to the PBC summary data, only those SNPs which were present in both the PBC summary data and in the proteomics genotype data were used. The prediction models were fitted using elastic net, with the alpha parameter set to 0.5.

2.3 Transcriptome wide association study (TWAS) and similar methods

The TWAS method is used many times throughout the thesis as a method for investigating the role of gene expression in complex traits. The general principle of TWAS will be described here.

The first step in TWAS analysis is to train a set of gene expression prediction models using matched genotype and gene expression data. These prediction models are usually trained by regressing gene expression on genotypes of SNPs proximal to the gene. These prediction models take the form:

$$y_{ig} \sim \sum_{l=1}^{p} x_{il}\beta_{lg} + \varepsilon_{ig}$$

where y_{ig} is the expression of gene *g* in individual i, x_{il} is the effect allele count (0, 1 or 2) of SNP *l* in individual *i*, β_{lg} is the weight of SNP l on gene g, *p* is the total number of SNPs in the prediction model and ε_{ig} is an error term that includes all non-genetic effects on expression.

Following on from this, the prediction models are then applied to GWAS genotype data to impute gene expression for the GWAS samples. Finally, the GWAS phenotype is then regressed on the predicted gene expression levels to obtain estimates of the effect size of gene expression on the phenotype, and the associated standard error, z score and p value.

Alternatively, the same results of a TWAS can be derived using the gene expression prediction models and summary statistics (betas and standard errors) from a GWAS, without actually doing the gene expression imputation and the regression. This summary statistics based approach does not require individual level genotype and phenotype data from the GWAS. This approach is taken by the MetaXcan and FUSION software packages. The equations used to derive TWAS z scores from

prediction models and GWAS summary data are specific to each software package and are given in Chapter 3.1.

2.3.1 Comparison of TWAS results using gene expression prediction models trained using seven different statistical methods

TWAS analysis is used in Chapter 4.4. First, gene expression prediction models were trained using seven different statistical approaches. The details of the seven methods are described in detail in Chapter 4.1. Each prediction model was trained using data from all 373 EUR samples from the Geuvadis project, using the genotypes of all SNPs within 1 Mb of the gene transcription start site. Following this, the gene expression prediction models were then applied to WTCCC1 T1D GWAS data using the PrediXcan package. Finally, the results were compared.

2.3.2 TWAS and MWAS of 30 complex traits

TWAS is used in Chapter 6, where it is applied to publicly available GWAS summary data for 30 complex traits using the MetaXcan software package to identify associations between predicted gene expression and the complex traits.

The prediction models used here were trained by the developers of the PrediXcan and MetaXcan packages. These prediction models were trained by regressing whole blood gene expression data on the genotypes of all SNPs within 1 Mb of the gene transcription start site, using elastic net (with alpha set to 0.5, and lambda determined by 10-fold cross-validation). These prediction models were trained using data from GTEx version 7. The prediction models were downloaded from predictdb.org and were applied to 30 sets of publicly available GWAS summary data, using the MetaXcan software package.

In addition to this, the TWAS method is adapted and used to conduct a methylome wide association study (MWAS) in Chapter 6, in which CpG methylation prediction models were applied to the same 30 sets of publicly available GWAS summary data to detect associations between predicted CpG methylation and the complex traits. The CpG methylation prediction models were trained using either all samples from ARIES, or all samples from Understanding Society. Prediction models were only trained for those CpG sites where a prediction accuracy ≥ 0.1 was achieved in the

training and testing procedure when using the optimal method and window size (described in Chapter 2.2.7). The CpG methylation prediction models were trained by regressing CpG methylation on genotypes of all SNPs within a CpG-specific distance (the procedure for determining this distance is described in Chapter 2.2.6) using elastic net, with the alpha parameter set to 0.5 and the lambda parameter determined by a 10-fold cross-validation. Any values of lambda that produced a final prediction model that did not contain any SNPs were excluded. Of the remaining values of lambda, the value at which the minimum mean squared error between predicted and observed methylation was achieved was then selected.

2.3.3 TWAS, MWAS and PWAS of PBC

In Chapter 7, a TWAS is conducted by applying gene expression prediction models to PBC meta-analysis summary data using the MetaXcan package. 48 sets of gene expression prediction models, each trained using data from a different GTEx tissue, were downloaded from the predictdb.org repository and applied to the PBC GWAS summary data using MetaXcan.

In addition, an MWAS is also conducted by applying CpG methylation prediction models to PBC summary data using the MetaXcan package. The same set of CpG methylation prediction models described in Chapter 2.3.2 are used here.

Finally, a PWAS is also conducted by applying protein level prediction models to the same PBC summary data using the MetaXcan package. Protein level prediction models were trained only for those proteins for which a prediction accuracy ≥ 0.1 was achieved in the 10-fold nested cross-validation (described in Chapter 2.2.8). For each of these proteins, protein level prediction models were fitted by regressing protein levels on genotypes of SNPs within 1 Mb of the transcription start site of the protein's corresponding gene. As the protein level prediction models were eventually intended to be applied only to the PBC summary data, only those SNPs which were present in both the PBC summary data and in the proteomics genotype data were used. The protein level prediction models were fitted using elastic net, with the alpha parameter set to 0.5, and with the lambda parameter determined by the inner cross-validation, with the value of lambda chosen as that at which the minimum mean squared error was achieved in the inner cross-validation.

2.4 Heritability estimation

The heritability of a trait is equivalent to the upper bound of the accuracy with which the trait can be predicted using only genetic information. In this thesis, gene expression and CpG methylation are predicted using only SNP genotype data. Estimating the heritability of these two traits provides an estimate of the prediction accuracy values that could be expected were the prediction operating perfectly.

The aim of the heritability estimation was to obtain values for the upper bound of prediction accuracy that could then be compared with the actual prediction accuracy values that were obtained from training and testing prediction models. These actual prediction accuracy estimates were obtained by training prediction models using only the SNPs proximal to the genes (or CpG sites) of interest. So, to obtain estimates of the heritability of gene expression (or CpG methylation) attributable to the same set of proximal SNPs (that would be comparable with the prediction accuracy estimates), the GCTA method was used to estimate heritability in this thesis.

GCTA fits the effects of SNPs on phenotype using the following linear mixed model:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{g} + \varepsilon$$

$$\boldsymbol{V} = \boldsymbol{A}\sigma_g^2 + \boldsymbol{I}\sigma_\varepsilon^2$$

$$\boldsymbol{A} = \boldsymbol{WW}'/N$$

Where ***y*** is a vector of phenotype measures (in this thesis, either gene expression of CpG methylation), ***X*** is a vector of covariates such as age and sex, ***β*** is a vector of fixed effects, ***g*** is a vector of the total genetic effects of the individuals, ***ε*** is an error term, **V** is the variance of the phenotype ***y***, ***A*** is a genetic relationship matrix (GRM) between individuals, ***W*** is a genotype matrix, N is the number of SNPs used to construct the GRM, σ_g^2 is the variance explained by all SNPs, ***I*** is an identity matrix and σ_ε^2 is the residual variance.

For each gene or CpG site of interest, ***W*** was restricted to be only the set of SNPs proximal to the gene/CpG of interest. By constructing the GRM using only the SNPs proximal to the gene/CpG, only the portion of heritability attributable to the SNPs proximal to the gene could be estimated. When estimating the heritability of gene expression in Chapter 4, the GRM for each gene was constructed using genotypes at

all SNPs within 1 Mb of the gene start or end site (the same set of SNPs used to generate gene expression prediction models in the 10-fold cross-validation procedure). When estimating the heritability of CpG methylation in Chapter 5, the GRM for each CpG site was constructed using genotypes at all SNPs within a CpG-specific distance from the CpG site. The procedure used to determine this CpG-specific distance is described in Chapter 2.2.6.

After constructing the GRM, restricted maximum likelihood analysis was implemented in GCTA to estimate the heritability. In this thesis, prediction accuracy was estimated as the correlation between predicted and observed gene expression (or CpG methylation). Squaring this correlation gives the R-squared estimate for a gene, which is the proportion of variance of the measured gene expression explained by the predicted gene expression (the value that is directly comparable with the heritability estimate). As this R-squared value is the square of a value bounded between -1 and 1, the R-squared estimate itself must fall within the [0,1] range. As the intention of heritability estimation was to compare the heritability estimates with the prediction accuracy estimates obtained from training and testing prediction models, heritability estimates from GCTA were restricted to fall within the [0,1] range.

2.5 Bayesian multi-trait colocalisation analysis

One of the vulnerabilities of the TWAS framework is that associations between predicted gene expression and phenotype can be caused by linkage disequilibrium (LD) between two different genetic variants, one of which affects gene expression and the other which affects the phenotype. These LD-induced associations are less biologically interesting than the associations that reflect true causality (or pleiotropy), and so identifying which associations have been induced by LD is important. To determine whether MWAS and TWAS associations detected in Chapter 6 had been induced by LD, a Bayesian colocalisation approach was taken. As there were more than two traits in each run of the colocalisation analysis, the multi-trait Bayesian colocalisation (Giambartolomei *et al.*, 2018) method was used.

Briefly, the approach aims to estimate the posterior probability (PPA) that multiple traits share the same causal SNP. Summary association statistics (betas and standard errors) from tests of associations between SNP genotypes and traits are used to compute approximate Bayes Factors. These Bayes Factors are then used to

estimate the posterior probability of a range of potential hypotheses for the sharing of causal SNPs between traits. Full details on how Bayes Factors are estimated and used to estimate posterior probabilities are given in (Giambartolomei *et al.*, 2018).

2.5.1 Colocalisation of CpG methylation, gene expression and complex traits

Multi-trait colocalisation (moloc) analysis was used in Chapter 6 to test for colocalisation between CpG methylation, gene expression and complex traits. Prior to colocalisation analysis, MWAS and TWAS were used to identify associations between CpG methylation, gene expression and complex traits. The colocalisation analysis was then used to identify which of these associations had been induced by LD.

To perform the analysis, per-SNP summary data (beta, standard error, sample size and MAF) from regression of CpG methylation, gene expression, and complex traits on genotypes of SNPs were required. Here, whole blood eQTL data downloaded from the GTEx portal, a self-generated set of mQTL summary data generated using ARIES and Understanding Society data, and the 30 sets of publicly available GWAS data were used. The sources for these summary data are outlined in Chapter 2.1.8.

For each trio of a CpG site, gene and complex trait, moloc analysis was carried using the *moloc.test* function implemented using the *moloc* package in R. The default options and priors were used for all moloc analyses conducted here. A hypothesis was considered to be strongly supported if the moloc analysis gave it a posterior probability >= 0.8.

2.5.2 Colocalisation of CpG methylation, gene expression, protein levels and PBC

Moloc analysis was used in Chapter 7 to test for colocalisation between CpG methylation, gene expression, protein levels and PBC. Here, moloc was not used to test a specific hypothesis, as done in Chapter 6, but was used in a more agnostic, broad scanning approach to detect as many colocalisations between PBC and omics traits as possible. Furthermore, this analysis was conducted using publicly available mQTL summary data, rather than the mQTL summary data generated by myself that were used in the moloc analysis in Chapter 6. These publicly available mQTL

summary data only contained SNPs that were significantly associated with CpG methylation, rather than all SNPs tested. As a result of these factors, a slightly different approach to conducting the moloc analysis was taken here.

First, 3 sets of pairwise colocalisation analyses were conducted, each between PBC and one of the omics traits (i.e. PBC and gene expression, PBC and methylation, PBC and proteomics). These pairwise colocalisation analyses were conducted using the R package *coloc*. The results for which at least 50 SNPs were used for colocalisation analysis, and which showed a posterior probability >= 0.8 for the hypothesis in which PBC colocalised with the omic trait were taken forward.

Following this, each possible multi-trait colocalisation involving PBC and two intermediate traits (PBC-expression-methylation, PBC-expression-protein, PBC-methylation-protein) was tested using moloc. This analysis was conducted using the *moloc.test* function implemented using the *moloc* package in R. Results in which at least 50 SNPs were used, which showed strong evidence of colocalisation between all 3 traits (PPA >= 0.8), and in which each of the two intermediate traits showed strong evidence of colocalisation in the previous pairwise test were taken forward.

Finally, multi-trait colocalisation involving PBC and all 3 intermediate traits was tested using moloc. This analysis was conducted using the *moloc.test* function implemented using the *moloc* package in R. Results in which at least 50 SNPs were used, which showed strong evidence of colocalisation between all 4 traits (PPA >= 0.8), and in which each of the three intermediate traits showed strong evidence of colocalisation in the previous pairwise tests were taken forward.

2.6 Mendelian Randomisation

Mendelian Randomisation is a form of instrumental variable analysis that is used to test for a causal relationship between an exposure and an outcome of interest. In Chapter 6, two-step Mendelian Randomisation analysis is performed to test for a causal effect of CpG methylation on gene expression, and a subsequent causal effect of gene expression on complex traits (Relton and Davey Smith, 2012). This analysis was carried out for each of the trios where CpG methylation, gene expression and the complex trait of interest were all found to colocalise to the same

causal SNP with a posterior probability >= 0.8 using the moloc analysis. The same per-SNP summary data used for moloc analysis were also used for MR.

Two step MR requires two independent instruments to work. To select independent instruments, the following algorithm was used:

1. Select all SNPs associated with methylation at the CpG site in question at $p<5x10^{-8}$ (from the previously determined ARIES or Understanding Society mQTLs) as step 1 instruments
2. Remove step 1 instruments where effect sizes and standard errors on gene expression are not available in GTEx data
3. Select all SNPs associated with gene expression at the gene in question at $p<5x10^{-8}$ (from GTEx eQTLs) as step 2 instruments
4. Remove step 2 instruments where effect sizes and standard errors on the complex trait of interest are not available in the GWAS data
5. Remove step 1 instruments in strong LD ($r^2 > 0.01$) with all potential step 2 instruments
6. Select the strongest remaining (lowest p value) step 1 instrument and take forward for MR analysis
7. Remove any step 2 instruments in strong LD ($r^2 > 0.01$) with this selected step 1 instrument as determined using European samples from 1000 Genomes Phase 1
8. Select the strongest remaining (lowest p value) step 2 instrument and take forward for MR analysis

After selecting independent instruments for both steps, the MR analysis was carried out using the *TwoSampleMR* R package. MR analysis was only carried out where a valid instrument was taken forward for both step 1 and step 2. After identifying valid instruments, the Mendelian Randomisation analysis was carried out using the Wald ratio test.

2.7 Enrichment testing

2.7.1 Gene set enrichment testing

Gene set enrichment testing is used in Chapter 4 to test whether a common function/role exists among the set of genes for which gene expression could be predicted well from SNP genotypes.

The gene set enrichment analysis was conducted using the “gene2func” option in the FUMA software (Watanabe *et al.*, 2017). This software tests a list of input gene against curated sets of genes (e.g. Gene Ontology gene sets, GWAS catalogue gene sets) taken from MSigDB and WikiPathways using hypergeometric tests. Here, the list of genes for which a prediction accuracy estimate >= 0.5 was achieved in the 10-fold nested cross-validation with data from the 373 Geuvadis samples of European ancestry was used as the input for this test, and the default set of background genes were used. Genes in the MHC regions were not excluded from this analysis. The gene sets with a Benjamini-Hochberg adjusted p value < 0.05 were reported as significantly enriched.

2.7.2 CpG site enrichment testing

Enrichment testing is also used in Chapter 5 to test whether a common role/feature exists among the set of CpG sites for which methylation could be predicted well from SNP genotypes. CpG sites were considered well predicted if the correlation between their predicted methylation and their measured methylation was greater than or equal to 0.5. Similarly, enrichment testing is also used in Chapter 6 to test for a common feature among the set of CpG sites associated with complex disease.

Here, the enrichment tests were conducted using annotations taken from in the Illumina manifest files. As the annotations listed in the 450k chip and EPIC chip manifest files were slightly different, separate enrichment tests were performed for the results obtained using the ARIES data and the results obtained using the Understanding Society data.

For each CpG site, annotations were defined in the following way:

- Genic – The CpG site was tagged to a gene in the manifest file;
- Island – The CpG site was tagged to a CpG island in the manifest file;
- Promoter – The CpG had either the “Promoter_Associated” or the “Promoter_Associated_Cell_type_specific” annotation in the “RegulatoryFeatureGroup” column;

- Enhancer – The “Enhancer” column (450k chip) or the “Phantom5_Enhancers” column (EPIC chip) were not empty for the given CpG site;
- DHS – The “DHS” column (450k chip) or the “DNase_Hypersensitivity_NAME” column (EPIC chip) were not empty for the given CpG site.

In Chapter 5, enrichment of CpG sites with each of the five annotations listed above among the set of well-predicted CpG sites was tested using a two-sided Fisher’s exact test, using the background set of all CpG sites that passed quality control. In Chapter 6, enrichment of CpG sites with each of the five annotations listed above among the set of trait-associated CpG sites identified through MWAS was tested using a two-sided Fisher’s exact test, using the background set of CpG sites that were tested in the MWAS. Odds ratios, 95% confidence intervals and p values were reported for all enrichment tests.

Chapter 3. Comparison of Transcriptome Wide Association Study Software Packages

Since the release of the first TWAS software package, PrediXcan (Gamazon *et al.*, 2015) in 2015, a number of other similar software packages have been released. While these packages enable the user to conduct TWAS analysis, each package conducts the analysis in a slightly different way. Additionally, there has been no formal comparison of these packages on the same data, so the extent to which the slight methodological differences between packages impact the results of TWAS analysis is unclear. In this chapter, I address this problem by applying four TWAS software packages (PrediXcan, MetaXcan, FUSION and SMR) to GWAS data for CD and T1D from WTCCC1, and comparing the results obtained.

The results described in this chapter are an updated version of those presented in (Fryett *et al.*, 2018).

The WTCCC data and Geuvadis data used to perform this analysis are described in detail in Chapter 2.

3.1 TWAS software packages used in the comparison

Details of the general TWAS analysis procedure are given in Chapter 2. Descriptions of the software packages and the prediction models used to perform TWAS are given below:

3.1.1 PrediXcan

PrediXcan analysis consists of two stages. The first stage is the derivation of gene expression prediction models from matched SNP genotype and gene expression data. For each gene, gene expression measures are regressed on genotypes at all SNPs within 1 Mb of either the gene transcription start site or the gene transcription end site using elastic net. The elastic net model is fitted using the R package *glmnet*, with the model tuning parameter α set to 0.5, and the model tuning parameter λ determined by performing a 10-fold cross-validation procedure on the training data,

and choosing the value of λ at which the minimum mean squared error is achieved. This gives a model of the form:

$$y_{ig} \sim \sum_{l=1}^{p} x_{il}\beta_{lg} + \varepsilon_{ig}$$

where y_{ig} is the expression of gene *g* in individual i, x_{il} is the effect allele count (0, 1 or 2) of SNP *l* in individual *i*, β_{lg} is the weight of SNP l on gene g, *p* is the total number of SNPs in the prediction model and ε_{ig} is an error term that includes all non-genetic effects on expression.

During the model training procedure, a cross-validation is performed on the model training data to evaluate how accurately the prediction model predicts gene expression. Prediction models where the correlation between predicted expression and measured expression is significant at a false discovery rate (FDR) of less than 5% were taken forward, with other prediction models discarded.

The prediction models are then applied to the GWAS data to impute gene expression levels into the GWAS samples for which genotype and phenotype have been measured, but gene expression has not. Finally, for each gene, the GWAS phenotype is regressed on the predicted gene expression values, using either simple linear regression (for continuous GWAS phenotypes) or simple logistic regression (for binary GWAS phenotypes), giving an effect size, standard error and p value of predicted expression on phenotype.

Two sets of prediction models were used in the analysis here, both of which were generated by the developers of the PrediXcan software using their standard approach as described on the previous page. The first set of models was derived from matched imputed SNP genotype data (at the set of SNPs present in HapMap2) and normalised gene expression data for 922 individuals from the Depression and Genes Network (DGN) project (Battle *et al.*, 2014). After elastic net regression and cross-validation to determine prediction accuracy, the set of prediction models where correlation between predicted and observed expression was significant at an FDR of 5% were uploaded to predictdb.org. I then downloaded these models and used them in this analysis.

The second set of models was derived from matched imputed SNP genotype data (at the set of SNPs present in 1000 Genomes phase 1) and normalised gene expression data from version 6p of the GTEx project (Battle *et al.*, 2017). 48 sets of gene expression prediction models were generated, each using gene expression data from a different GTEx tissue. After elastic net regression and cross-validation to determine prediction accuracy, the set of prediction models where correlation between predicted and observed expression was significant at an FDR of 5% were uploaded to predictdb.org. I then downloaded these models and used them in this analysis.

3.1.2 MetaXcan

MetaXcan is a summary statistics level adaptation of PrediXcan, which aims to impute the association statistics (beta, standard error, z score and p value) between predicted expression and phenotype that would be produced by PrediXcan, using only GWAS summary statistics and gene expression prediction models. The equation for deriving the association between predicted expression and phenotype as implemented in the MetaXcan package takes the form:

$$Z_g = \sum_{l=1}^{p} w_{lg} \frac{\hat{\sigma}_l}{\hat{\sigma}_g} \frac{\hat{\beta}_l}{se(\hat{\beta}_l)}$$

where Z_g is the z score for the association between the phenotype and the predicted expression of gene g, w_{lg} is the effect size of SNP l on gene g, $\hat{\sigma}_l$ and $\hat{\sigma}_g$ are the estimated variances of SNP l and the predicted expression of gene g respectively, $\hat{\beta}_l$ is the estimated effect size of SNP l on the phenotype of interest, $se(\hat{\beta}_l)$ is the standard error of this beta, and p is the total number of SNPs in the gene expression prediction model.

The values used for these parameters are derived from two main sources. The weights of SNPs on gene expression (w_{lg}) are taken from the gene expression prediction models trained on the reference panel. The estimated effect size ($\hat{\beta}_l$) and standard error ($se(\hat{\beta}_l)$) of SNPs on phenotype are taken from GWAS summary statistics for the phenotype of interest. The estimated SNP variances ($\hat{\sigma}_l$) are calculated from the reference data used to train prediction models. The estimated predicted expression variance ($\hat{\sigma}_g$) is estimated as follows:

$$\hat{\sigma}_g^2 = \boldsymbol{W}'_g \Gamma_g \boldsymbol{W}_g$$

where $\boldsymbol{W}_g$ is the vector of SNP weights (w_{lg}) on the expression of gene g and Γ_g is the sample covariance of X_g, the matrix of genotypes of SNPs that have weights in the prediction models. This matrix is calculated from samples in the model training set.

In attempting to modify the PrediXcan approach to work with GWAS summary statistics rather than individual level genotype and phenotype data, the developers made a number of simplifications. One such simplification was the removal of a term in the calculation of the MetaXcan z score. When the equation for calculating the MetaXcan z score is derived in full, it takes the form:

$$Z_g = \sum_{l=1}^{p} w_{lg} \frac{\hat{\sigma}_l}{\hat{\sigma}_g} \frac{\hat{\beta}_l}{se(\hat{\beta}_l)} \sqrt{\frac{(1 - R_l^2)}{(1 - R_g^2)}}$$

where the Z_g, w_{lg}, $\hat{\sigma}_l$, $\hat{\sigma}_g$, $\hat{\beta}_l$, $se(\hat{\beta}_l)$ and p terms are as previously described, R_l^2 is the proportion of the variance of the phenotype explained by SNP *l*, and R_g^2 is the proportion of the variance of the phenotype explained by predicted expression of gene *g*. The authors of MetaXcan state that correct estimation of R_l^2 and R_g^2 would require information not usually available in GWAS summary data. So to get around this, the $\sqrt{\frac{(1-R_l^2)}{(1-R_g^2)}}$ term is replaced with the value 1, and so is in effect dropped from the equation. This could theoretically cause MetaXcan to give slightly different results to PrediXcan, although the authors of the method state that they anticipate this approximation would have little effect on the estimated z score.

Another such simplification comes from using linear regression properties to derive the MetaXcan formula. The MetaXcan effect size and standard error of predicted gene expression on the phenotype are derived by using known properties of linear regression. This means that even when the phenotype being studied with MetaXcan is binary (and would normally be studied using logistic regression), the estimation of MetaXcan effect size and standard error is done using linear regression properties. Linear and logistic regression typically give similar results when the proportion of cases and controls are similar, yet give increasingly divergent results as the ratio of cases to controls (or vice-versa) increases. In many recent GWAS, such as those

using large population-based biobanks, the numbers of cases and controls are often highly discordant, and so in these situations it could be expected that MetaXcan would give incorrect results. However, given that WTCCC1 data (which have similar number of cases and controls) are used here, this is unlikely to affect the results too much, but it is worth noting as a potential cause of differences.

Here, MetaXcan version 0.5 was used for analysis, using the exact same set of gene expression prediction models derived from GTEX data that were used with PrediXcan.

3.1.3 FUSION

The primary purpose of FUSION is (like MetaXcan) to impute the association statistics between predicted gene expression and a phenotype of interest using only gene expression prediction models and GWAS summary data.

As with PrediXcan and MetaXcan, the first step in the FUSION analysis is the training of gene expression prediction models. For each gene, gene expression data is regressed on genotypes at all SNPs within 500 Kbs of the gene using four methods – BSLMM, BLUP, Elastic net and LASSO. When developing their prediction models, the FUSION developers implemented elastic net using the R package *glmnet*, with the model tuning parameter α set to 0.5, and the model tuning parameter λ determined by performing a 5-fold cross-validation procedure on the training data, and choosing the value of λ at which the minimum mean squared error is achieved. LASSO was implemented using the *–lasso* option in PLINK. The BSLMM was implemented using the GEMMA package, with the default number of MCMC burn-in and sampling iterations, and default priors used. The BLUP was also implemented using GEMMA, with default options used.

A 5-fold cross-validation is then performed on the model training data using each of the four models, and the squared correlation (and a corresponding p value) between predicted and observed expression is calculated.

Following this, the narrow-sense heritability of gene expression attributable to the SNPs within 500 Kbs of the gene is estimated by fitting a genetic relationship matrix and using restricted maximum likelihood analysis as implemented in GCTA. Gene

expression prediction models are then discarded for genes with non-significant (p>0.05) heritability.

The final step is the application of gene expression prediction models to GWAS summary data. For each gene, the prediction model (LASSO, elastic net, BSLMM or BLUP) that achieved the lowest p value in the 5-fold cross-validation stage is taken forward as the model for application. After the best performing model is chosen, SNPs in the prediction model are matched up with the SNPs in the GWAS summary data, and the ImpG-summary algorithm (Pasaniuc *et al.*, 2014) is used to impute GWAS summary statistics at any SNPs that are present in the gene expression prediction model but are missing from the GWAS summary data. Following this, the FUSION test statistic is calculated as:

$$z = \frac{\boldsymbol{W}Z}{\sqrt{\boldsymbol{W}\sum g\,\boldsymbol{W}'}}$$

Where z is the z score for the association between predicted expression of the gene of interest and the phenotype of interest, $\boldsymbol{W}$ is a vector of the effect sizes of the SNPs in the prediction model on expression of the gene of interest, $\boldsymbol{Z}$ is a vector of z scores of SNPs in the prediction model on the phenotype of interest (taken from GWAS summary data), and $\sum g$ is the covariance between SNPs in the prediction model (estimated using data from the model training set).

Here, FUSION version 0.6 was used to conduct analysis. The prediction models used were generated by the FUSION developers from matched imputed SNP genotype data and normalised gene expression data from the GTEX project version 6.

3.1.4 SMR

The SMR method approaches TWAS from the perspective of two-sample Mendelian Randomisation, in which genetic variants are used as instrumental variables to test for a causal relationship between an exposure of interest and an outcome. In the case of SMR, gene expression is used as the exposure, and the user-defined phenotype is used as the outcome.

SMR derives its χ^2 test statistic as:

$$T_{SMR} = \frac{z_{zy}^2 z_{zx}^2}{z_{zy}^2 + z_{zx}^2}$$

Where, z_{zx} is the z score (β divided by its standard error) of association between SNP and gene expression (taken from eQTL studies) and z_{zy} is the z score (β divided by its standard error) of association between SNP and phenotype of interest (taken from GWAS summary statistics).

In practice, this analysis is akin to performing a TWAS in which the gene expression prediction model contains only one SNP, with its weight on gene expression equivalent to the effect size of the SNP as estimated from regressing expression on SNP genotype in an eQTL study.

Here, SMR version 1.03 was used to conduct analysis. eQTL summary statistics and SNP annotation information for GTEx version 6p were downloaded from the GTEx portal website and converted to the SMR BESD file format, before being used in analysis here. The default settings of the package were used.

3.2 Replication of PrediXcan findings from Gamazon et al.

The WTCCC1 data used here for the comparison of software packages underwent similar quality control procedures to the data used in Gamazon *et al.* (2015). To establish that these data were clean and ready to be used, some of the analyses performed in Gamazon *et al.* (2015) were repeated here, and results were compared to those of the original analyses in Gamazon *et al.* (2015). To do this, PrediXcan was applied to the WTCCC CD and T1D GWAS data using gene expression prediction models trained with DGN whole blood data. The original DGN-trained gene expression prediction models used in Gamazon *et al.* (2015) were not available, so a more recently updated version of these models was used.

Overall, similar results were achieved here as in Gamazon *et al.* (2015) (Table 3.1). Of the 8 genes reported to be significantly associated with CD by Gamazon *et al.* (2015), 6 also reached a Bonferroni-corrected significance threshold ($p<5.61 \times 10^{-6}$) here, while the *BSN* and *GPX1* genes just missed this threshold. Of the 29 genes

reported to be significantly associated with T1D by Gamazon *et al.* (2015), 21 also reached a Bonferroni-corrected significance threshold ($p<5.61x10^{-6}$) here, indicating broad agreement between the results observed here and the results found in Gamazon *et al.* (2015), and suggesting that TWAS results can be similar despite small differences in data QC.

Gene	**Phenotype**	**p value in Gamazon et al.**	**p value in this analysis**
ATG16L1	CD	1.94E-10	2.97E-10
IL23R	CD	1.74E-07	2.11E-08
APEH	CD	2.77E-07	2.23E-07
ZNF300	CD	6.29E-07	4.86E-07
NKD1	CD	8.91E-07	4.15E-06
BSN	CD	2.89E-06	9.54E-06
GPX1	CD	3.87E-06	6.11E-06
SLC22A5	CD	5.75E-06	4.86E-06
DCLRE1B	T1D	4.34E-15	2.25E-15
ZNF165	T1D	2.92E-13	1.80E-12
ERBB3	T1D	1.01E-11	8.55E-12
EGFL8	T1D	2.52E-10	3.77E-05
C6orf136	T1D	2.52E-10	1.71E-03
HCG27	T1D	2.52E-10	3.99E-04
GTF2H4	T1D	2.52E-10	1.59E-13
DDR1	T1D	2.52E-10	2.54E-12
AGER	T1D	2.52E-10	1.37E-05
POU5F1	T1D	2.52E-10	3.22E-05
ATP6V1G2	T1D	2.52E-10	4.96E-08
TUBB	T1D	2.52E-10	9.58E-04
AIF1	T1D	2.52E-10	4.19E-06
CYP21A2	T1D	2.52E-10	8.27E-44
LSM2	T1D	2.52E-10	5.82E-16
VARS2	T1D	2.52E-10	1.43E-09
APOM	T1D	2.52E-10	1.57E-30
DDAH2	T1D	2.52E-10	3.73E-17
NCR3	T1D	2.52E-10	4.21E-31
ZSCAN16	T1D	7.37E-10	4.77E-10
ZKSCAN4	T1D	7.73E-10	8.39E-10
PTPN22	T1D	5.41E-09	5.58E-10
RPS26	T1D	6.00E-09	1.67E-08
GDF11	T1D	9.11E-09	3.45E-08
SUOX	T1D	4.49E-08	4.00E-08
BTN3A2	T1D	3.30E-07	2.87E-07
PRSS16	T1D	1.34E-06	4.24E-05
FAM109A	T1D	1.94E-06	6.46E-07
SH2B3	T1D	3.05E-06	5.92E-06

Table 3.1. P values for genes significantly associated with CD or T1D from Gamazon et al., and their p values in this analysis.

3.3 Comparison of TWAS software packages using Geuvadis data

To establish which of the TWAS software packages most accurately predicted gene expression from SNP genotypes, gene expression prediction models from the PrediXcan and FUSION packages were applied to data from the Geuvadis project. While the expected input for FUSION is GWAS summary data, the developers of FUSION provide an R script that allows the FUSION gene expression prediction models to be applied to individual level genotype data to obtain predicted gene expression values. As the MetaXcan and SMR packages can only use GWAS summary data as its input, they were not included in this analysis. For each package, prediction models trained using gene expression data from LCLs were applied to Geuvadis genotype data to predict gene expression. The squared correlation between the predicted expression levels and the measured levels was then calculated for each gene.

Prediction accuracy (squared correlation) estimates achieved by prediction models from the two software packages were highly correlated and concordant (Figure 3.1), indicating that both packages predicted gene expression with similar accuracy. For many genes, the prediction accuracy achieved by each package was poor, indicating that these methods do not predict gene expression well for many genes. This point will be explored further in Chapter 4.

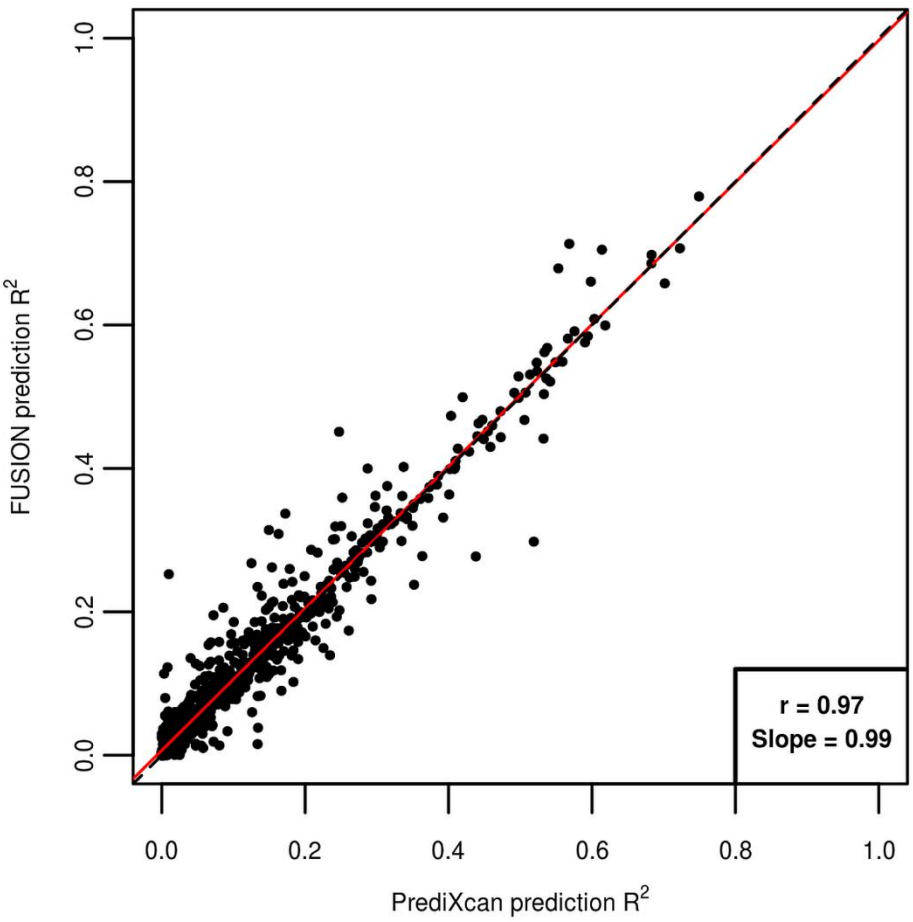

Figure 3.1. Comparison of PrediXcan and FUSION from application to Geuvadis data. Each point represents a single gene, and displayed are the squared Pearson correlation coefficients between measured Geuvadis expression and expression predicted by PrediXcan prediction models (x axis) and the squared Pearson correlation coefficient between measured Geuvadis expression and expression predicted by FUSION prediction models (y axis). The dotted line is the line of equality, and the solid red line is a best fit line. The correlation between x and y values and the slope of the best fit line are shown in the bottom right corner.

3.4 Comparison of TWAS software packages using WTCCC1 data

To compare how the different TWAS software packages performed at detecting associations between predicted expression and phenotype, the packages were next applied to GWAS data for CD and T1D from WTCCC1. For PrediXcan, individual level genotype and phenotype data were used as the input to the package. For the remaining packages, GWAS summary statistics were used as the input. To create GWAS summary statistics for the WTCCC1 data, a GWAS of each phenotype was conducted using the SNPTEST software package, using the *–frequentist 1 –method score* options to conduct logistic regression under an additive genetic model. As the genotype and phenotype data used for input to the PrediXcan software package

were not adjusted for any covariates, no covariates were used in the GWAS of the WTCCC1 CD and T1D data to avoid differences in the input for the different software packages. The resulting summary statistics for each SNP were then used as input for the MetaXcan, FUSION and SMR software packages. The z scores for the association between predicted gene expression and phenotype produced by each of the TWAS packages were then compared.

When applied to CD data, the software packages produced broadly similar results (Figure 3.2). Both PrediXcan and MetaXcan identified associations on chromosomes 3, 5 and 17, while the FUSION and SMR packages identified associations approaching significance on chromosomes 3 and 5. However, the genes identified as significant were not the same for each software package. For example, on chromosome 5 the PrediXcan and MetaXcan packages both identified significant associations between CD and predicted expression of *IRGM* (PrediXcan $p = 2.77 \times 10^{-8}$, MetaXcan $p = 2.83 \times 10^{-8}$), while the result for this gene from the SMR package did not reach significance (SMR $p = 1.41 \times 10^{-4}$), and this gene was not tested by the FUSION package, as there was no prediction model for this gene available in the FUSION package. In total, 6160 genes were tested by the PrediXcan and MetaXcan packages, 3293 genes were tested by SMR and 2041 genes were tested by FUSION, indicating that the PrediXcan and MetaXcan packages tested the broadest range of genes.

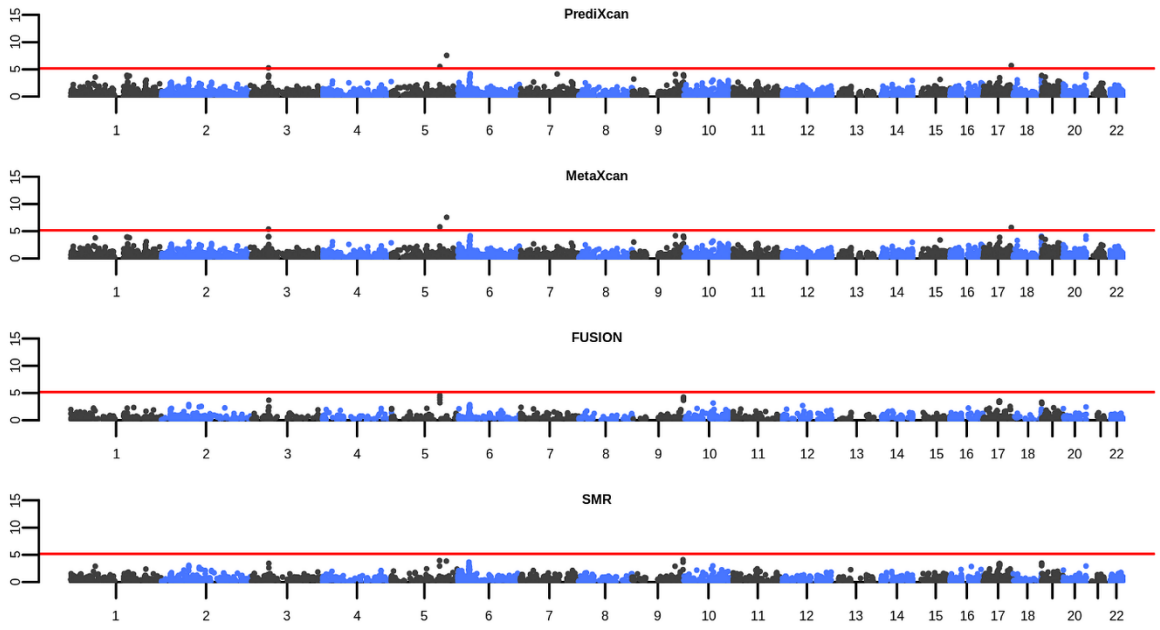

Figure 3.2. Comparison of results from applications of four TWAS methods to imputed WTCCC1 CD data. Manhattan plots showing *p* values of associations between predicted expression and CD from applications of PrediXcan, MetaXcan, FUSION and SMR to imputed WTCCC1 CD data using prediction models trained in GTEx whole blood data. *P* values are plotted against the transcription start site for each gene. The red line on each plot shows the Bonferroni-corrected significance threshold at 6.87×10^{-6}.

Again, similar results were achieved by all packages when applied to the T1D data. All software packages identified associations in the same regions on chromosomes 6 and 12, while the PrediXcan and MetaXcan packages detected an additional association on chromosome 16 that was not found by either the FUSION or SMR package (Figure 3.3). As observed in the analysis of CD data, some of these differences were because each software package tested a different set of genes. For example, both PrediXcan and MetaXcan found a significant association between T1D and predicted expression of *CLEC16A* on chromosome 16 (PrediXcan $p = 1.57 \times 10^{-7}$; MetaXcan $p = 1.46 \times 10^{-7}$), but this gene was not tested by either the FUSION or the SMR package.

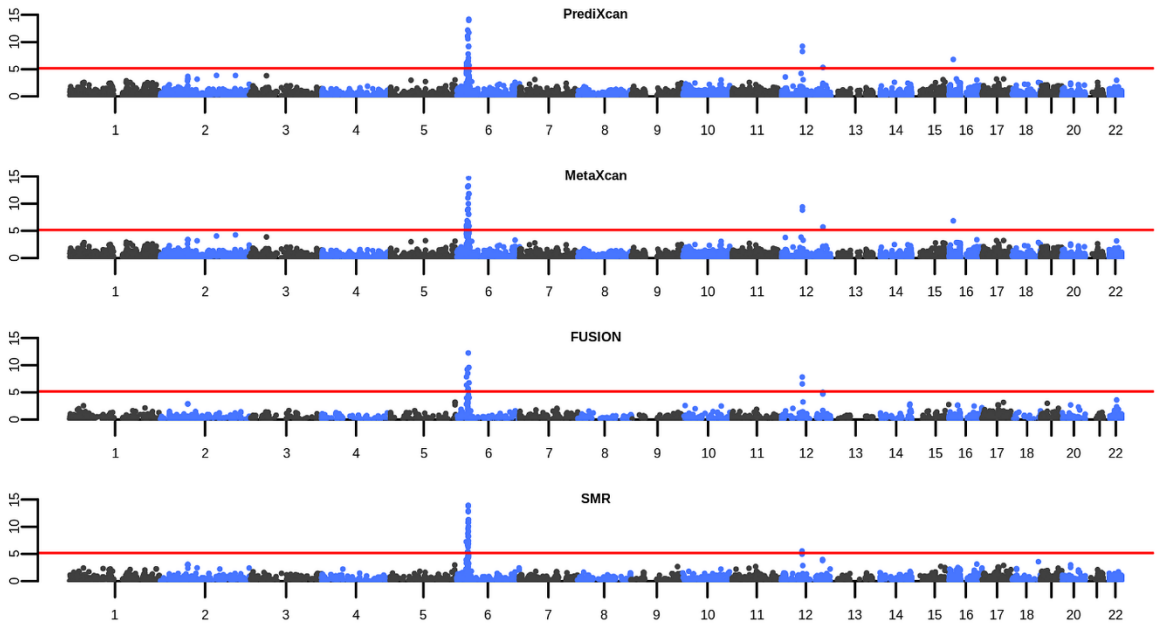

Figure 3.3. Comparison of results from applications of four TWAS methods to imputed WTCCC1 T1D data. Manhattan plots showing *p* values of associations between predicted expression and T1D from applications of PrediXcan, MetaXcan, FUSION and SMR to imputed WTCCC1 T1D data using prediction models trained in GTEx whole blood data. *P* values are plotted against the transcription start site for each gene. The red line on each plot shows the Bonferroni-corrected significance threshold at 6.87×10^{-6}.

For the set of genes tested by all four software packages, the z scores produced by the packages were highly correlated in both the CD (Figure 3.4) and T1D (Figure 3.5) analyses, indicating broad agreement between methods. As expected, PrediXcan and MetaXcan produced near-identical results, while the FUSION and SMR packages produced similar, but slightly more different results to those from PrediXcan and MetaXcan. Despite the broad similarity, there were some genes for which the different packages obtained strongly discordant results. One such gene was *HLA-DQA1*, the predicted expression of which was found to be significantly associated with T1D by all four software packages, but for which PrediXcan and MetaXcan found strongly positive z scores (PrediXcan z score = 7.80; MetaXcan z score = 7.04), while SMR and FUSION obtained strongly negative z scores (SMR z score = -6.49; FUSION z score = -17.34). Interestingly, the 20 genes for which the largest pairwise difference between results of the different packages were achieved in the T1D analysis were all located in the MHC region.

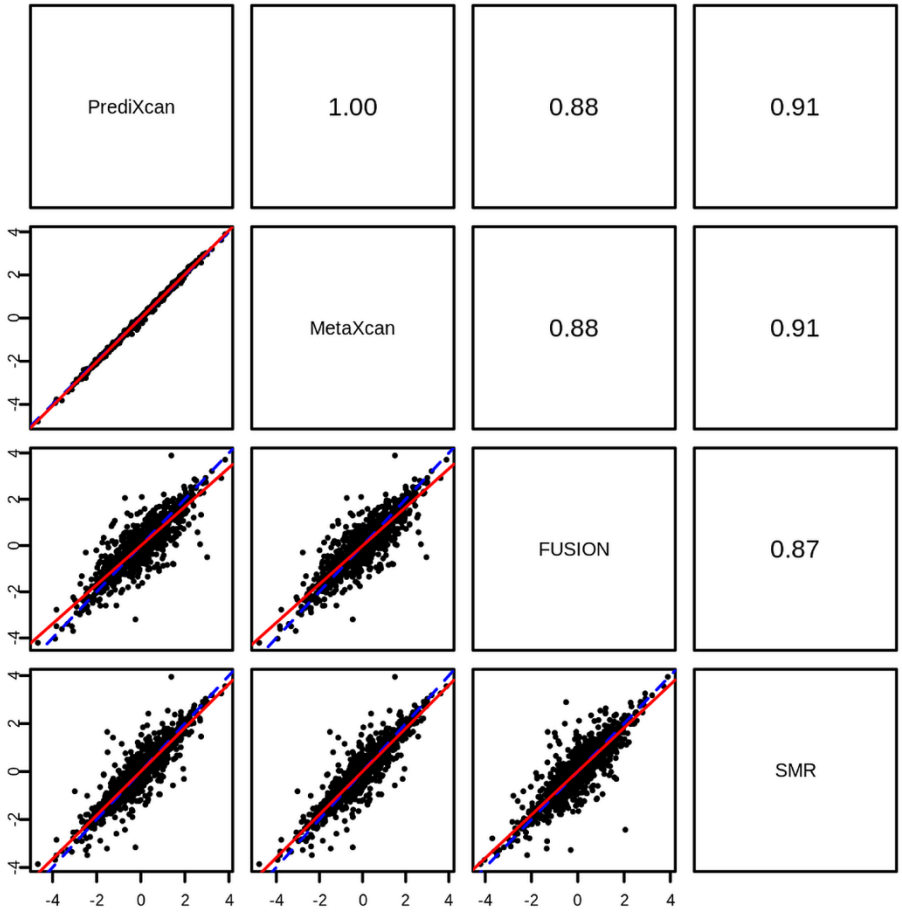

Figure 3.4. Comparison of results for genes tested by all four TWAS packages when applied to WTCCC1 CD data. Lower panels show scatter plots of z scores obtained for the genes tested by all four packages when applied to WTCCC CD GWAS data. The red line is a best fit line, and the blue dashed line is the line of equality (y=x). Upper panels show the pairwise Pearson correlation coefficients between the z scores obtained by the four packages for the set of genes.

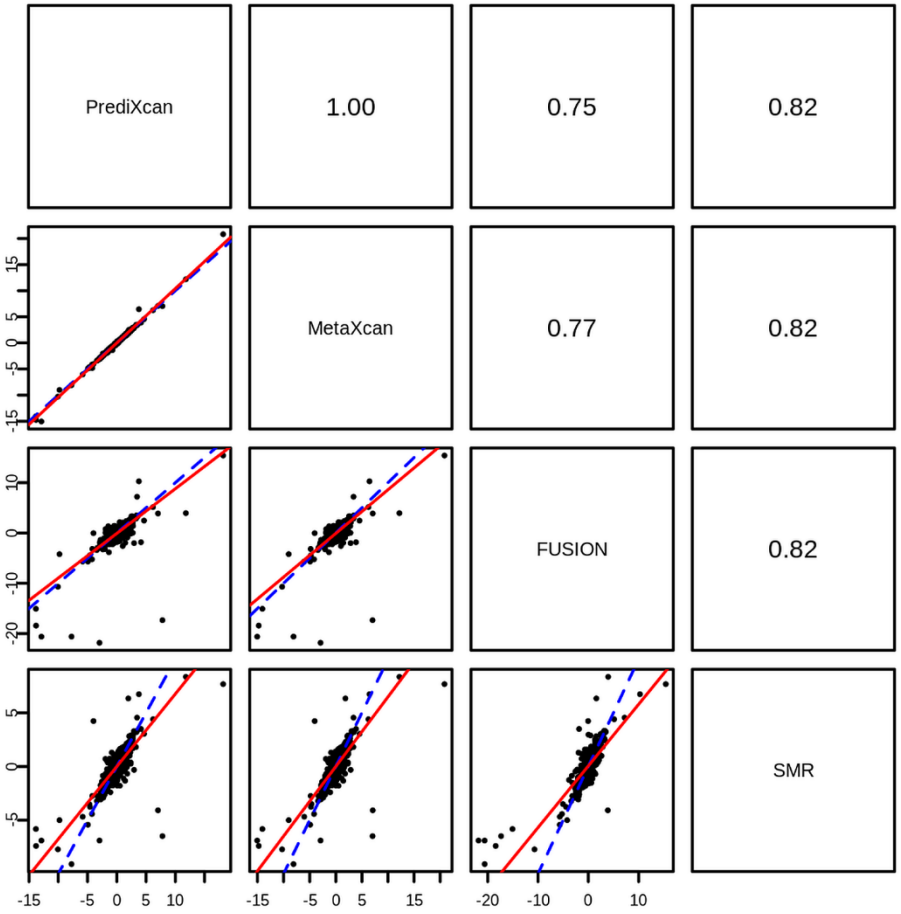

Figure 3.5. Comparison of results for genes tested by all four TWAS packages when applied to WTCCC1 T1D data. Lower panels show scatter plots of z scores obtained for the genes tested by all four packages when applied to WTCCC T1D GWAS data. The red line is a best fit line, and the blue dashed line is the line of equality (y=x). Upper panels show the pairwise Pearson correlation coefficients between the z scores obtained by the four packages for the set of genes.

One methodological difference between the packages that could cause differences in their results is the way in which missing SNP data is dealt with. For the PrediXcan and MetaXcan packages, SNPs that are present in a gene expression prediction model but not in the GWAS data (for which expression is being predicted) are considered missing and are not used for prediction. Conversely for the FUSION method (when using summary level GWAS data as input), GWAS summary data are imputed at the SNPs present in the prediction model but missing from the GWAS data using the ImpG-summary algorithm, and these imputed summary statistics are used for prediction. To examine whether the observed z score differences across the genes tested by all methods may have been influenced by SNP missingness in the PrediXcan and MetaXcan analyses, the proportion of each MetaXcan gene

expression prediction model's SNPs that were missing from the WTCCC1 genotype data was calculated. Genes were then sorted according to the difference between their MetaXcan and FUSION z scores and placed into ten bins of equal size, with genes showing the largest difference placed into bin ten, and those with the smallest difference placed into bin one. For each bin, the average proportion of the SNPs in the MetaXcan prediction models that were missing from the WTCCC1 genotype data was then calculated. Overall, the bin corresponding to genes with the greatest z score differences also showed the highest average SNP missingness (Figure 3.6), indicating that SNP missingness may have caused some of the differences observed here.

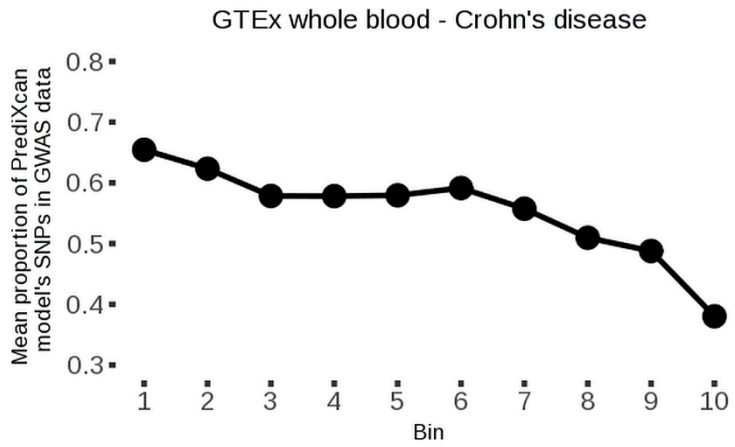

Figure 3.6. SNP missingness versus bin (denoting difference between MetaXcan and FUSION z scores) for MetaXcan prediction models when applied to the WTCCC1 CD data.

3.5 Comparison of TWAS analysis results across different tissues

The developers of the TWAS software packages tested here have each released sets of prediction models trained using data from different GTEx tissues. To compare how the prediction models derived from data from different tissues perform, the PrediXcan software package was applied to the WTCCC1 GWAS data for CD and T1D using gene expression prediction models for a range of tissues. For CD, prediction models trained using data from GTEx whole blood, GTEx EBV-transformed lymphocytes and GTEx sigmoid colon tissue were applied. For T1D,

prediction models trained using data from GTEx whole blood, GTEx EBV-transformed lymphocytes and GTEx pancreas tissues were applied. These sets of gene expression prediction models were chosen for their relevance to the trait studied.

Overall, the gene expression prediction models trained using data from different GTEx tissues produced similar results when applied to the WTCCC CD GWAS data (Figure 3.7). While there were few associations detected at a Bonferroni-corrected significance threshold in any tissue, prediction models from all three tissues identified suggestive associations on chromosomes 3 and 6. For the set of genes for which a prediction model was available for all three GTEx tissues tested here, the resulting z scores showed a mildly positive correlation (all pairwise correlations between 0.49 and 0.56) (Figure 3.8). Despite these similarities, there were some observable differences between the results from application of models for the different tissues. As seen in the comparison of different software packages, many of the differences between the results for different tissues were related to the set of genes tested for a given tissue. For example, association between *SLC22A5* predicted expression and CD approached significance when using models trained with data from GTEx whole blood ($p = 3.07 \times 10^{-6}$) and GTEx EBV-transformed lymphocytes ($p = 3.85 \times 10^{-5}$), yet there was no prediction model available for this gene in GTEx sigmoid colon, explaining why no suggestive association was detected on chromosome 5 by the GTEx sigmoid colon models.

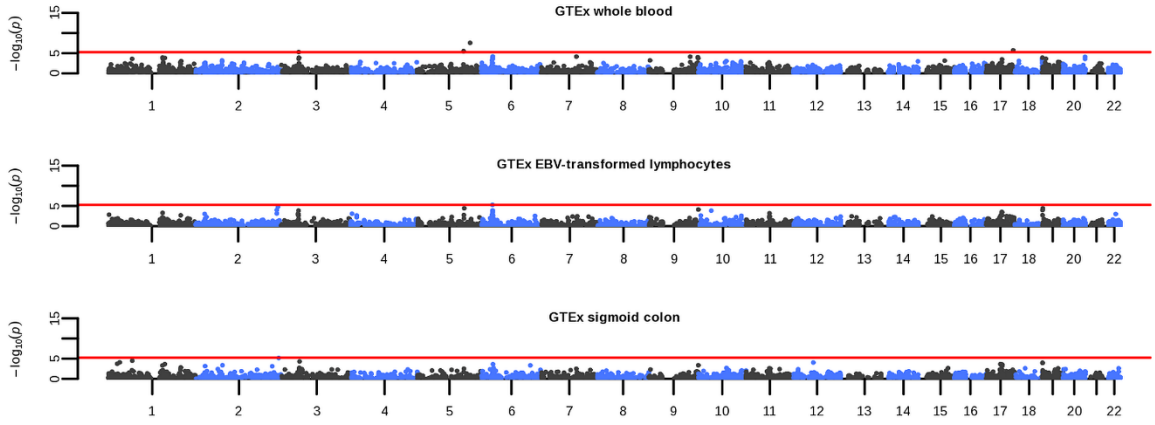

Figure 3.7. Results from applications of PrediXcan to WTCCC CD data using prediction models based on 3 tissues. Manhattan plots showing p values of associations between predicted expression and CD from applications of PrediXcan to imputed WTCCC1 CD data, using prediction models trained

in GTEx data for different tissues. P values are plotted against the transcription start site for each gene. The red line on each plot shows the Bonferroni-corrected significance threshold at (a) 5.78×10^{-6}.

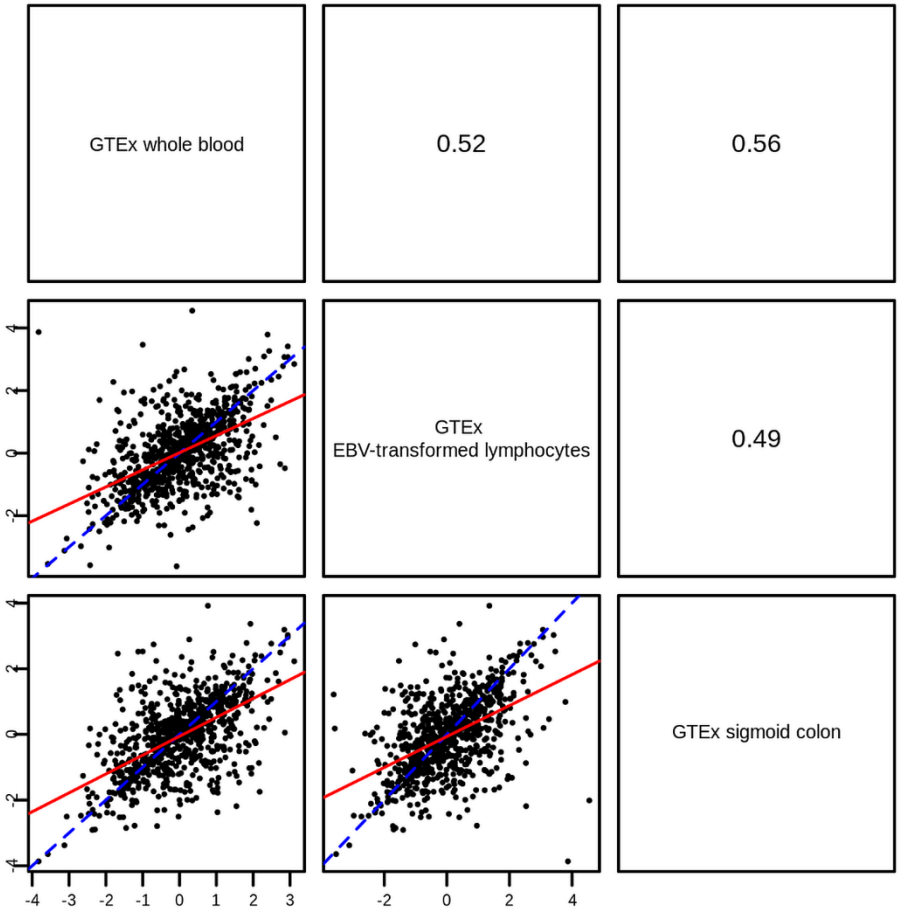

Figure 3.8. Comparison of results for genes tested in all three tissues from application of PrediXcan to WTCCC1 CD GWAS data. Lower panels show scatter plots of z scores obtained for the genes tested in all three GTEx tissues when applied to WTCCC CD GWAS data. The red line is a best fit line, and the blue dashed line is the line of equality (y=x). Upper panels show the pairwise Pearson correlation coefficients between the z scores obtained for the three tissues.

As with the application to CD data, gene expression prediction models trained using data from different GTEx tissues produced similar results when applied to the WTCCC T1D data (Figure 3.9), with all tissues detecting significant associations on chromosomes 6 and 12. For the genes tested in all three tissues, z scores from each tissue were positively correlated, with all pairwise tissue-tissue correlations between 0.56 and 0.60 (Figure 3.10). The largest difference between results from different

tissues were for genes located in the MHC region. Other differences were again related to which set of genes was tested for each tissue. For example, a significant association between T1D and predicted expression of *CLEC16A* on chromosome 16 was detected using prediction models trained with GTEx whole blood data ($p = 1.57x10^{-7}$), yet there was no prediction model available for this gene for either the GTEx EBV-transformed lymphocytes or the GTEx pancreas, meaning that only GTEx whole blood models detected an association on chromosome 16.

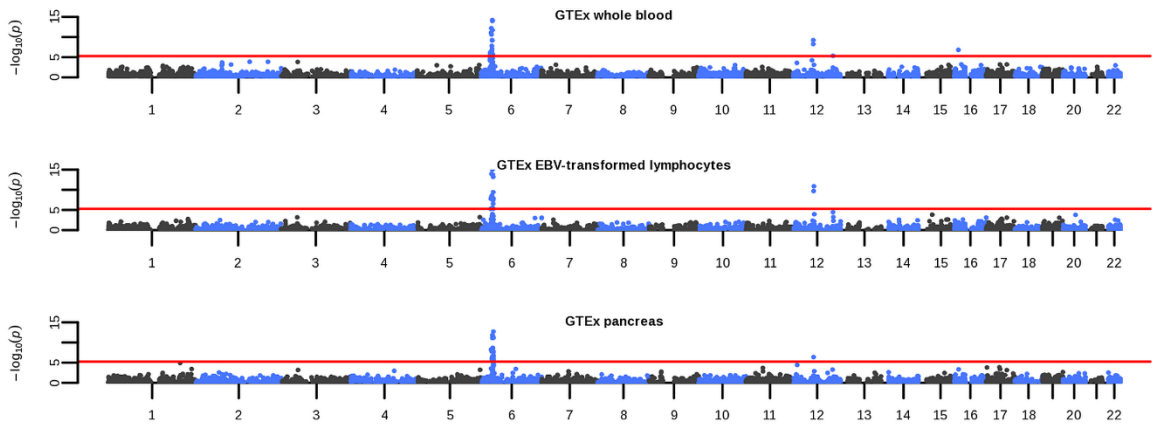

Figure 3.9. R**esults from applications of PrediXcan to WTCCC T1D data using prediction models based on 3 GTEx tissues.** Manhattan plots showing p values of associations between predicted expression and T1D from applications of PrediXcan to imputed WTCCC1 T1D data, using prediction models trained in GTEx data for different tissues. P values are plotted against the transcription start site for each gene. The red line on each plot shows the Bonferroni-corrected significance threshold at (a) 5.51×10^{-6}.

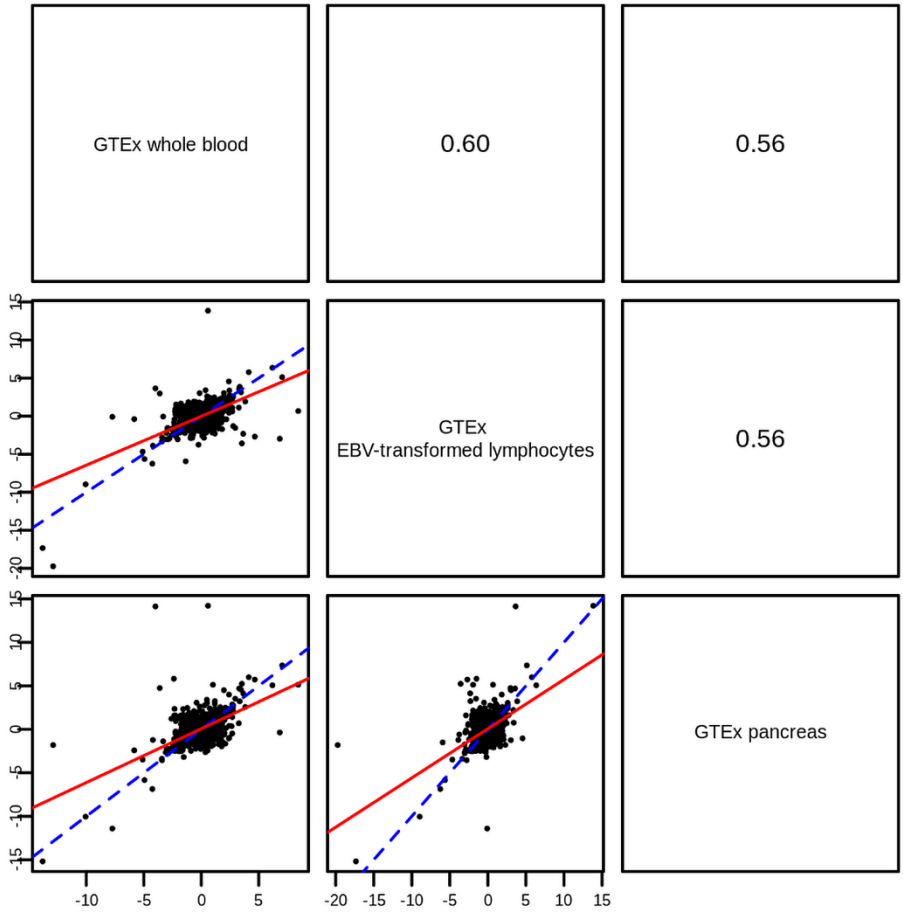

Figure 3.10. Comparison of results for genes tested in all three tissues from application of PrediXcan to WTCCC1 T1D GWAS data. Lower panels show scatter plots of z scores obtained for the genes tested in all three GTEx tissues when applied to WTCCC T1D GWAS data. The red line is a best fit line, and the blue dashed line is the line of equality (y=x). Upper panels show the pairwise Pearson correlation coefficients between the z scores obtained for the three tissues.

3.6 Comparison of TWAS analysis results with those from GWAS

While TWAS methods have been suggested as a complementary approach to GWAS, they can theoretically discover additional risk loci not found through GWAS. To investigate how these two approaches compare with respect to detection and localisation of associations, a GWAS was conducted with each of the imputed WTCCC1 CD and T1D datasets. As expected, most of the significantly associated genes found through TWAS analysis (Figures 3.7 and 3.9) were located in the same genomic loci as the observed GWAS hits (GWAS Bonferroni significance threshold $p<5\times10^{-8}$) (Figure 3.11), with only 2 loci significantly associated with predicted

expression not identified through GWAS. In contrast, 9 of the 14 loci that attained genome-wide significance for either CD or T1D through GWAS showed no significant association signal for predicted expression, implying that TWAS may not be as powerful for the discovery of new associations as GWAS, and reinforcing its role as being complementary to (rather than a replacement for) GWAS.

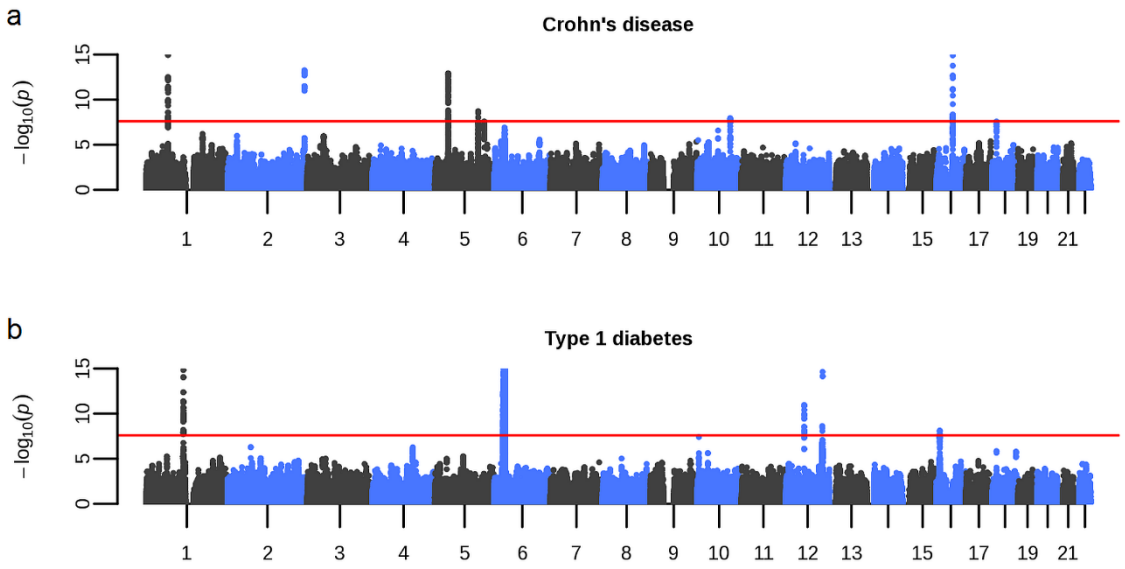

Figure 3.11. Manhattan plots of GWAS of (a) WTCCC CD data and (b) T1D data, with the Bonferroni-corrected significance threshold at (a) 9.12 x 10^{-9} for CD and (b) 9.11 x 10^{-9} for T1D.

3.7 Application of MetaXcan to more recent CD and T1D genome-wide meta-analysis data

So far, GWAS data from WTCCC1 have been used to perform most of the comparisons between the different TWAS software packages. However, these data are now approximately fifteen years old and their sample sizes are relatively small compared to those used in more recent GWAS, meaning that novel associations are unlikely to be detected using these data. In an attempt to find more novel associations between predicted expression and phenotypes, TWAS was next performed using publicly available summary statistics from more recent, better powered GWAS of CD (Liu *et al.*, 2015a) and T1D (Cooper *et al.*, 2017).

In the comparisons performed so far, the TWAS software packages performed similarly at prediction of gene expression and at detection of associations between predicted expression and phenotype. However, as the PrediXcan and MetaXcan packages tested more genes than the other software packages, and because the imputation of missing GWAS summary data by FUSION could result in uncertainty in the gene expression prediction, for the analyses in this section, the MetaXcan package was chosen to take forward.

Thus, MetaXcan was applied to summary statistics from a recent meta-analysis of CD comprising 5,956 CD cases and 14,927 controls using three sets of gene expression prediction models, each trained using a different GTEx tissue – GTEx whole blood, GTEx EBV-transformed lymphocytes and GTEx sigmoid colon.

In total, 54 significant associations between predicted expression and CD were detected at a Bonferroni-corrected significance threshold of $p<5.15x10^{-6}$ (Figure 3.12). This included 31 associations with predicted whole blood expression, 13 associations with predicted EBV-transformed lymphocyte expression and 10 associations with predicted sigmoid colon expression. Of these 54 significant associations, 45 were for genes that had previously been suggested as disease-relevant in previous CD GWAS and meta-analyses, 7 were for genes located in CD risk loci identified in earlier CD GWAS, but had not been suggested as potential effector genes, and 2 (*NPIPB6* and *NPIPB7*) were in a previously undiscovered CD risk locus.

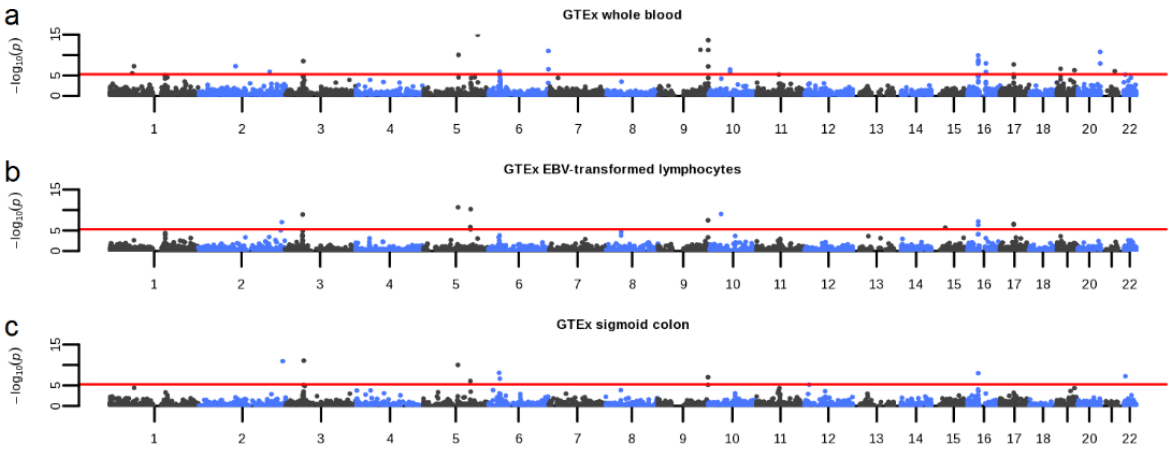

Figure 3.12. Application of MetaXcan to summary statistics from a meta-analysis of CD using prediction models for three tissues. Manhattan plots showing *p* values of associations between

predicted expression and CD from applications of MetaXcan to summary statistics from a CD meta-analysis using prediction models trained in **a** GTEx whole blood data, **b** GTEx EBV-transformed lymphocytes and **c** GTEx sigmoid colon. *P* values are plotted against the transcription start site for each gene. The red line on each plot shows the Bonferroni-corrected significance threshold at 5.15×10^{-6}

MetaXcan was also applied to summary statistics from a recent meta-analysis of T1D comprising 5,913 T1D cases and 8,829 controls using three sets of gene expression prediction models, each trained using a different GTEx tissue - GTEx whole blood, GTEx EBV-transformed lymphocytes and GTEx pancreas.

In total, 154 significant associations between predicted expression and T1D were detected at a Bonferroni-corrected significance threshold of $p<5.01x10^{-6}$ (Figure 3.13). This included 63 associations with predicted whole blood expression, 47 associations with predicted EBV-transformed lymphocyte expression and 44 associations with predicted pancreas expression. Most of the significant associations were located in risk loci previously implicated in T1D through GWAS, including the MHC region (in which 119 of the 154 associations were located) and 12q13, showing how important these regions are in T1D.

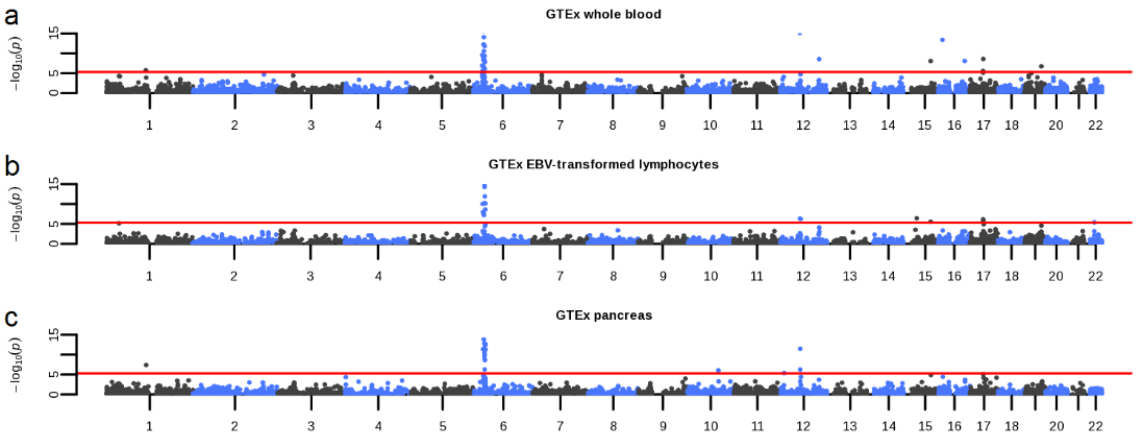

Figure 3.13. Application of MetaXcan to summary statistics from a meta-analysis of T1D using prediction models for three tissues. Manhattan plots showing *p* values of associations between predicted expression and T1D from applications of MetaXcan to summary statistics from a T1D meta-analysis using prediction models trained in **a** GTEx whole blood data, **b** GTEx EBV-transformed lymphocytes and **c** GTEx pancreas. *P* values are plotted against the transcription start site for each gene. The red line on each plot shows the Bonferroni-corrected significance threshold at 5.01×10^{-6}.

3.8 Discussion

Overall, the software packages tested here predicted Geuvadis gene expression with similar accuracy and identified associations between predicted gene expression and phenotype in the same genomic regions when applied to the WTCCC1 GWAS data, indicating broad similarity between packages.

Despite the packages tending to identify associations in similar regions of the genome, these associations were not necessarily for the same genes. In fact, most of the differences between the results from each software packages were due to each of the software packages testing a different set of genes. This may be related to the manner in which the developers of each package selected which genes to use in their analysis. PrediXcan/MetaXcan prediction models were created for all genes for which expression was measured in GTEx, but only models where the correlation between predicted and observed expression (as determined by a cross-validation procedure) was significant at a false discovery rate of less than 5% were made available. FUSION prediction models were created and uploaded for all genes where the heritability of gene expression attributable to SNPs local to the gene in question was significant (at $p<0.05$). In this application of SMR, the default settings of the package were used which resulted in the package testing only the genes for which there was a significant eQTL (at $p<5x10^{8}$) in the GTEx data. Given that these are all measures of the strength of the relationship between genotype and expression, it is perhaps slightly concerning that they result in such different sets of genes being made available. Nevertheless, there existed a set of genes that were tested by all methods, and which could be used for comparative purposes.

When looking at the set of genes tested by all methods, all methods performed similarly to each other, with z scores highly correlated and concordant between all methods in both the CD and T1D analyses. Reassuringly, PrediXcan and MetaXcan showed near perfect concordance, corroborating the result shown in (Barbeira *et al.*, 2018). This is unsurprising, given that the same set of gene expression prediction models are used for the two packages, and also given that the differences between PrediXcan and MetaXcan would be unlikely to manifest when applied to the WTCCC data. It is possible that if the packages were applied to data for a case/control

phenotype with a more unequal ratio of cases to controls, greater differences between the packages might be expected. Results for these genes were also similar for the FUSION and SMR packages, again corroborating the results shown in (Barbeira *et al.*, 2018) and suggesting that software users would obtain similar results regardless of which package was chosen.

However, despite these similarities, the packages obtained discordant results for some genes. One factor that could have caused this is the way in which missing SNP data is treated by the different software packages. For the PrediXcan and MetaXcan packages, SNPs that are in the gene expression prediction models but that are missing from the GWAS data are ignored during prediction. However, the FUSION package imputes the GWAS summary statistics at these missing SNPs using the ImpG-summary algorithm and then uses them for gene expression prediction. Indeed, some of the largest differences were observed in instances where more SNPs were missing from the PrediXcan and MetaXcan models. While it may seem attractive to impute missing data, the reason the data were missing was because the SNPs were poorly imputed in the genotype imputation of WTCCC data, and were excluded in the subsequent quality control. As this feature could not be switched off, PrediXcan and MetaXcan seemed more compelling as their predictions used only the most reliable SNPs. It is worth noting that since this analysis was performed, additional sanity checks related to the use of the Imp-G summary algorithm have been added to the FUSION package. This includes a summary statistics imputation quality filter, and a filter that removes genes where more than a given proportion (usually 50%) of the gene expression prediction model's SNPs are missing. These additional filters may help to mitigate some of the issues identified here.

Another factor that may have resulted in the discordances observed here was the location of the genes. The largest discordances were observed for tests of association between T1D and predicted expression of genes located in the MHC region. As there are known to be strong effects of variants located in the MHC on T1D, it is possible that the strength of these effects amplified the differences between the packages, making them larger than would otherwise be expected.

Following the comparison of TWAS software packages, a TWAS was performed using gene expression prediction models for different GTEx tissues. As with the comparison of software packages, many differences between the results for different

tissues were attributable to a different set of genes being tested for each tissue. For the set of genes tested in all tissues examined here, TWAS results were strongly correlated. Given that genetic effects on gene expression are known to be similar across the GTEx tissues (Mele *et al.*, 2015) (Battle *et al.*, 2017), this is unsurprising. This suggests that associations can be detected even when not using the true causal tissue of interest for the trait. For those genes tested across multiple tissues, it also raises the possibility of multivariate testing using known correlations between expression from different tissues, as implemented in the MultiXcan package (Barbeira *et al.*, 2019).

Although TWAS has sometimes been sold as a method for the discovery of new genetic risk loci not identified through GWAS alone, due to the reduced multiple testing burden and the summing of multiple SNPs within a single test, application of TWAS to the WTCCC data here found only two loci missed by GWAS, and did not identify many of the loci found through GWAS. This is to be expected, as TWAS should only identify associations where SNPs act on the phenotype by affecting gene expression, whereas GWAS should identify both these associations and those at which a SNP changes the protein code. Notably, the two “new” loci found through TWAS contained SNPs that reached a suggestive significance threshold of $p<1e-05$ in the GWAS. This reinforces the role of TWAS as complementary to, rather than a replacement for, GWAS, and suggests that the real utility of TWAS is not necessarily the discovery of new risk loci, but the identification of the potential risk genes within previously known loci. Theoretically, TWAS has more power to detect an association than GWAS when multiple SNPs affecting the expression of the same gene (with the same direction of effect) are each independently associated with the phenotype. From the results observed here, there were no obvious instances of this occurring in regions that would not be identified through GWAS. It is possible that instances of this could be detected in other phenotypes not considered here, or in CD or T1D if larger sample sizes were used in the GWAS analysis.

In an attempt to identify more associations than were detected using the WTCCC1 GWAS data, TWAS was applied to summary statistics from larger, more recent GWAS of CD and T1D. For CD, 54 associations between predicted gene expression and disease status were identified, including at genes that have been previously suggested to be involved in CD (*SLC22A5*, *IRGM* and *ATG16L1*), reinforcing the role of these genes in CD. Additionally, some genes that have never previously been

suggested were identified (*NPIPB6* and *NPIPB7*), although the role of these genes is unknown, and further research will be required to determine if they truly play a role in CD. Among the other interesting associations identified were *ETS2* and *ICAM1*, of which the gene expression has been implicated in CD (van der Pouw Kraan *et al.*, 2009; de Lange *et al.*, 2017). Interestingly, an application of FUSION to the same GWAS summary data identified similar results to those found here (Mancuso *et al.*, 2017), giving additional confidence to the results found here and suggesting that the similarities between TWAS software packages that we observed when using the WTCCC1 data hold true when using data other than the WTCCC1 data. For T1D, 154 associations were identified, mostly in regions known to be heavily involved in T1D risk. Further investigation will be required to identify the true causal genes within these risk regions.

I conclude by drawing attention to some caveats of this analysis. It is worth noting that most of this analysis was performed in 2016 and 2017, and a number of other TWAS software packages that are not included in this comparison have since been released. These packages include TIGAR (Nagpal *et al.*, 2019), CoMM (Yang *et al.*, 2019a) and its summary statistics based extension CoMM-S2 (Yang *et al.*, 2019b) and SLINGER (Vervier and Michaelson, 2016). However, neither the TIGAR nor CoMM software packages make any of their prediction models, or the data that could be used to train prediction models, publicly available. As the aim of this project was to compare the software packages using their publicly available prediction models, and no such models were provided for these packages, they could not be included. Additionally, the developers of the SLINGER package only developed and released prediction models trained using data from the DGN project, which would not have been suitable for comparison with the other software packages, and so SLINGER was not included in the comparison. These packages may be worth revisiting in the future if gene expression prediction models trained using data from the GTEx project are released in the future.

Another caveat of this analysis is that real GWAS data were used to compare the methods, meaning the true effects of gene expression on the phenotype are unknown. This means there is no guarantee that the associations detected by each method are real, and that it cannot be determined which method identified the most “correct” set of associations. Indeed, the associations in the MHC region identified by all the methods seem unlikely to be true, as it is widely thought that variation at

coding SNPs is responsible for the effects of genes in the MHC on autoimmune disease observed in this genomic region. Due to the high levels of linkage disequilibrium observed across the MHC, it is possible that genotypes at SNPs affecting the expression of genes in the MHC region are highly correlated with genotypes at coding SNPs, resulting in TWAS associations being detected.

Chapter 4. Investigation of Factors Affecting Prediction Accuracy in Transcriptome Imputation

A crucial aspect of TWAS that has received little attention in the literature is the issue of prediction accuracy. As seen in Chapter 3, the accuracy with which gene expression can be predicted from SNP genotype data appears to be low for many genes. As the accuracy with which gene expression can be predicted from SNP genotypes is related to the power in the subsequent test of association between predicted expression and phenotype, this issue is important. To investigate this issue further, here I carry out a 10-fold nested cross-validation experiment using seven different statistical approaches to identify the best approach for model building. I then investigate how a number of factors related to the data used to train and test gene expression prediction models can affect the accuracy of gene expression prediction.

The results described in this chapter are those described in (Fryett *et al.*, 2020)

4.1 Description of statistical methods being compared

The ability of seven different statistical approaches for the prediction of gene expression from local SNP genotype data was compared. These different methods are described below:

4.1.1 Ridge regression

Ridge regression (Hoerl and Kennard, 2000) allows for coefficients in the prediction model to be shrunk by applying an L2 penalty. When using ridge regression, coefficients can be shrunk to near zero, but not to exactly zero, meaning that no variable selection is performed and that a polygenic model (where many SNPs each have a small effect) is produced.

Ridge regression aims to minimise:

$$(y - X\beta)^2 + \lambda \sum_{j=1}^{p} {\beta_j}^2$$

where *y* is the gene expression, *X* is a matrix of SNP genotypes, *β* represents regression coefficients, and *λ* is the regularisation parameter. Here, ridge regression was implemented using the cv.glmnet function in the *glmnet* package, with the value of λ determined by 10-fold cross-validation.

4.1.2 LASSO

The LASSO (Tibshirani, 1996) allows for coefficients in the prediction model to be shrunk by applying an L1 penalty when fitting the model. When using LASSO, the model coefficients can be shrunk to exactly zero, which means that LASSO tends to produce a sparse model in which there are relatively few SNPs with non-zero effect sizes, but that these effect sizes tend to be quite large. The LASSO aims to minimise:

$$(y - X\beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

Where *y*, *X*, *β* and *λ* are defined in the same way as for ridge regression in section 4.1.1. Here, LASSO was implemented using the cv.glmnet function in the *glmnet* R package, with the value of λ determined by 10-fold cross-validation.

4.1.3 Elastic net

The elastic net (Zou and Hastie, 2005) is a mixture model that uses both the L1 and L2 penalties that are used by LASSO and ridge regression respectively. The degree of mixture between these two penalties is determined by the value of the *α* parameter, and so the degree of shrinkage applied to the model coefficients and the sparsity of the model depends on the value chosen for *α*. The elastic net aims to minimise:

$$(y - X\beta)^2 + \lambda(\frac{1}{2}(1 - \alpha) \sum_{j=1}^{p} \beta_j^2 + \alpha \sum_{j=1}^{p} |\beta_j|)$$

Where *y*, *X*, *β* and *λ* are defined in the same way as for ridge regression in section 4.1.1, and *α* determines the degree of mixture between the L1 and L2 penalties. Here, elastic net was used in two different ways. In the first instance, the *α* parameter was set to 0.5, and the *λ* parameter was determined by 10-fold cross-validation. In the second instance, the values of *α* and *λ* were both determined by 10-fold cross-validation.

4.1.4 BSLMM

The BSLMM (Zhou *et al.*, 2013) combines the standard linear mixed model (LMM) with a Bayesian variable selection regression (BVSR) model. The BSLMM assumes that all SNPs each have a small effect on the phenotype, and so is in a sense polygenic, but also assigns a small group of SNPs an additional large effect on the phenotype, and so can also be considered to have sparse properties. The model is defined as:

$$y = 1_n\mu + X\beta + u + \varepsilon$$

$$\beta \sim \pi N(0, \sigma_a^2\tau^{-1}) + (1-\pi)\partial_0$$

$$u \sim MVN(0, \sigma_b^2\tau^{-1}K)$$

$$\varepsilon \sim MVN(0, \tau^{-1}I_n)$$

where μ is the mean value of expression, β is a vector of fixed effects, u is a vector of random effects, ε is a vector of errors, π is the proportion of variants assigned a non-zero fixed effect, τ^{-1} is the residual variance, σ_a is the magnitude of non-zero fixed effects, δ_0 is a point mass at zero , σ_b is the magnitude of the random effects, K is a variance-covariance matrix of genotypes, and I_n is an identity matrix.

In practice, the model is re-parameterised in terms of PVE (ρ), which is the proportion of variance explained by the fixed and random effects together, and PGE (h), which is the proportion of variance explained by only the fixed effects. ρ and h are model hyperparameters estimated through Markov chain Monte Carlo (MCMC). Here, BSLMM was implemented using GEMMA, with 1,000 burn-in iterations and 10,000 iterations used for the MCMC settings. Additionally, I tested the BSLMM using 10,000

burn-in iterations and 100,000 MCMC iterations for 250 genes on chromosome 18 (Figure 4.1), although as the prediction accuracy estimates observed with this longer run were nearly identical to those observed with the shorter run, I chose to use the shorter run for the full analysis in the interest of time.

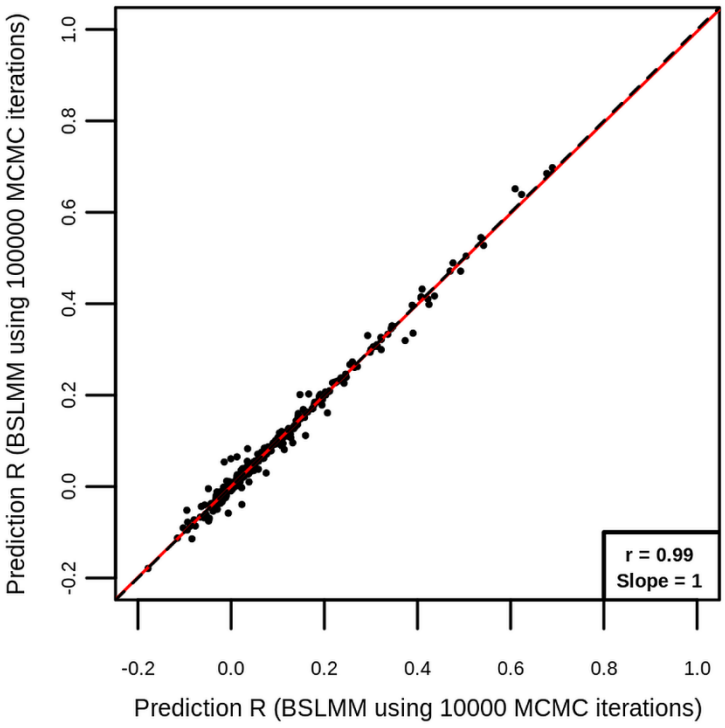

Figure 4.1. Comparison of BSLMM performance at two different MCMC lengths. R estimates from 10-fold cross-validation on EUR Geuvadis samples using BSLMM with an MCMC length of 10000 (x axis) and an MCMC length of 100000 (y axis) are shown for each gene on chromosome 18. The line of equality (dashed black) and a best fit line (solid red) are also shown.

4.1.5 BLUP

The Best Linear Unbiased Predictor (BLUP) is derived from a standard random effects regression model:

$$y_i = u_i + \varepsilon_i$$

where y_i is the phenotype of individual i, u_i is a random effect representing the genetic effect summed over all loci in individual i and ε_i is the residual. These are assumed to follow a normal multivariate distribution:

$$\begin{bmatrix} \boldsymbol{u} \\ \boldsymbol{\varepsilon} \\ \boldsymbol{y} \end{bmatrix} \sim MVN \begin{bmatrix} 0, & \boldsymbol{G}\sigma_u^2 & 0 & \boldsymbol{G}\sigma_u^2 \\ & 0 & \boldsymbol{I}\sigma_\varepsilon^2 & \boldsymbol{I}\sigma_\varepsilon^2 \\ & \boldsymbol{G}\sigma_u^2 & \boldsymbol{I}\sigma_\varepsilon^2 & \boldsymbol{G}\sigma_u^2 + \boldsymbol{I}\sigma_\varepsilon^2 \end{bmatrix}$$

$$\boldsymbol{G} = \boldsymbol{X}\boldsymbol{X}^T/p$$

where $\boldsymbol{G}$ is a genetic relationship matrix (GRM) defined using genotypes of SNPs within 1 Mb of the gene, p is the number of SNPs used to generate the GRM, and $\mathbf{I}$ is an identity matrix. This model produces a polygenic solution, where all SNPs have a small effect on the phenotype of interest. Further details on the model can be found in (de Los Campos *et al.*, 2013). Here, the BLUP model was implemented using GEMMA.

4.1.6 Random Forests

Random Forests is a tree-based machine learning method proposed by (Breiman, 2001). The first step in the standard Random Forests algorithm consists of bagging, which consists of randomly splitting the data to create many different training data sets from the initial data. Each of these data sets is then used to construct a decision tree. When constructing a tree, feature bagging is also performed, meaning for each tree a different random subset of the covariates (SNPs) is used to construct the tree. This procedure is performed for all the training data sets, resulting in a "forest" of decision trees. The mean of the predictions from each tree in the forest is then used as the overall prediction from the Random Forests approach. Here, Random Forests were implemented using the R package *ranger*.

4.2 Comparison of statistical methods for the prediction of gene expression from SNP genotype data through 10-fold nested cross-validation

Gene expression prediction models used in TWAS can be (and have been) constructed using a range of statistical methods. The ability of a method to predict

gene expression from SNPs will depend (at least in part) on how well the method's assumptions match the genetic architecture of gene expression. To date, there has been limited comparison of different methods on the same data to establish which gives the most accurate predictions. To address this, a comparison of seven different model training methods was performed. These methods were LASSO, elastic net (with $\alpha = 0.5$), elastic net (with α determined by cross-validation), ridge regression, BSLMM, BLUP and Random Forests. For each of the seven different methods, 10-fold nested cross-validation was performed for each gene in the Geuvadis data set. For each gene, prediction accuracy was calculated as the mean of the 10 estimates of correlation between predicted and observed expression from each of the outer folds of the nested cross-validation.

Prediction accuracy estimates were obtained by all seven methods for 22,218 genes. On average the BSLMM performed the best across these 22,218 genes, achieving a mean R = 0.0743 (Table 4.1). Behind the BSLMM, the Random Forests and the penalised regression approaches that assumed sparsity (LASSO and elastic net) outperformed the more polygenic approaches (BLUP and ridge regression).

Method	**Mean R (across 22,218 genes)**
Ridge regression	0.0587
Elastic net (α=0.5)	0.0634
Elastic net (α tuned by cross-validation)	0.0656
LASSO	0.0626
BSLMM	0.0743
BLUP	0.0608
Random Forests	0.0641

Table 4.1. Mean R estimates across 22,218 genes from 10-fold nested cross-validation using 7 different statistical methods.

Overall, estimates of prediction accuracy achieved by each of the seven methods were highly correlated with each other (Figure 4.2). As expected, the greatest pairwise correlations were observed between pairs of sparse methods (e.g. elastic net and LASSO) or pairs of polygenic methods. Despite these high correlations between the results from different methods, for some specific genes the results achieved by the methods were quite different. One such gene was *HSPA12B*, which

showed R = 0.877 with elastic net (α = 0.5) and R = 0.885 with LASSO, but only R = 0.425 with ridge regression.

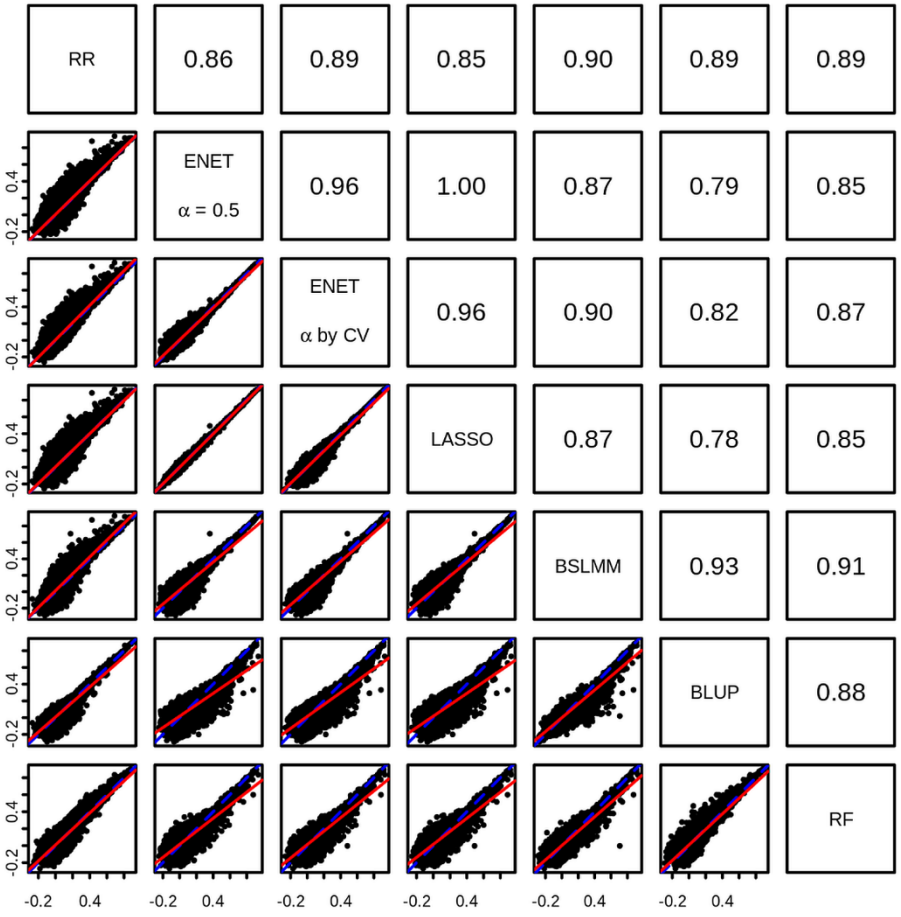

Figure 4.2. Correlation between R estimates from 7 different modelling approaches. In the lower panels, each point represents a single gene, and the R estimate obtained from the 2 corresponding methods are shown on the x and y axes. Also shown are the line of equality (blue dashed line) and a best fit line between x and y (red solid line). In the upper panels, Pearson correlation coefficients between the R estimates from pairs of methods are shown.

While the BSLMM appears to perform the best of all methods tested here, closer inspection of its MCMC chain revealed that it often showed a failure to converge. For each run of the BSLMM, the mixture of the Markov chain was evaluated by calculating the autocorrelation between Markov chain states. For each chain, good mixing was considered to be achieved when the autocorrelation at lag 100 was between -0.1 and 0.1, while chains showing autocorrelation values outside this range

were considered to have mixed poorly. Of the 225,170 BSLMM models generated as part of the 10-fold cross-validation, only 16,468 (7.31%) showed consistent convergence of all six hyperparameters, indicating a lack of reliability. An example of the MCMC chains of the six hyperparameters for the *PARP4P3* gene is given in Figure 4.3, and is representative of the MCMC chains observed for most genes.

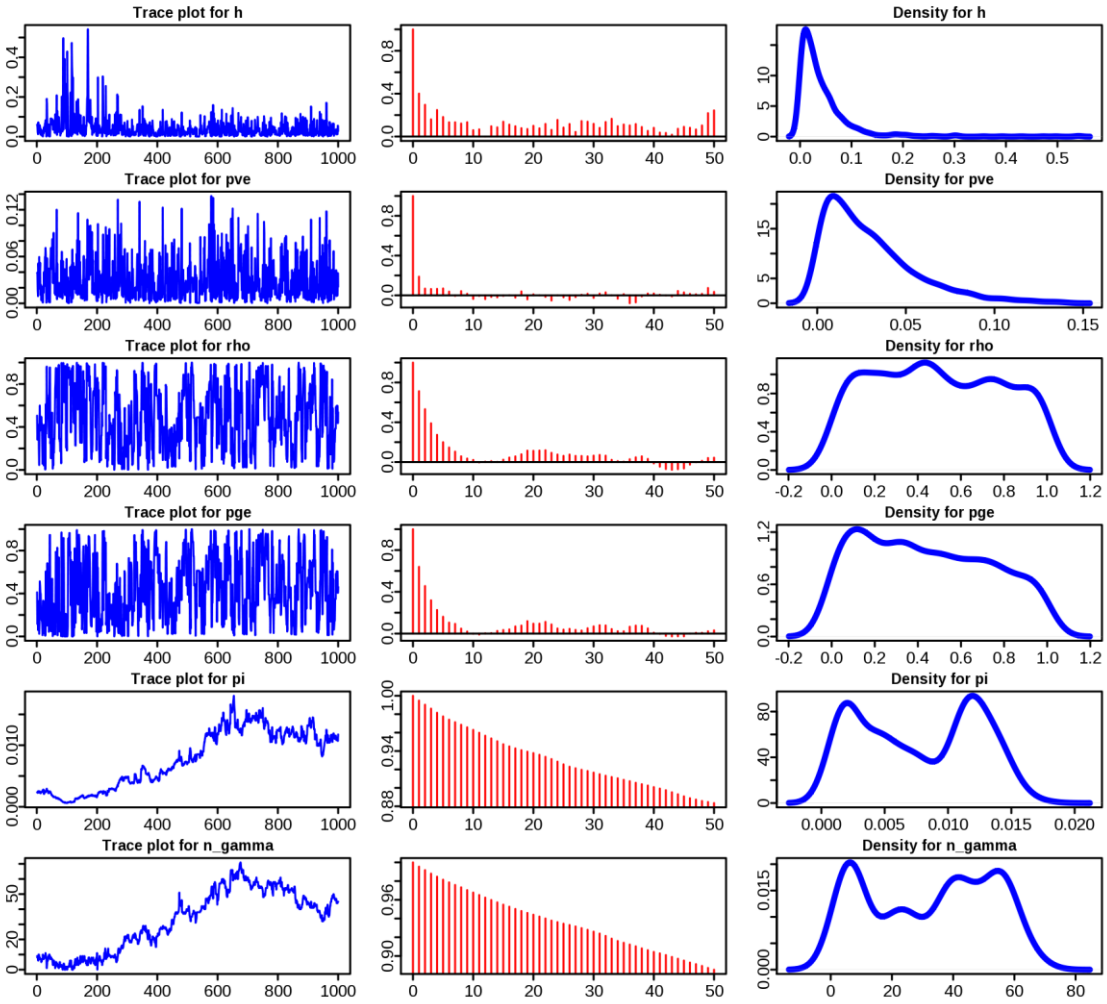

Figure 4.3. Convergence of hyperparameters for BSLMM. This plot shows convergence of BSLMM hyperparameters for the *PARP4P3* gene. Each row contains plots for one of the BSLMM hyperparameters (h, pve, rho, pge, pi, n_gamma). The leftmost graphs show trace plots for these hyperparameters, showing the values of the hyperparameter selected in each step of the MCMC. The central plots show autocorrelation. The rightmost plots show the density of hyperparameter values chosen across the MCMC. It is expected that a hyperparameter should show a trace that travels across the parameter space but hovers around a mean, and autocorrelation that quickly approaches zero.

When looking at results on a gene-by-gene basis, it was clear that the expression of most genes could not be accurately predicted by any of the methods (Figure 4.4), with distributions of R from all methods heavily skewed towards zero, and many genes showing a negative correlation between predicted and observed expression. Despite this, there existed a subset of genes for which expression could be reasonably well predicted from SNP genotypes. A total of 480 genes achieved R ≥ 0.5 with any of the 7 methods, and these genes were subsequently defined as “well-predicted”.

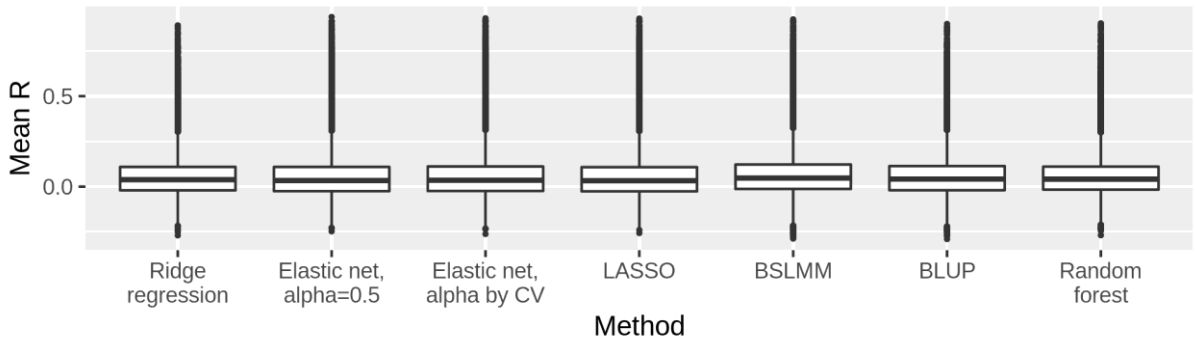

Figure 4.4. Boxplots of gene expression prediction accuracy estimates from 7 methods. Each boxplot shows the distribution of R estimates (between predicted and observed expression) for 22218 genes from 10-fold nested cross-validation for 1 statistical method. The central line within the box represents the median, with the upper and lower quartiles shown as the hinges.

When looking at only these well-predicted genes, the difference between the sparse methods and the polygenic methods was more stark, with the sparse method outperforming the polygenic methods even more strongly (Figure 4.5).

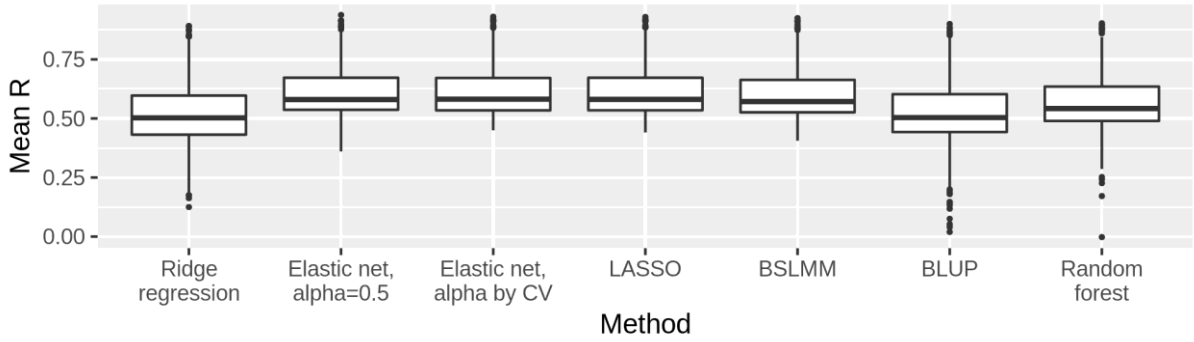

Figure 4.5. Boxplots of gene expression prediction accuracy estimates from 7 methods for well-predicted genes. Each boxplot shows the distribution of R estimates (between predicted and observed expression) for 480 genes from 10-fold nested cross-validation for 1 statistical method. These genes had R ≥ 0.5 from at least one of the 7 methods. The central line within the box represents the median, with the upper and lower quartiles shown as the hinges.

To investigate these genes further, a gene set enrichment analysis was conducted using the FUMA software. In total, 11 gene sets were significantly enriched (Table 4.2). Most of these gene sets related to immune functions, potentially due to the immune nature of the LCLs in which Geuvadis expression was measured.

Category	Gene set	Number of genes in set	Number of well predicted genes in set	p	Bonferroni adjusted p
GO_bp	GO_ANTIGEN_PROCESSING_AND_PRESENTATION_OF_ENDOGENOUS_PEPTIDE_ANTIGEN	14	5	3.87E-07	0.002847112
GO_bp	GO_INTERFERON_GAMMA_MEDIATED_SIGNALING_PATHWAY	88	9	9.30E-07	0.003418405
GO_bp	GO_ANTIGEN_PROCESSING_AND_PRESENTATION_OF_PEPTIDE_ANTIGEN	186	12	2.18E-06	0.004257968
GO_bp	GO_ANTIGEN_PROCESSING_AND_PRESENTATION	221	13	2.32E-06	0.004257968
GO_bp	GO_ANTIGEN_PROCESSING_AND_PRESENTATION_OF_ENDOGENOUS_ANTIGEN	21	5	3.68E-06	0.005410801
GO_bp	GO_RRNA_METHYLATION	27	5	1.38E-05	0.016886592
GO_bp	GO_RESPONSE_TO_INTERFERON_GAMMA	194	11	1.95E-05	0.020482831
GWAScatalog	Myositis	15	7	1.66E-10	1.77E-07
GWAScatalog	Pneumonia	9	6	1.98E-10	1.77E-07
GWAScatalog	Lymphoma	16	7	2.92E-10	1.77E-07
GWAScatalog	Response to hepatitis B vaccine	22	7	4.10E-09	1.86E-06

Table 4.2. Gene set enrichment analysis on 480 well-predicted genes.

4.3 Comparison of prediction accuracy estimates with heritability estimates

In TWAS, gene expression is predicted using only genetic variants. The upper bound of the accuracy with which gene expression can be predicted using only genetic information is equal to the heritability of gene expression. To examine how close the prediction accuracy estimates were to this upper bound, the heritability of Geuvadis gene expression was estimated using GCTA. The heritability of gene expression was estimated using only SNPs within 1Mb of each gene (the same SNPs used to predict gene expression in the 10-fold cross-validation). The prediction accuracy estimates obtained from 10-fold nested CV were then compared with these heritability estimates.

Reassuringly, prediction accuracy estimates from elastic net (α =0.5) were highly correlated with point estimates of local heritability obtained using GCTA (Figure 4.6), although there was a set of genes with large point estimates of heritability but poor estimates of prediction accuracy. On average, the prediction accuracy estimates tended to be slightly smaller than the heritability estimates, suggesting there may be some room for improvement in prediction accuracy.

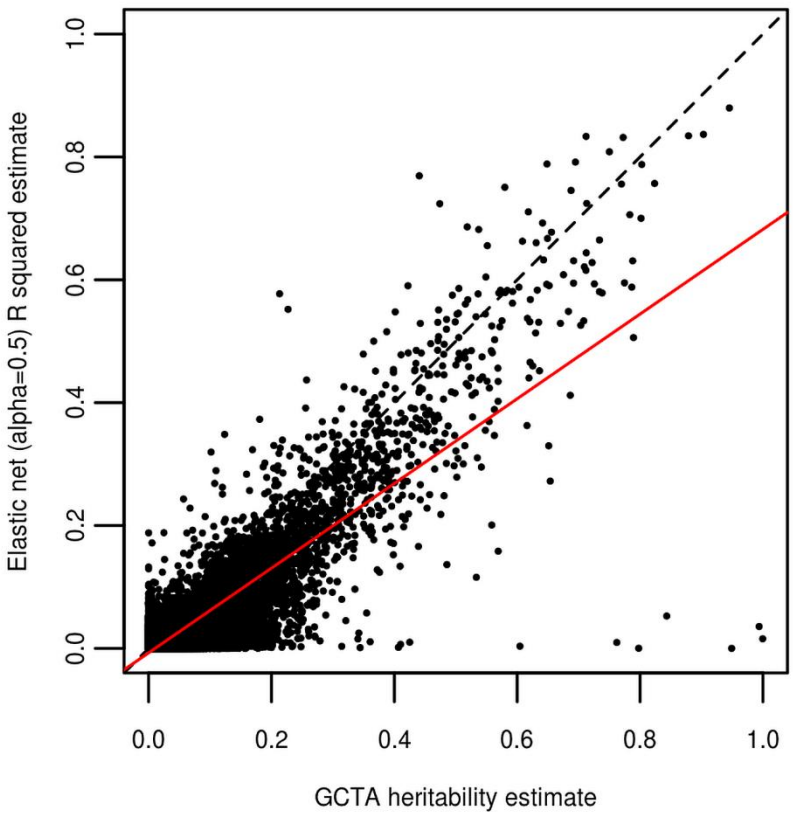

Figure 4.6. Comparison of prediction accuracy estimates with heritability. Each point shows the estimate of the heritability of gene expression attributable to SNPs within 1Mb of the gene obtained using GCTA (x axis) and the prediction accuracy estimate from the 10-fold nested cross-validation using elastic net (α=0.5) (y axis). In this plot, prediction accuracy is shown as R^2 rather than R, as the heritability is the upper bound on the estimate of R^2. Also shown are the line of equality (black dashed) and a line of best fit (red).

4.4. Comparison of statistical methods through application to WTCCC1 data

To examine how gene expression prediction models trained using the different statistical approaches compared at detection of predicted expression – trait associations, gene expression prediction models were trained using all European ancestry Geuvadis samples using each of the 7 methods, and were applied to T1D GWAS data from WTCCC1.

The seven methods detected associations between predicted gene expression and T1D in the same genomic regions (Figure 4.7), including the MHC on chromosome 6, and at 12q13 and 12q24 on chromosome 12, all of which are known T1D risk loci

that have been previously identified in T1D GWAS. Z scores achieved by all seven approaches were highly correlated with each other (Figure 4.8), and no approach achieved a greater average z score than another, indicating further similarity between the methods.

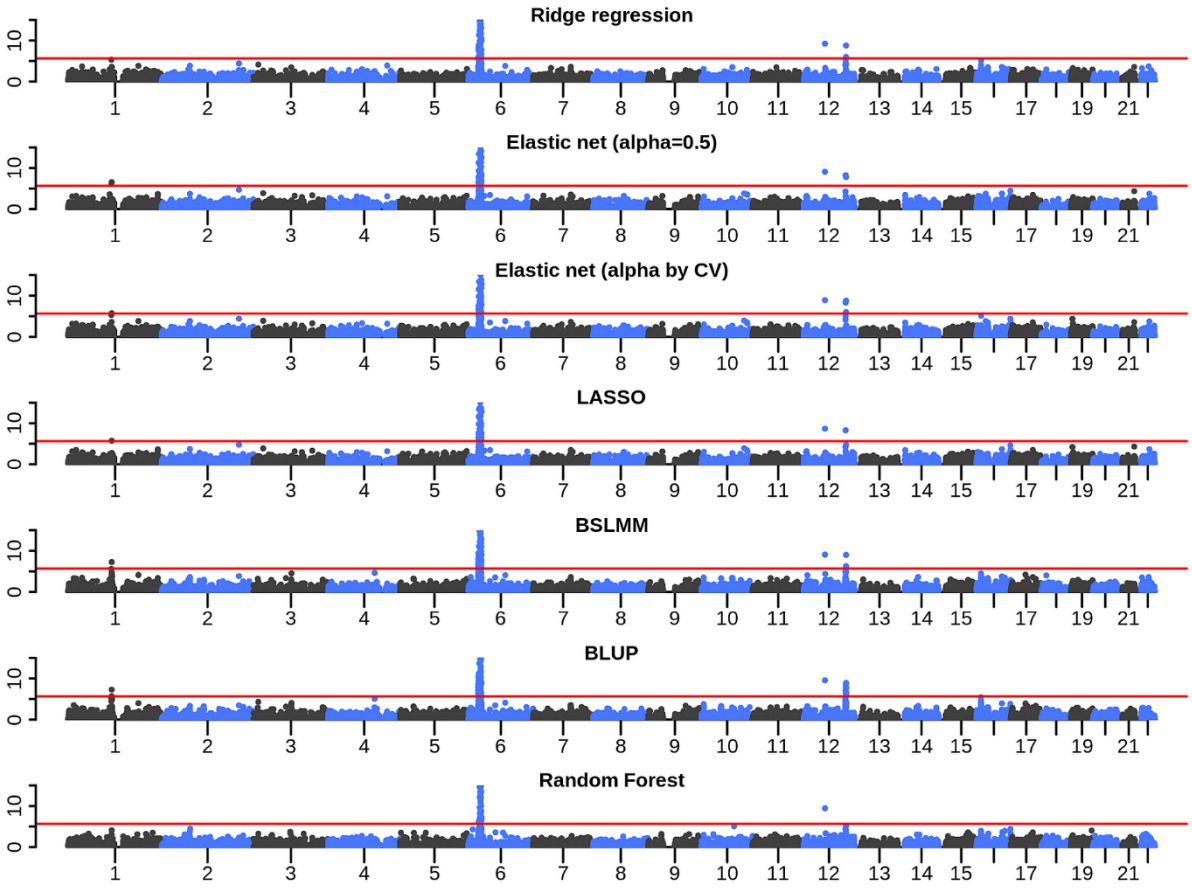

Figure 4.7. Manhattan plots from application of gene expression prediction models to WTCCC T1D GWAS data. Each plot shows the results of a TWAS on WTCCC T1D data using gene expression prediction models trained with a different statistical method. In each plot, each point represents a gene, plotted by its genomic position (defined by the TSS) on the x axis, and its p value in the TWAS on the y axis. The red lines indicate Bonferroni-corrected significance thresholds.

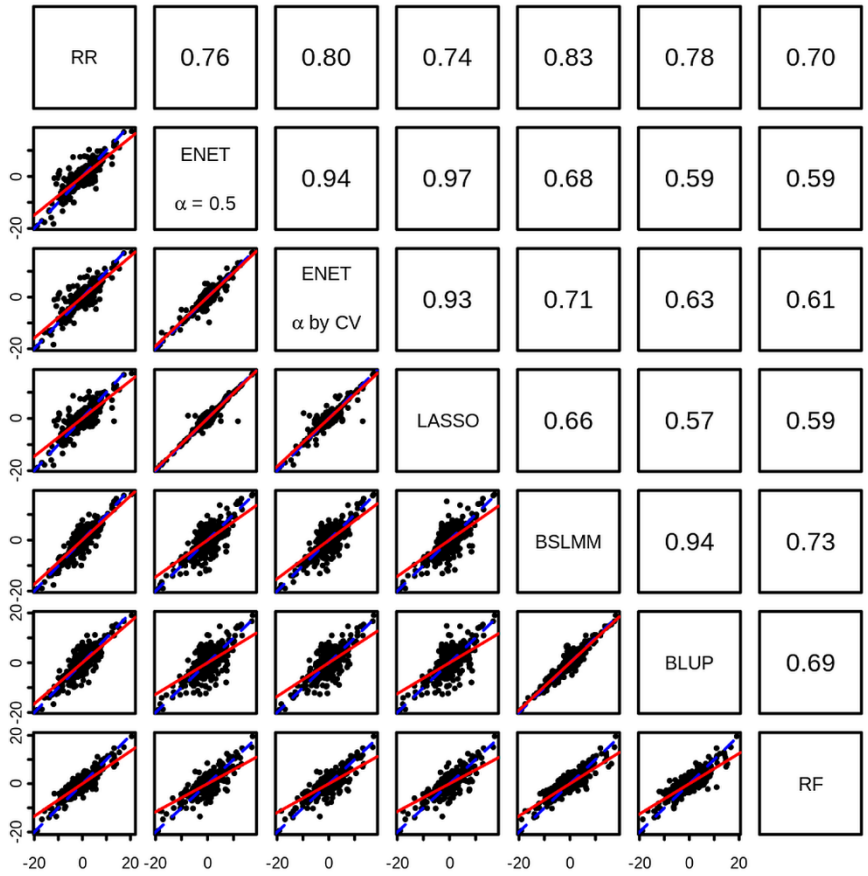

Figure 4.8. Correlation between z scores from TWAS on WTCCC T1D data using 7 different modelling approaches. In lower panels, each point shows a single gene, with the z scores from TWAS using models trained with 2 different statistical approaches shown on the x and y axes. Also shown are the line of equality (blue dashed line) and a best fit line between x and y (red solid line). Upper panels show Pearson correlation estimates between z scores from TWAS using the 2 corresponding methods.

4.5. Investigation of the effect of sample size on prediction accuracy

To investigate how the sample size of the reference panel used to train gene expression prediction models affects prediction accuracy in TWAS, the 10-fold nested cross-validation was repeated using 10% of samples as the training set and the remaining 90% as the test set in each fold. Elastic net with α set to 0.5 was used for this analysis. Estimates of prediction accuracy from this analysis were compared with those from the cross-validation using the larger training set.

Prediction accuracy estimates tended to be smaller when using fewer samples to train gene expression prediction models (Figure 4.9). Of the 22,490 genes for which gene expression prediction accuracy could be estimated in this analysis and the previous analysis, 14,019 (62.3%) showed reduced prediction accuracy in the analysis using the smaller training set, with genes showing higher prediction accuracy in the analysis using the smaller training set tending to have low prediction accuracy in both analyses. However, the prediction accuracy estimates achieved at the reduced sample size were still highly correlated (r = 0.79) with those achieved using the larger sample size. Furthermore, many genes for which large R estimates were reached in the 90% training set analysis also reached similarly large R estimates at the reduced sample size. One such gene was *RPS26*, which achieved prediction accuracy R = 0.913 in the 90% training set analysis and prediction accuracy R = 0.888 in the 10% training set analysis. Another gene, *AC008957.1* showed prediction accuracy R = 0.915 when using the larger training set, and R = 0.889 when using the smaller training set. For genes such as these where local SNP genotypes have strong enough effects on gene expression, it seems that small sample sizes are sufficient for constructing models that can predict expression well.

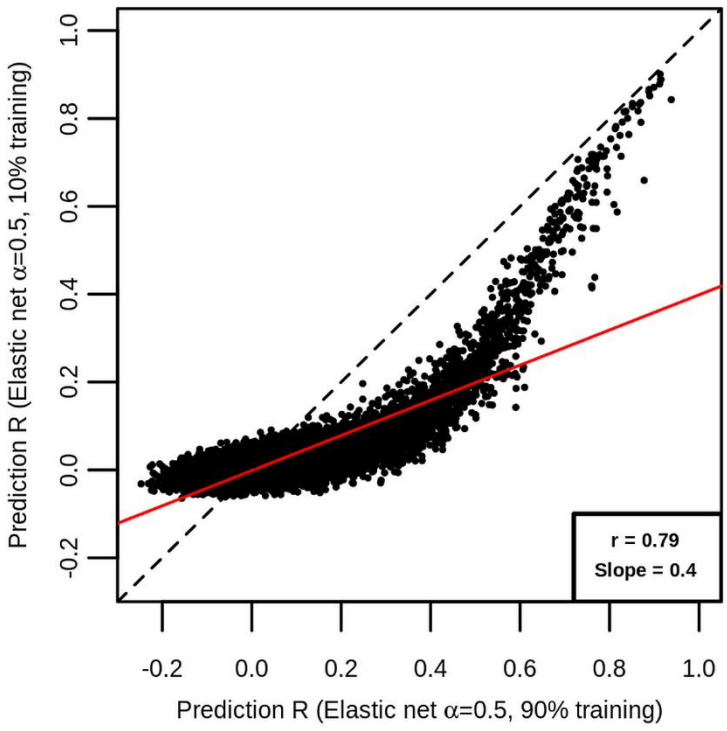

Figure 4.9. Comparison between prediction accuracy estimates at large and small samples sizes. Each point represents a gene, with its R estimate from 10-fold nested cross-validation using 90% of EUR samples as the model training set on the x axis, and the R estimate from 10-fold nested cross-validation using 10% of EUR samples as the model training set on the y axis. All R estimates were obtained using elastic net with α = 0.5. Also shown are the line of equality (black dashed) and a line of best fit (red solid), with the correlation between x and y and the slope of the best fit line shown in the bottom right corner.

Following this, 10-fold nested cross-validation was repeated a further 7 times, each time using a different proportion of the samples as the gene expression prediction model training set. The proportions used for prediction model training in each analysis were 20%, 30%, 40%, 50%, 60%, 70% and 80%. Overall, a clear improvement in average prediction accuracy was observed with increasing sample size used for prediction model training (Figure 4.10A). Additionally, there was no plateau to the average prediction accuracy increase at the sample sizes tested here, indicating that larger average prediction accuracy estimates could be achieved by increasing the sample size further. However when examining this on a gene-by-gene basis, a plateau was observed for some individual genes, with no increase in prediction accuracy beyond a certain sample size (Figure 4.10B). For these genes, it is possible that the limit on prediction accuracy imposed by the heritability of gene expression attributable to the local SNP genotypes had been reached.

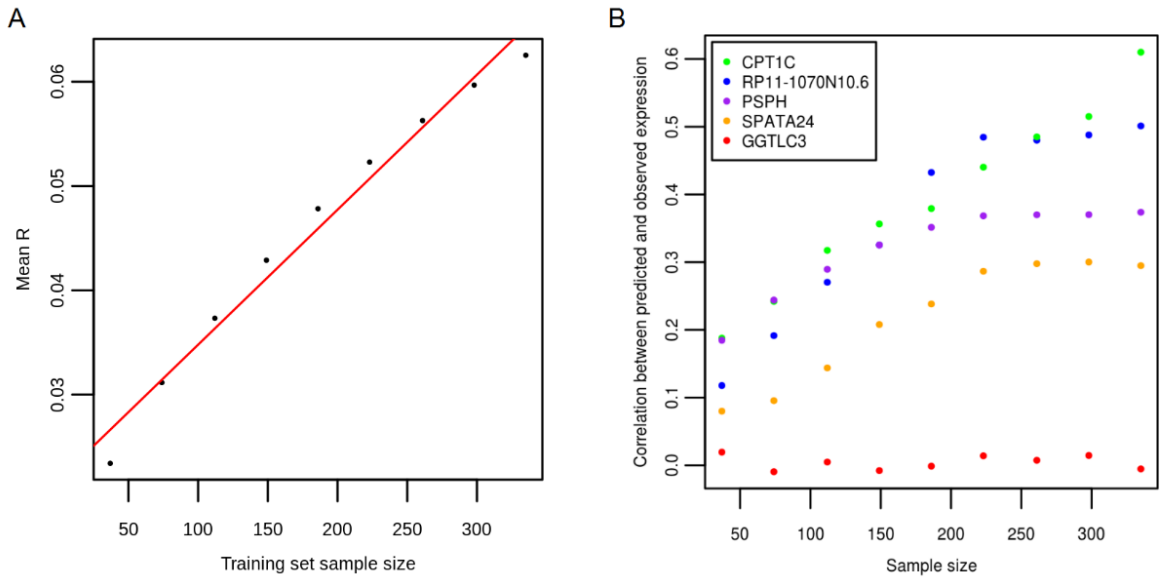

Figure 4.10. Prediction accuracy estimates at a range of samples sizes. Prediction models were trained using 10%, 20%, 30% … of EUR samples, and tested on the remaining samples. In plot A, each point shows the mean R across genes (y axis) and the sample size at which models were trained (x axis). The red line indicates a best fit line between x and y. There is a clear increase in average prediction accuracy with increasing sample size. In plot B, each point shows the prediction accuracy estimate (y axis) achieved for the specified gene, and the sample size at which this prediction accuracy estimate was obtained (x axis). For some of the genes, the prediction accuracy estimates do not continue to increase with increasing sample size.

4.6. Investigation of the effect of ancestry on prediction accuracy

To investigate how accurately prediction models trained using data from samples of one ancestry were able to predict the gene expression of samples of a different ancestry, prediction models were trained using all 373 EUR Geuvadis samples using elastic net (with $\alpha = 0.5$) and were applied to genotype data for 89 YRI Geuvadis samples.

On average, prediction models tended to perform more poorly when predicting expression of samples of a different ancestry than those used to train the prediction model (Figure 4.11A). Across the genes for which prediction accuracy estimates were available from both analyses, the average prediction accuracy obtained from 10-fold cross-validation using only EUR samples (mean R = 0.0625) was greater than the average prediction accuracy from application of prediction models trained with EUR samples to YRI samples (mean R = 0.0332). However, when looking on a gene-by-gene basis there was no consistent pattern to the results, with some genes showing greater R from application of EUR models to YRI samples than from the 10-fold nested cross-validation.

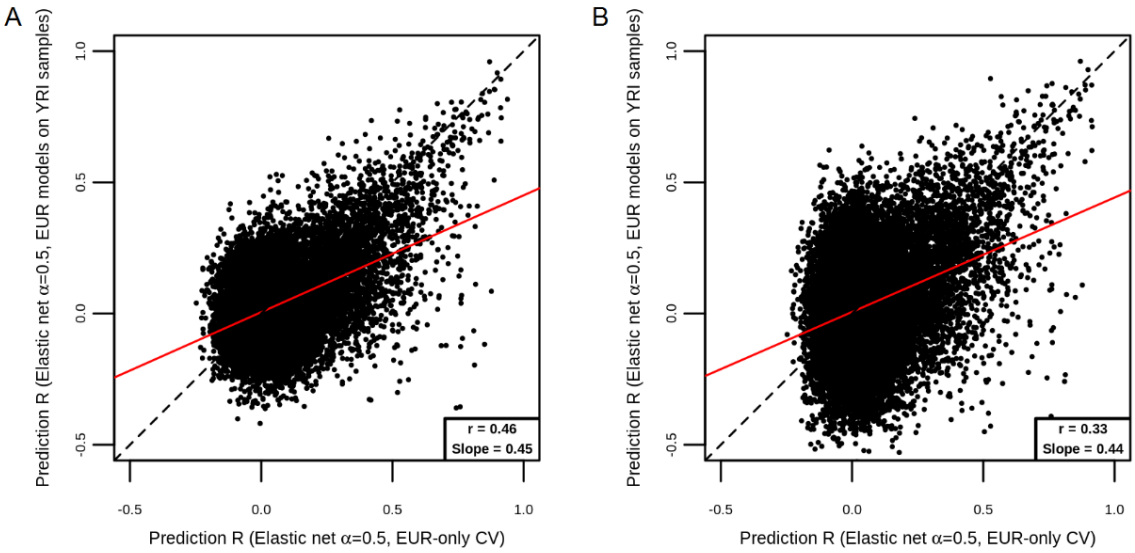

Figure 4.11. Comparison of prediction accuracy estimates when EUR-trained models are applied to EUR and YRI populations. On both plots, each point represents a gene, and shown are the R estimate from 10-fold nested cross-validation within EUR samples (x axis), and the R between expression predicted using models trained on EUR samples and applied to YRI samples, and measured YRI expression (y axis). Plot (A) corresponds to the analysis where sample sizes of the model training and testing sets used for the within-EUR analysis was not the same as the sample sizes used in the across-ancestries analysis. Plot (B) corresponds to the analysis where the sample sizes of the model training and testing sets was the same in both the within-EUR and across-ancestries analyses. Also shown are the line of equality (black dashed) and a line of best fit (red solid), with the correlation between x and y and the slope of the best fit line shown in the bottom right corner.

As shown previously, the sample size of the prediction model training data set is known to have an effect on prediction accuracy. However, this was not accounted for in the above analysis in Figure 4.11A. To account for sample size differences between the 10-fold cross-validation on EUR samples and the application of EUR prediction models to YRI samples, EUR prediction models were re-trained using 90% of the EUR samples and were applied to 37 YRI samples, matching the sample sizes used for the prediction model training and testing sets used for one fold of the 10-fold nested cross-validation on EUR samples. Again, average prediction accuracy from application of the EUR trained models to YRI samples was smaller than average prediction accuracy from the 10-fold nested cross-validation on EUR samples (Figure 4.11B), although there was no consistent pattern to the results on a gene-by-gene basis. It is worth noting that there was higher variation in the prediction accuracies observed from this reduced sample analysis (compared with the prediction

accuracies achieved when using all available samples). This is likely reflective of the smaller sample size used for the model testing set.

Following this, the reverse analysis was performed by training gene expression prediction models using the 89 YRI samples and applying them to the EUR samples. A 10-fold nested cross-validation was also performed using the 89 YRI samples, and the prediction accuracy estimates from this were compared with those from application of YRI-trained models to EUR samples. Overall, the average prediction accuracy from 10-fold cross-validation on YRI samples (mean R = 0.0317) was slightly greater than the average accuracy from application of YRI-trained models to EUR samples (mean R = 0.0184) (Figure 4.12A). As above, to account for the difference in sample sizes between analyses, the analysis was repeated by training prediction models on 90% of YRI Geuvadis and applying them to 9 EUR samples, matching the sample sizes to one fold of the 10-fold cross-validation on the 89 YRI samples. Again, the average prediction accuracy was poorer when predicting across populations (Figure 4.12B). It is worth noting that the prediction accuracy estimates obtained from this repeated analysis showed much variation, likely reflecting the smaller sample sizes used for both prediction and testing.

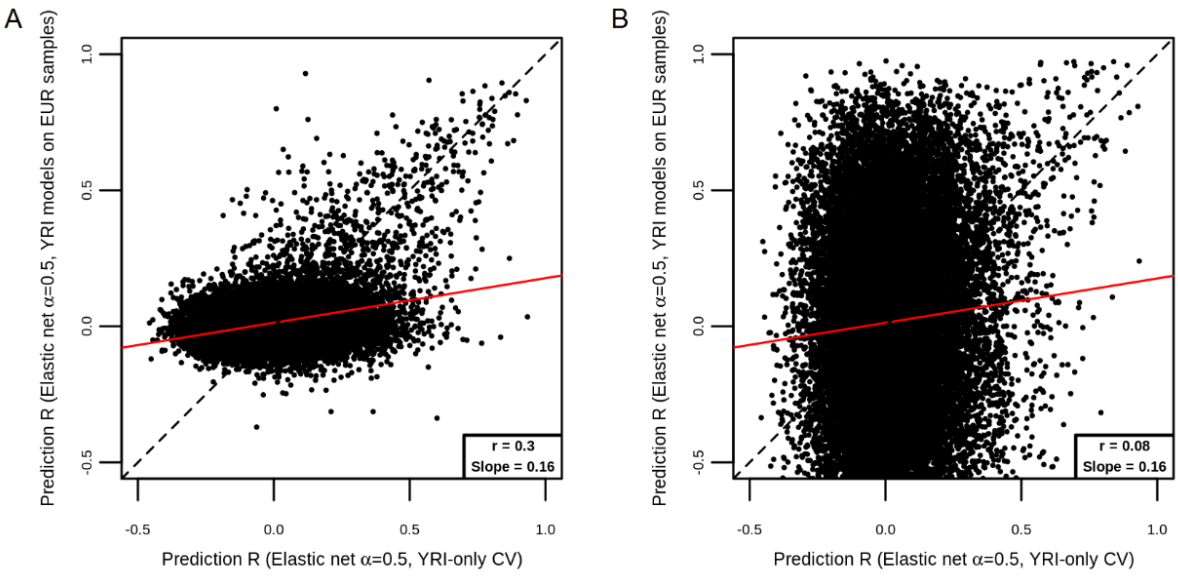

Figure 4.12. Comparison of prediction accuracy estimates when YRI-trained models are applied to EUR and YRI populations. On both plots, each point represents a gene, and shown are the R estimate from 10-fold nested cross-validation within YRI samples (x axis), and the R between expression predicted using models trained on YRI samples and applied to EUR samples, and

measured EUR expression (y axis). Plot (A) corresponds to the analysis where sample sizes of the model training and testing sets used for the within-YRI analysis was not the same as the sample sizes used in the across-ancestries analysis. Plot (B) corresponds to the analysis where the sample sizes of the model training and testing sets was the same in both the within-YRI and across-ancestries analyses. Also shown are the line of equality (black dashed) and a line of best fit (red solid), with the correlation between x and y and the slope of the best fit line shown in the bottom right corner.

Finally, the EUR and YRI samples were combined into a single group of 462 samples. This group was then down-sampled to a group of 373 samples, with the relative proportions of EUR and YRI samples kept the same as in the larger group of 462 samples. Using this mixture of EUR and YRI samples, a 10-fold nested cross-validation was performed using elastic net (with $\alpha = 0.5$). The prediction accuracy estimates achieved with this mixed sample were highly correlated ($r = 0.780$) with those achieved from 10-fold nested cross-validation using only the 373 EUR samples (Figure 4.13). The average prediction accuracy estimate obtained from this mixed sample 10-fold cross-validation (mean $R = 0.0609$) was similar to, but marginally smaller than the average prediction accuracy estimate obtained from 10-fold cross-validation on EUR samples (mean $R = 0.0625$). This demonstrates that even when the population contains samples from different ancestries, the prediction accuracy can be similar to that achieved using a single ancestry, as long as the composition of the training population matches that of the testing population.

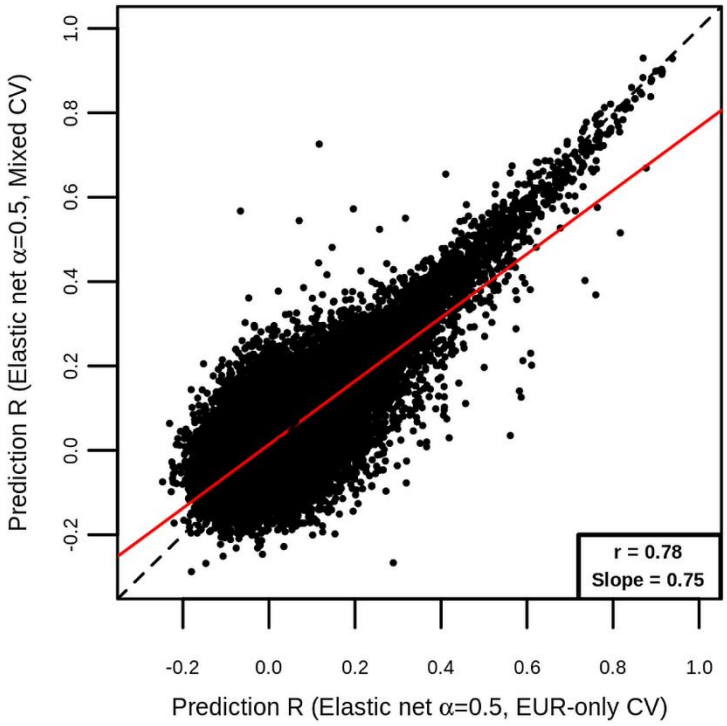

Figure 4.13. Comparison of prediction accuracy estimates when using an EUR-ancestry population and a population of mixed ancestry. Each point represents a gene, and shown are the R estimate from 10-fold nested cross-validation within EUR samples (x axis), and the R from 10-fold nested cross-validation using the EUR and YRI samples combined into a single group (y axis). Also shown are the line of equality (black dashed) and a line of best fit (red solid), with the correlation between x and y and the slope of the best fit line shown in the bottom right corner.

4.7. Prediction using models trained with data from GTEx

All analyses conducted up to this point have been performed by splitting the Geuvadis data into training and testing sets, and then using prediction models trained with samples from one set to predict gene expression in another set. While this is a convenient way of conducting the analysis, it is likely to be over-optimistic as the different subsets of Geuvadis data were collected and processed in the same way, and so are highly similar. It does not necessarily reflect a more realistic application of TWAS methods, in which the data sets used for model training and for application may be quite different to one another. To investigate a more realistic scenario, gene expression prediction models trained using GTEx EBV-transformed LCL gene expression data were downloaded from predictdb.org and were applied to the 373 Geuvadis EUR samples. The prediction accuracy achieved by these models was

then calculated as the Pearson correlation coefficient between their predicted expression and measured Geuvadis expression.

Prediction accuracy estimates achieved by the prediction models trained using GTEx LCL data were highly concordant with those from the 10-fold nested cross-validation using the 373 EUR Geuvadis samples with elastic net (α=0.5) (Figure 4.14). This was especially the case for the well-predicted genes, including *RPS26*, which achieved prediction accuracy R = 0.905 from application of the GTEx LCL gene expression prediction model to Geuvadis, and prediction accuracy R = 0.913 from 10-fold nested cross-validation using EUR Geuvadis samples. On average, the prediction accuracy estimates achieved by the GTEx-trained models (mean R = 0.188) were slightly smaller than those achieved by the 10-fold nested cross-validation (mean R = 0.199), indicating that Geuvadis-informed models were able to predict Geuvadis expression marginally better than GTEx informed models.

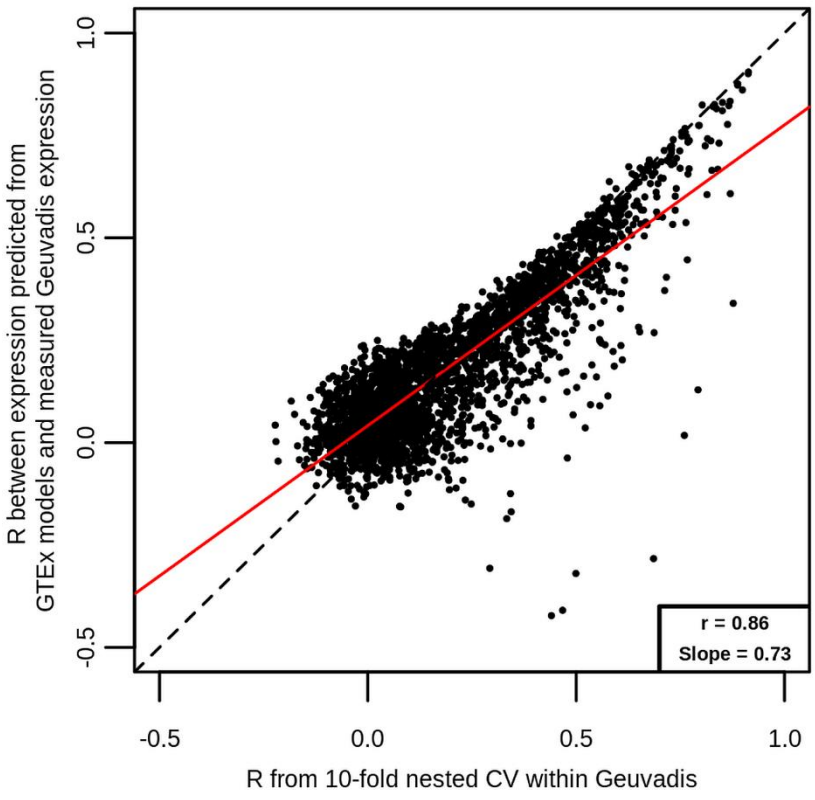

Figure 4.14. Comparison of Geuvadis-trained models and GTEx-trained models at predicting Geuvadis expression. Each point represents a gene, and shown are the R estimates from 10-fold

nested cross-validation within EUR samples on Geuvadis data (x axis), and R between measured Geuvadis expression and expression predicted using GTEx-trained models. Also shown are the line of equality (black dashed) and a line of best fit (red solid), with the correlation between x and y and the slope of the best fit line shown in the bottom right corner.

4.8. Investigation of the effect of tissue on prediction accuracy

Gene expression prediction models trained using GTEx data from a range of tissues were next applied to the Geuvadis data to examine the portability of prediction models across tissues. In total, 47 sets of gene expression prediction models downloaded from predictdb.org, each trained using GTEx data from a tissue other than LCLs (in which Geuvadis expression was measured). These 47 sets of gene expression prediction models were applied to the Geuvadis data. The prediction accuracy achieved by these sets of prediction models was then estimated as the Pearson correlation coefficient between the expression they predicted and the measured Geuvadis gene expression. The average level of prediction accuracy achieved by prediction models for each of the non-LCL GTEx tissues was lower than that achieved (average R=0.188) with the GTEx LCL prediction models (Table 4.3), indicating that correct tissue matching leads to more accurate prediction of gene expression.

GTEx tissue	**Number of genes with predicted tissue expression and measured Geuvadis expression**	**Average correlation between predicted tissue expression and measured Geuvadis expression**
Adipose_Subcutaneous	6623	0.0886
Adipose_Visceral_Omentum	5228	0.0985
Adrenal_Gland	3748	0.0971
Artery_Aorta	5415	0.0898
Artery_Coronary	2802	0.1045
Artery_Tibial	6758	0.0809
Brain_Amygdala	1846	0.0882

Brain_Anterior_cingulate_cortex_BA24	2586	0.0886
Brain_Caudate_basal_ganglia	3233	0.0877
Brain_Cerebellar_Hemisphere	3773	0.0716
Brain_Cerebellum	4852	0.0666
Brain_Cortex	3383	0.0832
Brain_Frontal_Cortex_BA9	2765	0.0868
Brain_Hippocampus	2189	0.0926
Brain_Hypothalamus	2195	0.0915
Brain_Nucleus_accumbens_basal_ganglia	2778	0.0885
Brain_Putamen_basal_ganglia	2505	0.0882
Brain_Spinal_cord_cervical_c-1	1974	0.0877
Brain_Substantia_nigra	1581	0.0903
Breast_Mammary_Tissue	4241	0.1037
Cells_EBV-transformed_lymphocytes	2737	0.1878
Cells_Transformed_fibroblasts	6226	0.0983
Colon_Sigmoid	4211	0.0991
Colon_Transverse	4406	0.1098
Esophagus_Gastroesophageal_Junction	4275	0.1000
Esophagus_Mucosa	6672	0.0907
Esophagus_Muscularis	6290	0.0890
Heart_Atrial_Appendage	4811	0.0932
Heart_Left_Ventricle	4393	0.0916
Liver	2708	0.0926
Lung	6186	0.0959
Minor_Salivary_Gland	1770	0.1075
Muscle_Skeletal	6263	0.0782
Nerve_Tibial	7440	0.0746
Ovary	2379	0.0961
Pancreas	4328	0.0942
Pituitary	3655	0.0893
Prostate	2453	0.1070
Skin_Not_Sun_Exposed_Suprapubic	6034	0.0867
Skin_Sun_Exposed_Lower_leg	7142	0.0814
Small_Intestine_Terminal_Ileum	2443	0.1202
Spleen	3712	0.1136
Stomach	3853	0.1089
Testis	5844	0.0634
Thyroid	7481	0.0772
Uterus	1923	0.1007
Vagina	1857	0.1090
Whole_Blood	5376	0.0959

Table 4.3. Mean R estimates from application of 48 sets of GTEx-trained prediction models to Geuvadis data.

The prediction accuracy achieved by models trained using data from each of the 47 GTEx non-LCL tissues was then directly compared with the accuracy achieved by models trained in the GTEx LCL data. To do this, only genes for which a prediction model was present for both GTEx LCL and in the non-LCL GTEx tissue of interest were used. When looking at the results of this comparison on a gene-by-gene basis, for many genes the prediction models trained using data from a GTEx non-LCL tissue could predict Geuvadis expression with similar accuracy to the prediction model trained using GTEx LCL data (Figure 4.15). 53.3% of the prediction accuracy estimates achieved by non-LCL prediction models trained using GTEX data from a non-LCL tissue were within 0.05 of the corresponding prediction accuracy estimate achieved for the same gene by the prediction model trained using GTEx LCL data. This indicated that in many instances, a gene expression prediction model trained using data from one tissue could proxy quite well for another tissue. However, there existed a set of genes for which the prediction model trained using GTEx LCL data strongly outperformed the prediction model trained using GTEx data from a non-LCL tissue. In total, 10.3% of prediction accuracy estimates achieved by non-LCL prediction models were at least 0.2 less than that achieved by the LCL model for the same gene. One example of this was *NDUFAF1*, for which the GTEx LCL model achieved a prediction accuracy estimate of 0.686, whereas the GTEx transverse colon model achieved an estimate of 0.128. In instances such as this, the LCL models clearly outperformed the non-LCL models, showing that correct tissue matching is important for some genes.

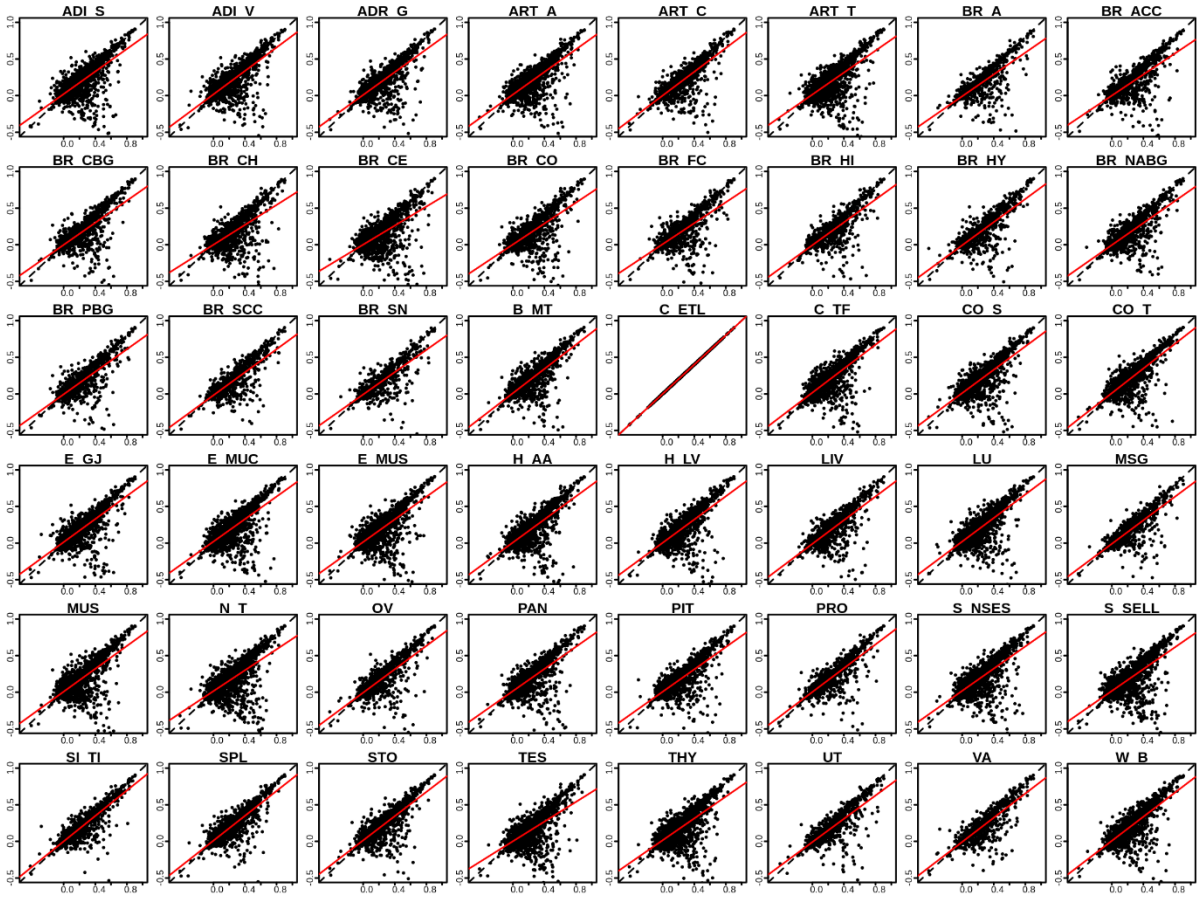

Figure 4.15. Comparison of prediction accuracy achieved by GTEx LCL-trained models and GTEx non-LCL-trained models. In each plot, the x axis shows the R between measured Geuvadis expression and expression predicted using models trained with GTEx LCL expression data. The y axis in each plot shows the R between measured Geuvadis expression and expression predicted using models trained with GTEx data from a tissue other than LCLs (the tissue is given in the plot sub-heading). Each point represents a gene.

Tissues are: adipose subcutaneous (ADI_S), adipose visceral omentum (ADI_V), adrenal gland (ADR_G), artery aorta (ART_A), artery coronary (ART_C), artery tibial (ART_T), brain – amygdala (BR_A), brain – anterior cingulate cortex (BR_ACC), brain – caudate basal ganglia (BR_CBG), brain – cerebellar hemisphere (BR_CH), brain – cerebellum (BR_CE), brain – cortex (BR_CO), brain – frontal cortex (BR_FC), brain – hippocampus (BR_HI), brain – hypothalamus (BR_HY), brain – nucleus accumbens basal ganglia (BR_NABG), brain – putamen basal ganglia (BR_PBG), brain – spinal cord cervical c-1 (BR_SCC), brain – substantia nigra (BR_SN), breast – mammary tissue (B_MT), cells – LCLs (C_ETL), cells – transformed fibroblasts (C_TF), colon – sigmoid (CO_S), colon – transverse (CO_T), esophagus – gastroesophageal junction (E_GJ), esophagus – mucosa (E_MUC), esophagus – muscularis (E_MUS), heart – atrial appendage (H_AA), heart – left ventricle (H_LV), liver (LIV), lung (LU), minor salivary gland (MSG), muscle – skeletal (MUS), nerve – tibial (N_T), ovary (OV), pancreas (PAN), pituitary (PIT), prostate (PRO), skin – not sun exposed suprapubic (S_NSES), skin – sun exposed lower leg (S_SELL), small intestine – terminal ileum (SI_TI), spleen (SPL), stomach (STO), testis (TES), thyroid (THY), uterus (UT), vagina (VA), whole blood (W_B).

4.9. Discussion

In this chapter, a range of different statistical approaches for the prediction of gene expression from SNP genotypes were compared, with the methods that assumed sparsity performing slightly better on average than those that assumed polygenicity, reinforcing similar findings from an earlier, more limited comparison of four statistical approaches (Zeng *et al.*, 2017). These differences between the sparse and polygenic methods were the greatest for the genes where expression could be predicted quite well (prediction accuracy ≥ 0.5) from SNP genotypes. Given that the genetic architecture of gene expression at SNPs proximal to genes is thought to be sparse (Wheeler *et al.*, 2016), and given that the gene expression prediction models here were trained by regressing gene expression on genotypes at SNPs most proximal to each gene, this result is perhaps unsurprising. In addition to the local architecture of gene expression, it is known that many distal SNPs act on the expression of genes, usually with much weaker effects than those observed for the proximal SNPs (Liu *et al.*, 2019). This is indicative that there is a more polygenic distal genetic architecture of gene expression, and so if these distal SNPs were to be used for gene expression prediction model training, it is possible that the more polygenic methods would outperform the more sparse methods. Further investigation would be required to determine if this is indeed the case.

On average, the best performing method here was the BSLMM, which showed a marginally higher average prediction accuracy than other methods including the Random Forests, both flavours of elastic net and LASSO. Currently, the most popular TWAS software packages (PrediXcan, MetaXcan and FUSION) use the elastic net (with the α parameter set to 0.5) to train their gene expression prediction models. Based on the results observed here, it seems that there could potentially be a slight gain in average prediction accuracy by fully tuning the α parameter of the elastic net, or by switching to Random Forests or the BSLMM.

However, there would be a number of issues with switching the methods. First, switching to the BSLMM would seem inappropriate given that major convergence issues were observed when using the BSLMM. These convergence issues imply that the parameters and hyperparameters used by BSLMM may not be correctly

estimated from the data, which raises concerns about the use of this method for the prediction of gene expression. Second, switching to the Random Forests approach would only work when performing a TWAS using individual level data, and would not be appropriate for use in a summary statistics based approach (such as MetaXcan). This is because the Random Forests approach does not produce estimates of the coefficients of SNP genotypes on gene expression, but instead produces a “forest” of prediction trees that require individual level genotype data as the input. Given that most TWAS conducted now use a summary statistics based method such as MetaXcan or FUSION, use of the Random Forests method would be inconvenient. Third, switching to tuning the α parameter in the elastic net would increase the time and computational resources required to train the gene expression prediction model. Given that the increased prediction accuracy achieved by tuning the parameter (compared to that achieved by setting to 0.5) was only marginal, the gain in TWAS power would likely only be marginal, and so it may not be worth doing this. The choice of whether or not to tune this parameter would likely come down to the size of the dataset used to train models and the computational resources available for the project.

A crucial observation from the results obtained here is that the prediction accuracy for many of the genes examined here was very low, with many genes showing a cross-validation R near 0, and some genes showing an R below 0. Similar results were shown in Chapter 3 of this thesis and in (Gamazon *et al.*, 2015; Wheeler *et al.*, 2016), suggesting that the prediction values observed here are realistic. As the accuracy with which gene expression can be predicted is related to the subsequent power for that gene in a test of association between predicted expression and phenotype, these results suggest that the TWAS power for many genes would be quite low. However, the power to detect an association between predicted expression and phenotype in a TWAS relies not only on the prediction accuracy of gene expression, but also on the sample size of the GWAS data being used. Thus, an association can still be detected for genes where expression cannot be predicted accurately, if the GWAS data being used has a sufficient sample size. Given that the sample sizes used in GWAS are ever-increasing, more and more associations for poorly predicted genes will likely be detected in the future.

One reason that the prediction accuracy estimates may be so low is that there is not much variation in gene expression attributable to the SNPs used for generating

prediction models. Indeed, the prediction accuracy estimates found here were mostly concordant with, although marginally smaller than, estimates of gene expression heritability calculated with GCTA, suggesting that the gene expression prediction models were performing nearly as well as could be expected. On average, the prediction accuracy estimates were smaller than the estimates of heritability, suggesting there may be room for improvement to the prediction accuracy estimates observed here. It is also worth noting that GCTA assumes that every SNP has an effect on the phenotype when estimating heritability, and so is polygenic. Given that the local architecture of gene expression is thought to be sparse, it is possible that GCTA is in fact underestimating the heritability, which would suggest that there is more room for improvement to the prediction accuracy estimates observed here. Further investigation using a heritability estimation model better suited to the data would be required to determine this.

Following the examination of prediction accuracy itself, the role of a number of factors on prediction accuracy was examined. The first to be examined was sample size. Unsurprisingly, the average prediction accuracy observed from 10-fold cross-validation increased as the sample size used to train the gene expression prediction models increased. Similar results have been found when examining the prediction of other complex traits (Wei *et al.*, 2013; Guo *et al.*, 2016). This highlights the need for development of reference panels with larger sample sizes to allow the training of more accurate gene expression prediction models.

Currently, the most popular TWAS software packages mostly use GTEx as the reference panel for prediction model training. Although GTEx has matched genotype and expression data for many tissues, the sample size available in GTEx is often quite small, meaning that the TWAS power in many tissues is small. By increasing the sample sizes available in GTEx, gene expression prediction models would be able to predict expression with greater accuracy, improving TWAS power. Additionally, this increase in prediction accuracy would likely increase the number of genes that could be examined in a TWAS. For example, consider the PrediXcan package, which only makes gene expression prediction models that are able to predict gene expression with accuracy above a certain threshold publicly available. By increasing sample size and thus prediction accuracy, more genes would likely pass this threshold and be included in the set of publicly available prediction models, allowing greater gene discovery in future TWAS.

A further reason to increase the sample size of reference panels is that it will allow for the inclusion of more distal SNPs in gene expression prediction models. The sample size currently available in reference panels such as GTEx prohibits the inclusion of SNPs across the whole genome in gene expression prediction models, so the effects of SNPs distal to the gene on expression are missed. Given that these distal SNPs are known to have effects on gene expression (Brynedal *et al.*, 2017), and many of these distal effects are thought to be key drivers of SNP - disease associations (Westra *et al.*, 2013; Kirsten *et al.*, 2015), their inclusion in prediction models would likely facilitate additional gene discovery through TWAS. Thus, increasing sample size to the point where these effects can be modelled would seem beneficial. However it is worth considering that because many of these distal effects act on multiple genes simultaneously (Brynedal *et al.*, 2017), their inclusion in gene expression prediction models also increases the likelihood of co-prediction of multiple genes by a single prediction model. This could lead to the identification of spurious TWAS associations, and so would lead to difficulty in interpretation of TWAS results.

The next factor to be examined was ancestry. Overall, a reduction in the average prediction accuracy was observed when using gene expression prediction models trained using data from samples of one ancestry to predict into samples of a different ancestry (compared to within-population prediction). This corroborates findings from (Mogil *et al.*, 2018) and (Mikhaylova and Thornton, 2019), and suggests that correct population matching will lead to better prediction accuracy and improve TWAS power. This reduction in average prediction accuracy may have been caused by a number of factors, including differences in linkage disequilibrium between the different populations, differences in data processing between the samples from different populations, differences in the SNPs used for prediction, or differences in allele frequencies between the populations as suggested by (Mogil *et al.*, 2018). GTEx, the most popular resource used to generate gene expression prediction models, consists primarily of samples of European descent. Using resources such as this may lead to inaccurate prediction of gene expression for samples of non-European ancestry, and even the detection of spurious TWAS associations due to mismatched linkage disequilibrium. While some effort has been made to create prediction models using data from non-European populations (Mogil *et al.*, 2018), more population-specific reference panels (especially with gene expression data

gathered from tissues other than whole blood) will need to be developed to allow better TWAS in populations of non-European ancestry.

Predictions accuracy estimates from 10-fold cross-validation were then compared with those achieved by the prediction models from the PrediXcan software package. Overall, the prediction accuracy estimates achieved by the prediction models from PrediXcan were highly similar, although on average slightly smaller than those achieved by 10-fold nested cross-validation. This may have been the result of slight differences in data collection and processing between the Geuvadis data and the GTEx data used to train PrediXcan models, or may have been caused by SNP missingness between the models and the Geuvadis data. Regardless of the reason, the result suggests that using data as similar as possible to the intended "test" data for the training of gene expression prediction models would likely give more accurate predictions. This may be a realistic prospect for large consortia where gene expression has been measured in a subset of the samples. However, even in circumstances such as these, it would still be important to consider the potential trade-off between using data as similar as possible to the intended test data, and using gene expression prediction models potentially trained using a larger sample size (which may be achieved by using a standard reference panel such as GTEx).

Finally, the issue of prediction of gene expression across tissues was examined. The average prediction accuracy was at its greatest when the data used to train the gene expression prediction models was derived from the same tissue as the data used to test the gene expression prediction models, corroborating similar findings in (Gamazon *et al.*, 2015). Although, when digging a bit deeper, for many genes the prediction models trained using data from a non-LCL tissue were able to predict Geuvadis gene expression with similar accuracy to the prediction models trained using LCL expression data. Given that many eQTLs show evidence of concordance of effects across many different tissues (GTEx Consortium, 2015), this result is perhaps not surprising. However, it is reassuring, as it implies that for some genes using gene expression prediction models from the "wrong" tissue can still provide a similar degree of prediction accuracy as if the "correct" tissue were used, meaning that TWAS power is likely to be similar. Despite this, there was a group of genes where the prediction models trained using data from a non-LCL tissue clearly performed more poorly than the models trained using LCL data. In these instances, using gene expression prediction models trained in the wrong tissue may result in

poorer prediction accuracy and reduced power in TWAS. Given this result, and considering that it would not be known a priori whether any given gene would be one that could proxy well or not, ideally it would be best to attempt to carefully match the tissue of the prediction model to the intended test data. The best publicly available resource for this would seem to be GTEx, which currently has data available for over 50 tissues, allowing the training of many tissue-specific sets of gene expression prediction models. Continuing improvements to GTEx by gathering more samples, especially those from non-EUR ancestry, and data from tissues not currently collected by GTEx would make this resource even better.

It is worth noting that this analysis is biased in a number of ways. The gene expression prediction models downloaded from predictdb.org and used in this analysis were a filtered set of gene expression prediction models that all showed prediction accuracy above a certain threshold in the model training phase. Thus, genes where expression could not be predicted with sufficient accuracy were not included in this analysis. These poorly predicted genes that were not tested here may behave differently to those well-predicted genes that were examined here. Further study with more data would be required to determine this. Furthermore, the Geuvadis LCL data were used to test the ability of all these prediction models to predict into another tissue. Thus, inference can only be made regarding how well these models predict into LCL data, not into data from other tissues.

In conclusion, this chapter has demonstrated that methods with assumptions of sparsity tend to achieve the best gene expression prediction accuracy estimates, although for the majority of genes these estimates are still low. This chapter has also demonstrated that increases to sample size and matching of ancestry and tissue between the data used to generate prediction models and the prediction model testing data improves the prediction accuracy. Further increases in the sample size of reference panels and gathering of samples from multiple ancestries and tissues will help to improve gene expression prediction accuracy, and thus improve the power of future TWAS.

Chapter 5. Investigation of Prediction of CpG methylation From SNP Genotypes

Although the TWAS methodology has mainly been used to explore the role of gene expression in complex traits, the same methodology can theoretically be used to explore the role that any other intermediate trait plays in complex traits, provided that trait is under genetic regulation. One such trait is DNA methylation, which is known to regulate gene expression and has been shown to affect complex traits. The power to detect associations between predicted DNA methylation and complex traits in a TWAS-like approach would depend partially on how accurately DNA methylation could be predicted from SNP genotypes. In this chapter, I investigate how accurately DNA methylation can be predicted by training and testing DNA methylation prediction models using data from ARIES and Understanding Society.

The ARIES and Understanding Society data sets used in this chapter are described in detail in Chapter 2.

5.1 Comparison of statistical methods for prediction of CpG methylation levels from SNP genotypes

First, the ability of three penalised regression approaches to predict CpG methylation from local SNP genotypes was compared. These three approaches were: LASSO, elastic net (with α set to 0.5) and ridge regression. Although the elastic net with alpha determined by cross-validation outperformed these methods for prediction of gene expression from SNP genotypes in Chapter 4, this method was not used here due to time constraints. To compare the three approaches, 50% of ARIES samples were designated as the prediction model training set, and 20% as the test set. For each CpG site, CpG methylation prediction models were trained using samples from the training set, using each of the three methods, using all SNPs within 1 Mb of the CpG site. For each of the three methods, the lambda parameter was determined by 10-fold cross-validation within the training set. Only values of lambda at which the final prediction model contained at least one SNP were considered, and the value of lambda was chosen as that at which the minimum mean squared error was achieved in the 10-fold cross-validation. The prediction models were then applied to the test

set, and the Pearson correlation coefficient (R) between predicted and observed methylation levels was calculated.

Overall, LASSO and elastic net performed highly similarly to each other, while both marginally outperformed ridge regression (Figure 5.1A). Across the CpG sites that were successfully modelled with all three methods, higher average prediction accuracy estimates were achieved with LASSO (mean R = 0.0973) and elastic net (mean R = 0.0975) than with ridge regression (mean R = 0.0799). In addition, estimates from all three methods were highly correlated (Figure 5.1B), with all pairwise correlations greater than 0.86.

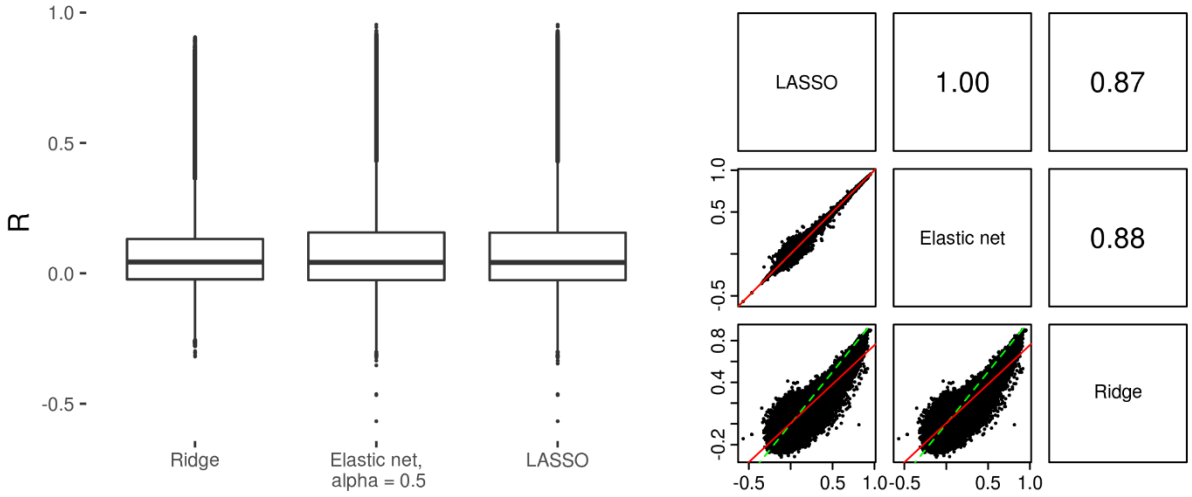

Figure 5.1. Comparison of penalised regression approaches for predicting CpG methylation. (A) Box plots of prediction accuracy estimates (R) from training and testing prediction models using 3 forms of penalised regression (ridge regression, elastic net, LASSO) on ARIES data. The line within the box represents the median, with the edges of the box the upper and lower quartiles. (B) Correlation plots between prediction accuracy estimates achieved using the 3 penalised regression approaches. In the lower panels, each point represents a CpG site, with the R achieved by 2 methods displayed on the axes. Also shown are the line of equality (green dashed line) and a best fit line between x and y (red solid line). Upper panels show the pairwise correlations between the R values achieved using the 3 methods.

When looking at the results on a CpG-by-CpG basis, methylation could not be predicted accurately at most CpG sites examined here, with only 10,004 CpG sites showing a prediction accuracy estimate greater than 0.5 from any of the three methods. When looking at these 10,004 CpGs for which an R estimate ≥ 0.5 was achieved by any of the 3 methods, the difference between the sparse methods and the polygenic method was clear, with the sparse methods outperforming the

polygenic methods even more strongly than observed previously (Figures 5.2A and 5.2B).

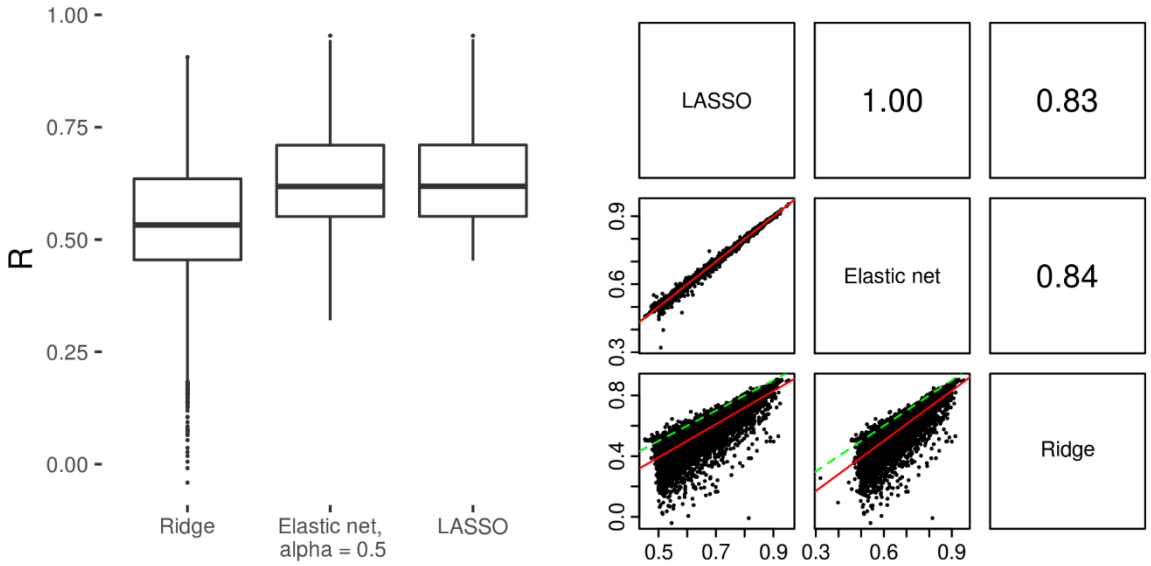

Figure 5.2. Comparison of penalised regression approaches for predicting CpG methylation for well predicted CpG sites. Displayed are CpGs that achieved R estimates of 0.5 or greater from any of the 3 penalised regression approaches. (A) Box plots of prediction accuracy estimates (R) from training and testing prediction models using 3 forms of penalised regression (ridge regression, elastic net, LASSO) on ARIES data. The line within the box represents the median, with the edges of the box the upper and lower quartiles. (B) Correlation plots between prediction accuracy estimates achieved using the 3 penalised regression approaches. In the lower panels, each point represents a CpG site, with the R achieved by 2 methods displayed on the axes. Also shown are the line of equality (green dashed line) and a best fit line between x and y (red solid line). Upper panels show the pairwise correlations between the R values achieved using the 3 methods.

5.2 Comparison of window sizes for prediction of CpG methylation levels from SNP genotypes

In the (Gaunt *et al.*, 2016) paper in which the ARIES data is described, the authors identified a number of trans-mQTLs, where SNPs distal to a CpG site are associated with its methylation levels. To investigate whether incorporating SNPs more distal to CpG sites into methylation prediction models could improve their prediction accuracy, CpG methylation prediction models were trained using a number of different window sizes (the distance from the CpG by which SNPs are selected for modelling). To do this, CpG methylation prediction models were trained by regressing CpG methylation on genotypes of all SNPs within either 250Kb, 500Kb, 1Mb, 2Mb or 3Mb upstream or downstream of the CpG site. This analysis was conducted using elastic net with

alpha set to 0.5 and with lambda determined by 10-fold cross-validation as in section 5.1. This analysis also used the same training and testing sets as used in section 5.1.

Overall, the prediction accuracy estimates obtained at the five window sizes were highly correlated with one another (Figures 5.3A and 5.3B), with all pairwise correlations greater than 0.78. On average, slightly greater prediction accuracy estimates were achieved at the smaller window sizes (Table 5.1), although when looking at the results on a CpG-by-CpG basis, no clear pattern was observed. Indeed, some CpG sites showed greater prediction accuracy at the larger window sizes, while other CpG sites showed greater prediction accuracy at the smaller window sizes. This indicates that the optimal window size for training CpG methylation prediction models appears to be CpG-specific.

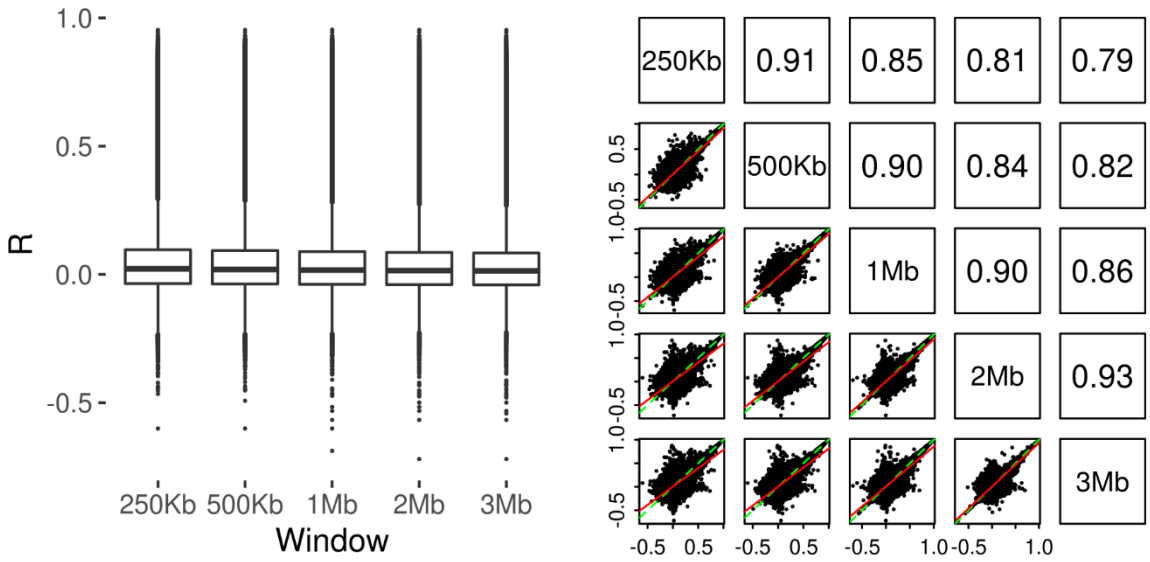

Figure 5.3. Comparison of window sizes for predicting CpG methylation. (A) Box plots of prediction accuracy estimates (R) from training and testing prediction models using elastic net with SNPs selected using 5 window sizes (250Kb, 500Kb, 1Mb, 2Mb and 3Mb) on ARIES data. The line within the box represents the median, with the edges of the box the upper and lower quartiles. (B) Correlation plots between prediction accuracy estimates achieved using the 5 window sizes. In the lower panels, each point represents a CpG site, with the R achieved at the 2 window sizes displayed on the axes. Also shown are the line of equality (green dashed line) and a best fit line between x and y (red solid line). Upper panels show the pairwise correlations between the R values achieved at the 5 window sizes.

Window size	Average prediction accuracy
250Kb	0.0548
500Kb	0.0525
1Mb	0.0502
2Mb	0.0481
3Mb	0.0467

Table 5.1. Average prediction accuracy estimates achieved when training and testing CpG methylation prediction models using five different window sizes.

When focussing on the set of CpG sites at which a prediction accuracy estimate of 0.5 had been achieved with at least one of the window sizes tested, a similar conclusion was reached. Overall, the prediction accuracy estimates tended to be similar for the majority of CpG sites, although there was a set of CpG sites where prediction accuracy was clearly better at some window sizes than others, further indicating that the optimal window size for training CpG methylation prediction models is a CpG-specific quantity (Figures 5.4A and 5.4B). Interestingly, when considering these well-predicted CpG sites, while the same trend in the average prediction accuracy was observed as when considering all CpG sites, the differences between the average prediction accuracy estimates were far smaller than when considering all CpG sites, regardless of prediction accuracy (Table 5.2).

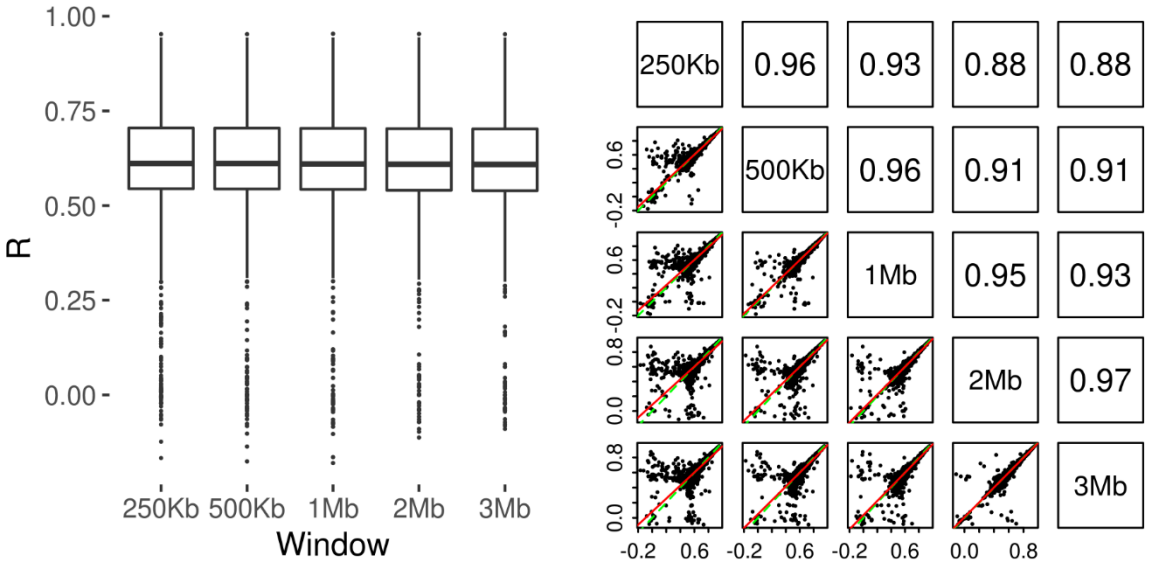

Figure 5.4. Comparison of window sizes for predicting CpG methylation for well-predicted CpG sites. Displayed are CpGs that achieved R estimates of 0.5 or greater from any of the 5 window sizes. (A) Box plots of prediction accuracy estimates (R) from training and testing prediction models using

elastic net with SNPs selected using 5 window sizes (250Kb, 500Kb, 1Mb, 2Mb and 3Mb) on ARIES data. The line within the box represents the median, with the edges of the box the upper and lower quartiles. (B) Correlation plots between prediction accuracy estimates achieved using the 5 window sizes. In the lower panels, each point represents a CpG site, with the R achieved at the 2 window sizes displayed on the axes. Also shown are the line of equality (green dashed line) and a best fit line between x and y (red solid line). Upper panels show the pairwise correlations between the R values achieved at the 5 window sizes.

Window size	Average prediction accuracy
250Kb	0.6279
500Kb	0.6286
1Mb	0.6276
2Mb	0.6267
3Mb	0.6256

Table 5.2. Average prediction accuracy estimates achieved when training and testing CpG methylation prediction models using five different window sizes, focussing on CpG sites at which a prediction accuracy estimate of greater than 0.5 was achieved at one or more of the window sizes.

Having determined that the optimal window size for predicting CpG methylation was CpG-specific, the same comparison of five window sizes was performed on the Understanding Society data to identify optimal window sizes for those CpG methylation measurements. A random 50% of the Understanding Society samples were designated as the prediction model training set, with a random 20% of samples assigned to the testing set. This analysis was conducted using elastic net with alpha set to 0.5 and with lambda determined by 10-fold cross-validation as in section 5.1.

As observed when using data from ARIES, prediction accuracy estimates at the five window sizes were highly correlated with one another (Figure 5.5), with all pairwise correlations greater than 0.87. Again, some CpG sites showed greater prediction at larger window sizes, while others showed greater accuracy at the smaller window sizes, reinforcing the conclusion that the optimal window size for CpG methylation prediction model training is CpG-specific.

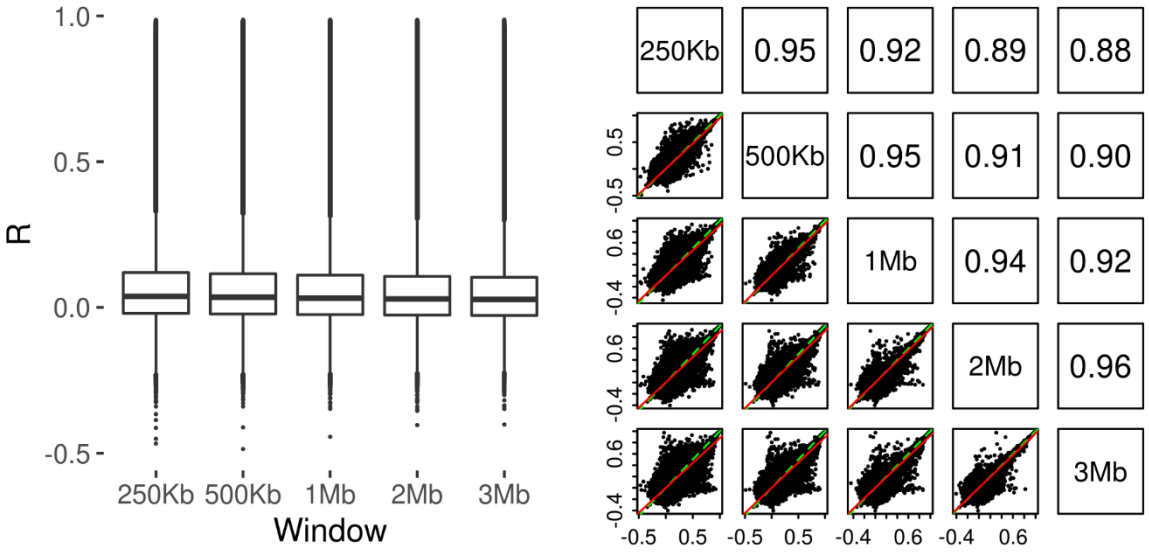

Figure 5.5. Comparison of window sizes for predicting CpG methylation using data from Understanding Society. (A) Box plots of prediction accuracy estimates (R) from training and testing prediction models using elastic net with SNPs selected using 5 window sizes (250Kb, 500Kb, 1Mb, 2Mb and 3Mb) on ARIES data. The line within the box represents the median, with the edges of the box the upper and lower quartiles. (B) Correlation plots between prediction accuracy estimates achieved using the 5 window sizes. In the lower panels, each point represents a CpG site, with the R achieved at the 2 window sizes displayed on the axes. Also shown are the line of equality (green dashed line) and a best fit line between x and y (red solid line). Upper panels show the pairwise correlations between the R values achieved at the 5 window sizes.

When considering only the CpG sites where methylation was well predicted using at least one of the window sizes, the same conclusion was reached, with some CpG sites clearly showing much greater prediction accuracy at the smaller window sizes, while others showed a much greater accuracy when using a larger window size (Figure 5.6).

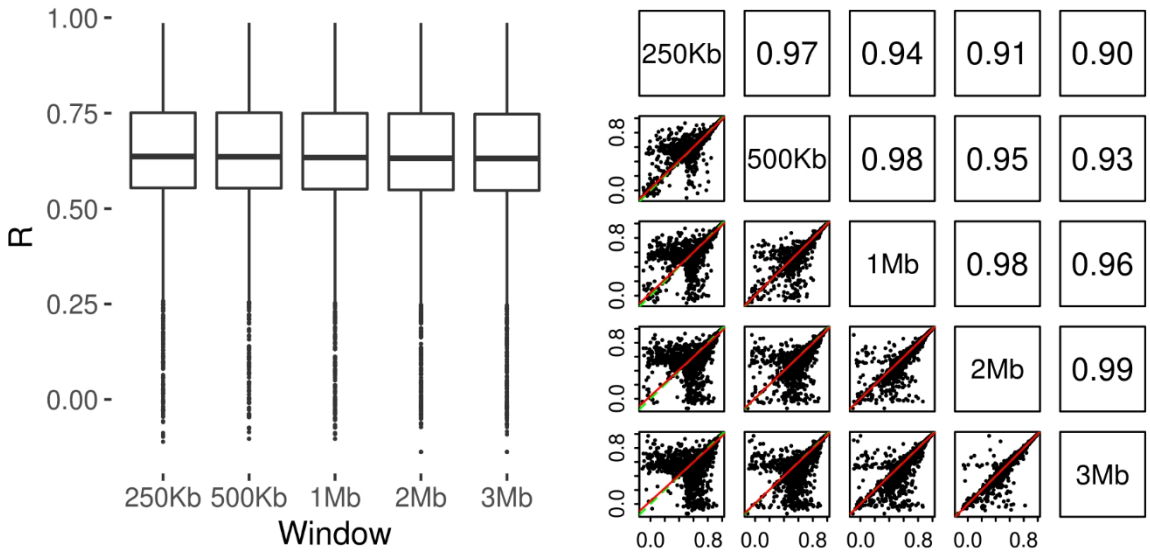

Figure 5.6. Comparison of window sizes for predicting CpG methylation for well-predicted CpG sites using data from Understanding Society. Displayed are CpGs that achieved R estimates of 0.5 or greater from any of the 5 window sizes. (A) Box plots of prediction accuracy estimates (R) from training and testing prediction models using elastic net with SNPs selected using 5 window sizes (250Kb, 500Kb, 1Mb, 2Mb and 3Mb) on ARIES data. The line within the box represents the median, with the edges of the box the upper and lower quartiles. (B) Correlation plots between prediction accuracy estimates achieved using the 5 window sizes. In the lower panels, each point represents a CpG site, with the R achieved at the 2 window sizes displayed on the axes. Also shown are the line of equality (green dashed line) and a best fit line between x and y (red solid line). Upper panels show the pairwise correlations between the R values achieved at the 5 window sizes.

For both the ARIES and Understanding Society data, the optimal window size for each CpG was then determined as the window size at which the maximum correlation between predicted and measured methylation was observed in the test set.

5.3 Evaluation of CpG methylation prediction accuracy using the optimal method and window size

Having identified an optimal method (elastic net) and window size for prediction model training, CpG methylation prediction models were next trained using the ARIES data and the Understanding Society data sets. The 50% of samples that had been used as a prediction model training set and the 20% that had been used for prediction model testing in Chapter 5.1 were combined, and were then used as the prediction model training set here. The remaining 30% of samples that had not been

used prior to this point were used as a prediction model testing set. The same procedure was used to generate training and testing sets with the Understanding Society data. Prediction models were then trained on each training set using elastic net, with alpha set to 0.5 and lambda determined by 10-fold cross-validation on the training set as described in section 5.1, and using the CpG-specific window size determined in Chapter 5.2. The prediction models were then assigned to their respective testing sets (i.e. ARIES-trained models applied to the ARIES testing set), and the Pearson correlation coefficient between predicted and measured methylation was calculated.

Within each data set, methylation at most CpG sites could not be accurately predicted from SNP genotypes, reinforcing findings from Chapter 5.1 (Figure 5.7). Yet there existed a subset of CpG sites for which methylation could be predicted with some accuracy. In total, 10,220 ARIES-trained models and 30,865 Understanding Society-trained models showed R ≥ 0.5 in their respective test sets, representing a set of well-predicted CpG sites.

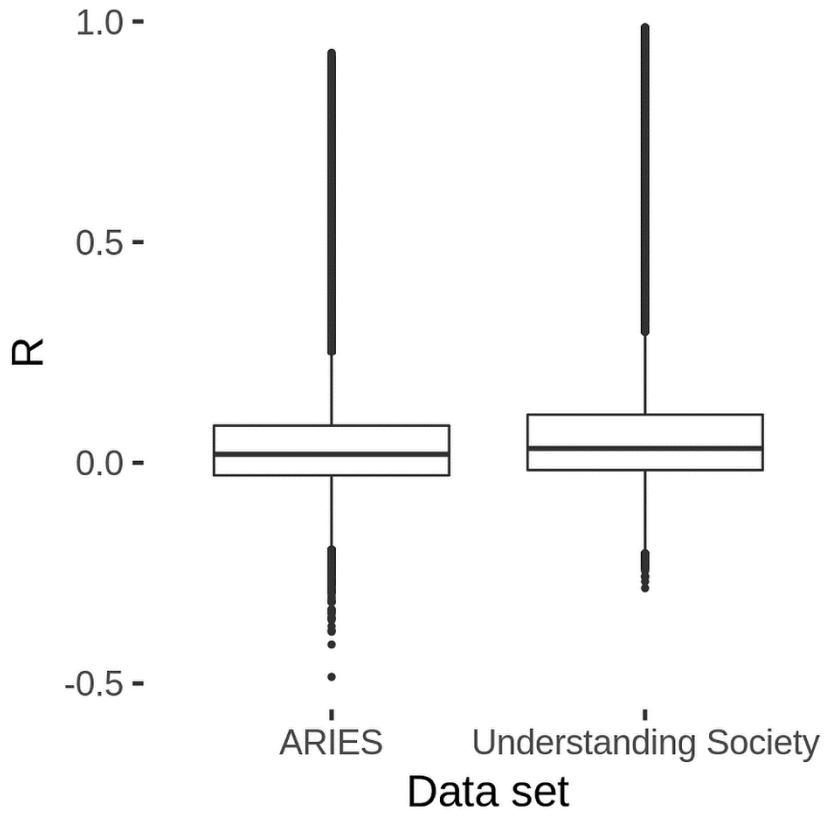

Figure 5.7. Prediction accuracy of CpG methylation prediction models trained using elastic net with alpha = 0.5 and with a CpG-specific window size. Box plots of prediction accuracy estimates

(R) from training and testing prediction models using elastic net (with alpha = 0.5) with a CpG-specific window size using data from ARIES and Understanding Society. The line within the box represents the median, with the edges of the box the upper and lower quartiles.

As a sanity check, estimates of prediction accuracy obtained from application of ARIES-trained prediction models to the ARIES testing set were next compared with the estimates of prediction accuracy from application of Understanding Society-trained prediction models to the Understanding Society test set. Overall, the estimates of prediction accuracy from the two data sets were highly correlated with each other (r = 0.757), indicating broad agreement and providing confidence in the estimates obtained here (Figure 5.8). This agreement was especially strong for the CpG sites that could be predicted with a high degree of accuracy. For example, the ARIES-trained model for cg16906346 showed a prediction accuracy of 0.924, while the Understanding Society-trained model for the same CpG showed a prediction accuracy estimate of 0.953. When considering the set of CpGs available in both the ARIES and Understanding Society data sets, the average prediction accuracy achieved was slightly greater when using the Understanding Society data (mean R = 0.0708) than when using the ARIES data (mean R = 0.0550), potentially reflecting the larger sample size of the Understanding Society data.

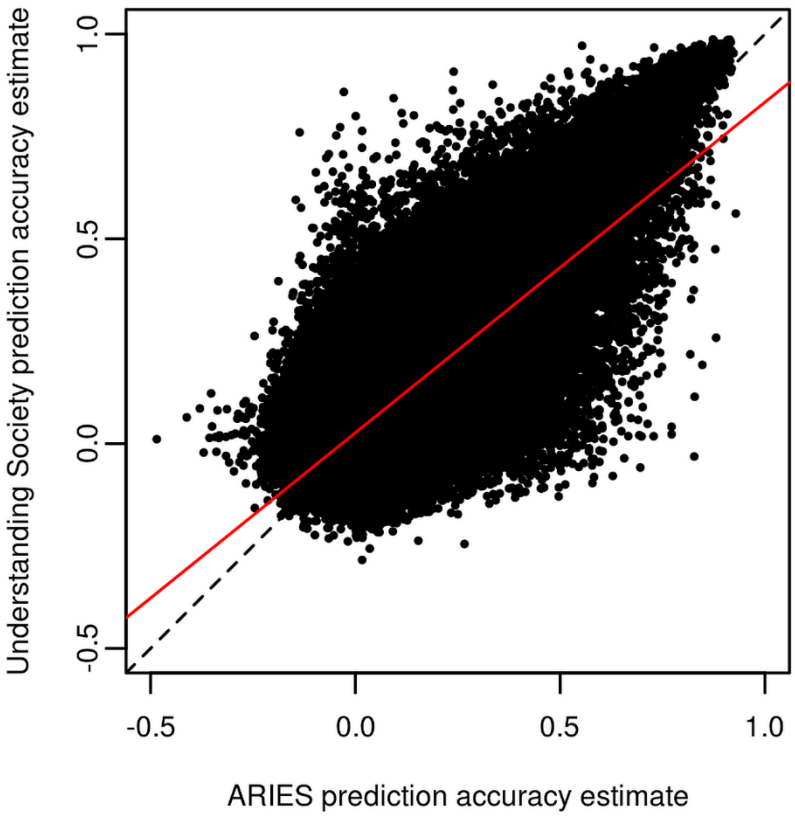

Figure 5.8. Comparison of prediction accuracy estimates from ARIES and Understanding Society data sets. Each point represents a CpG site, with its prediction accuracy estimate obtained from training and testing a prediction model using the ARIES data shown on the x axis, and its prediction accuracy estimate obtained from training and testing a prediction model using the Understanding Society data shown on the y axis. The red line represents a best fit line, and the dashed line represents the line of equality (y=x).

To learn more about the CpG sites where methylation could be predicted well from SNP genotypes, CpG sites were annotated using the manifest files for the 450k and EPIC chips used to measure methylation in the ARIES and Understanding Society studies respectively. Enrichment of key annotations among the well-predicted CpG sites was then tested using two-sided Fisher's exact tests. Overall, CpG sites tagged to genes (OR = 0.682, $p = 4.08 \times 10^{-68}$), CpG sites located at CpG islands (OR = 0.883, $p = 1.72 \times 10^{-9}$) and CpG sites tagged to promoters (OR = 0.636, $p = 1.13 \times 10^{-61}$) were all depleted among the set of well-predicted CpG sites (when compared to the background set of all CpG sites), while CpGs at enhancer regions (OR = 1.42, $p = 9.62 \times 10^{-53}$) and CpGs at DNAse 1 hypersensitivity sites (OR = 1.40, $p = 8.57 \times 10^{-33}$) were enriched among the well-predicted CpG sites. Crucially, these enrichments and depletions were replicated when looking at those CpG sites that were well-predicted

when using the Understanding Society data (compared to all CpGs on the EPIC chip) (Figure 5.9).

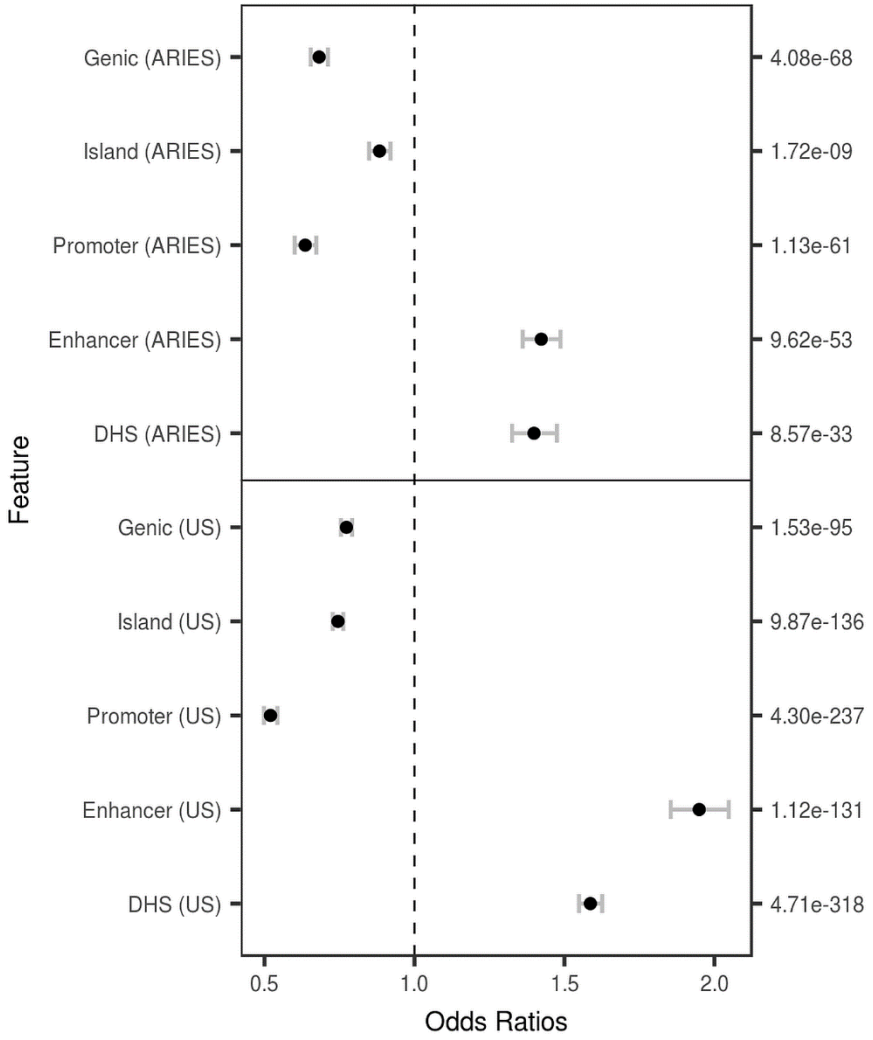

Figure 5.9. Enrichment of CpG annotations among CpGs tested in MWAS. CpG sites were annotated using manifest files downloaded from the Illumina website. For each annotation, enrichment of well-predicted (R >= 0.5) CpGs against the background of all CpGs on either the 450k or EPIC chip was tested using a two-sided Fisher's exact test. Odds ratios and 95% confidence intervals are shown on the x axis, with p values shown on the right.

5.4 Estimation of the heritability of CpG methylation

The maximum accuracy with which CpG methylation can be predicted from SNP genotypes is equivalent to its heritability. Thus, to establish the potential maximum accuracy achievable here, the heritability of CpG methylation attributable to local

SNPs was estimated using REML analysis as implemented in GCTA. For each CpG site, the heritability of methylation was estimated using all SNPs within the CpG-specific window size as determined in Chapter 5.2. The resulting heritability estimates were then compared with the prediction accuracy estimates obtained from application of CpG methylation prediction models to their testing sets in Chapter 5.3.

When using data from ARIES, CpG methylation prediction accuracy estimates were highly correlated and concordant with estimates of heritability for most CpG sites (Figure 5.10). On average, heritability estimates of CpG methylation were slightly greater (mean heritability = 0.0306) than squared prediction accuracy estimates (mean squared prediction accuracy estimate = 0.0253), suggesting that there was some potential room for improvement to the prediction accuracy estimates observed here. Interestingly, there was a small set of CpG sites for which high heritability, but low prediction accuracy estimates were achieved.

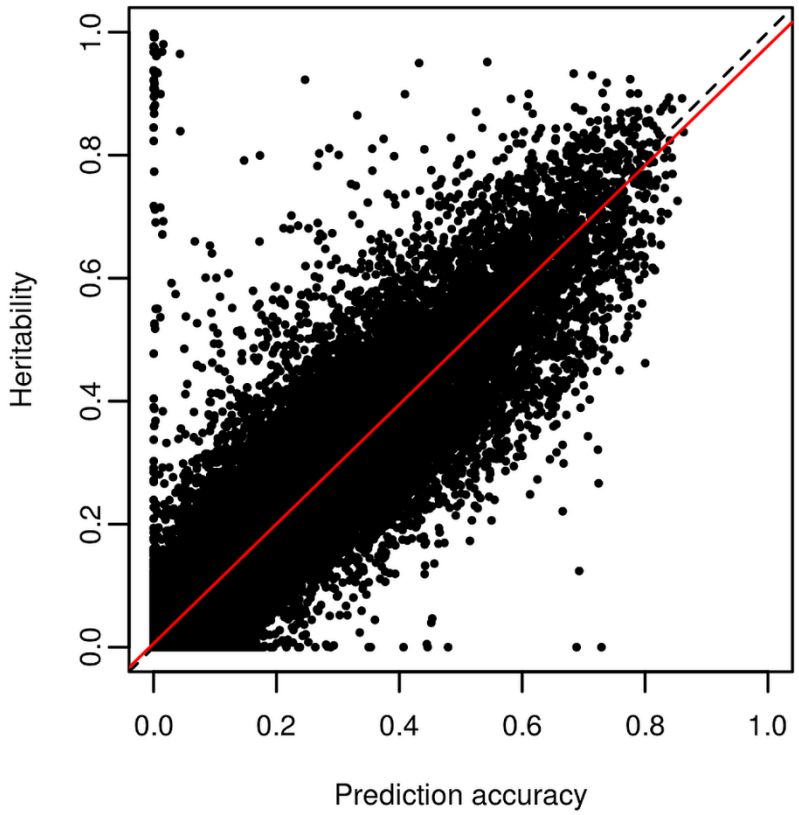

Figure 5.10. Comparison of CpG methylation prediction accuracy estimates with estimates of the heritability of CpG methylation, using data from ARIES. Each point represents a CpG site, with its prediction accuracy estimate obtained from training and testing a prediction model using the ARIES

data shown on the x axis, and its heritability estimate obtained using GCTA REML analysis on the ARIES data shown on the y axis. The red line represents a best fit line, and the dashed line represents the line of equality (y=x). In this plot, prediction accuracy is shown as R^2 rather than R, as the heritability is the upper bound on the estimate of R^2.

When using data from Understanding Society, again CpG methylation prediction accuracy estimates were highly correlated with the heritability estimates (Figure 5.11). However, there was a greater spread in the data, with more CpG sites showing a greater heritability estimate than the prediction accuracy estimate. Furthermore, the average heritability estimate (mean heritability = 0.0411) was again greater than the average prediction accuracy estimate (mean squared prediction accuracy estimate = 0.0344), indicating that there was even greater potential room for improvement with the Understanding Society data than seen with the ARIES data.

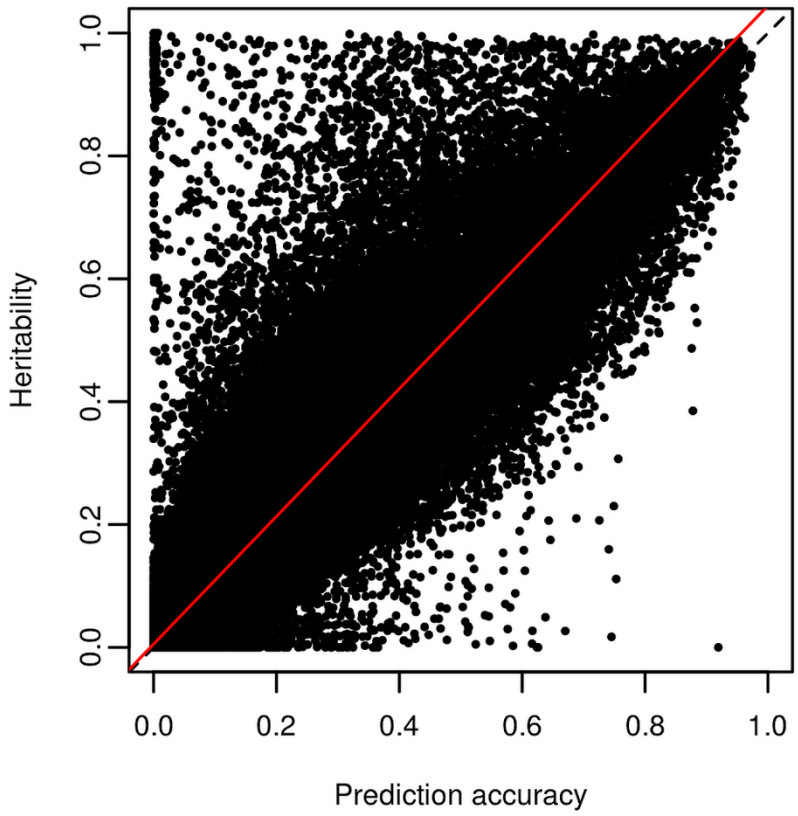

Figure 5.11. Comparison of CpG methylation prediction accuracy estimates with estimates of the heritability of CpG methylation, using data from Understanding Society. Each point represents a CpG site, with its prediction accuracy estimate obtained from training and testing a prediction model using the Understanding Society data shown on the x axis, and its heritability estimate obtained using GCTA REML analysis on the Understanding Society data shown on the y

axis. The red line represents a best fit line, and the dashed line represents the line of equality (y=x). In this plot, prediction accuracy is shown as R^2 rather than R, as the heritability is the upper bound on the estimate of R^2.

5.5 Training and validating a final set of CpG methylation prediction models

Having obtained an estimate of prediction accuracy using the optimal method and window size for each CpG site, the final step was to train a set of CpG methylation prediction models that could be used in an MWAS to investigate the relationship between predicted CpG methylation and complex traits. As shown in Chapter 4.5, there is a clear relationship between the sample size of the data set used to train a prediction model and the prediction accuracy of the resulting model. While Chapter 4.5 demonstrated this relationship was true for the prediction of gene expression, it has also been shown to be true for the prediction of a number of other complex traits (Wei *et al.*, 2013; Guo *et al.*, 2016), and so it seemed likely to be true for the prediction of CpG methylation too. Given this, in order to maximise the prediction accuracy of these final CpG methylation prediction models (and thus improve the power of the subsequent tests of association between predicted methylation and phenotype), the sample size of the prediction model training set was increased by combining the 70% training set and 30% testing set used in Chapter 5.3. This resulted in two training sets (one comprised of ARIES data and one of Understanding Society data), each consisting of 100% of their samples.

In total, 78,250 CpG sites from ARIES and 207,525 CpG sites from Understanding Society showed a prediction accuracy >= 0.1 when predicted using the optimal method and window size in Chapter 5.3. For these CpG sites, a prediction model was then trained using their respective training set, resulting in 78,250 CpG methylation prediction models trained using 100% of the ARIES data and 207,525 CpG methylation prediction models trained using 100% of the Understanding Society data.

As a validation step, the 78,250 CpG methylation prediction models trained using the ARIES data were next applied to Understanding society genotype data, and the Pearson correlation coefficient between the predicted and observed methylation was calculated. In total, only 68,230 of the 78,250 ARIES-trained could actually be validated through this approach, as the CpG sites for the remaining 10,020 ARIES-

trained CpG methylation prediction models were not measured (or did not pass QC) in the Understanding Society data.

For many of the CpG sites that underwent this validation procedure, the estimates of prediction accuracy achieved in validation were similar to those achieved in Chapter 5.3 (Figure 5.12). This was especially true for the well-predicted CpG sites, such as cg08103988, which showed a prediction accuracy of 0.873 when a prediction model trained using 70% of ARIES samples was applied to the remaining 30% of ARIES samples, and a prediction accuracy estimate of 0.984 when a prediction model trained using all 100% of ARIES samples was applied to the Understanding Society data. Overall, the two sets of prediction accuracy estimates showed a strong correlation (r = 0.778). Despite this, there also existed a small group of CpG sites that showed a strong prediction accuracy in Chapter 5.3, but a poor prediction accuracy in the validation step, raising doubts about how reliable the prediction accuracy estimates were for these CpG sites. In total, 42,371 of the 68,230 prediction models trained using 100% of ARIES samples showed a prediction accuracy >= 0.1 in the validation step, so were taken forward to the next stage of the analysis. The 25,859 remaining CpG methylation prediction models that did not meet this prediction accuracy threshold in the validation step were discarded at this stage.

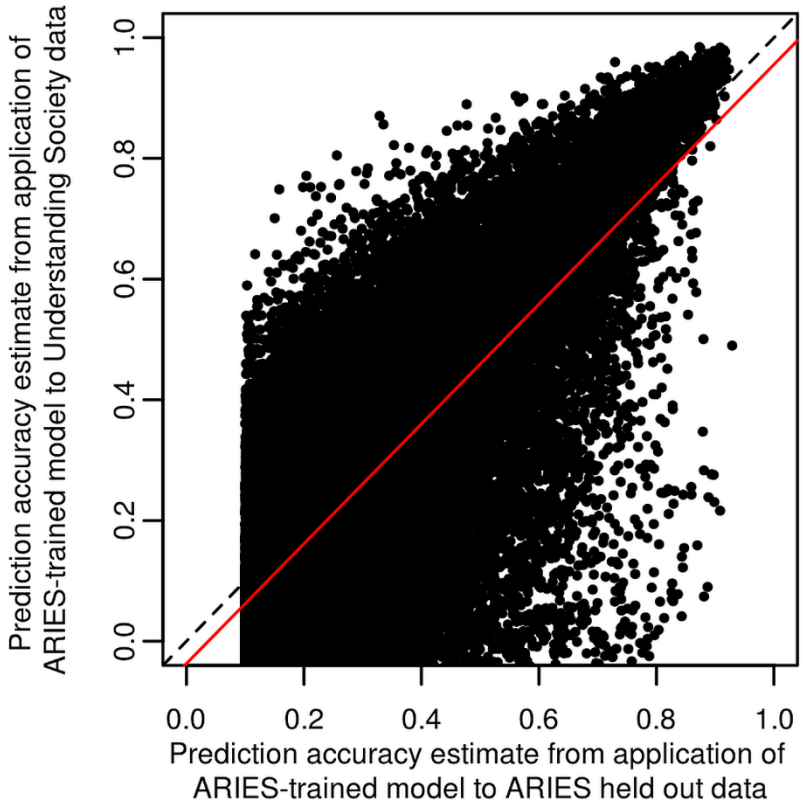

Figure 5.12. Validation of CpG methylation prediction models trained using ARIES data. Shown are comparisons between prediction accuracy estimates from application prediction models trained using 70% of ARIES data to the remaining 30% of ARIES data (x axis) and from application of prediction models trained using 100% of ARIES data to 100% of the Understanding Society data (y axis). On each plot, each point represents a CpG site, the dashed line represents the line of equality, and the solid red line is a best fit line.

Similarly, the 207,525 Understanding Society-trained models were applied to genotype data for the ARIES samples, and the Pearson correlation between predicted and observed methylation was calculated. In total, only 81,353 of the 207,525 Understanding Society-trained models could actually be validated through this approach, as the CpG sites for the remaining 126,172 Understanding Society-trained CpG methylation prediction models were not measured (or did not pass QC) in the ARIES data set.

Again, for most of the CpG sites that underwent this validation procedure, the estimates of prediction accuracy achieved in the validation stage were similar to those achieved in Chapter 5.3 (Figure 5.13). This was especially true for the well-predicted CpG sites, such as cg09035930, which showed a prediction accuracy of 0.980 when a prediction model trained using 70% of Understanding Society samples was applied to the remaining 30% of Understanding Society samples, and a

prediction accuracy estimate of 0.922 when a prediction model trained using all 100% of Understanding Society samples was applied to the ARIES data. Overall, correlation between the two sets of prediction accuracy estimates was strong (r = 0.845), although it is worth noting that there were a number of CpG sites where a strong prediction accuracy was achieved in one analysis, but not another. In total, 53,793 Understanding Society-trained models showed a prediction accuracy ≥ 0.1 in the validation step, so were taken forward to the next stage of the analysis. The 27,560 CpG methylation prediction models that did not meet this prediction accuracy threshold were discarded at this stage.

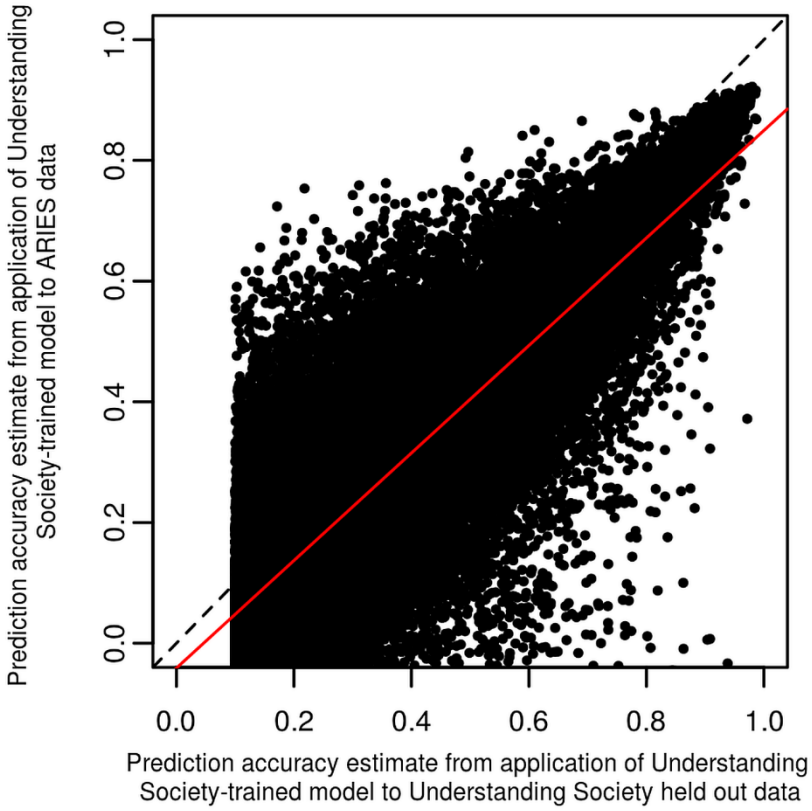

Figure 5.13. Validation of CpG methylation prediction models trained using Understanding Society data. Shown are comparisons between prediction accuracy estimates from application prediction models trained using 70% of Understanding Society data to the remaining 30% of Understanding Society data (x axis) and from application of prediction models trained using 100% of Understanding Society data to 100% of the ARIES data (y axis). On each plot, each point represents a CpG site, the dashed line represents the line of equality, and the solid red line is a best fit line.

In addition to the 42,371 ARIES-trained and 53,793 US-trained prediction models that passed this validation step, the 10,020 ARIES-trained and 126,172

Understanding Society-trained models that could not be validated were also taken forward, giving a total of 232,356 prediction models. These 232,356 prediction models covered 193,315 unique CpG sites, with 39,041 CpG sites represented by both an ARIES-trained and an Understanding Society-trained prediction model.

One factor that potentially impacted the prediction accuracy estimates achieved in the validation stage was SNP missingness. Due to genotyping differences, different sets of SNPs were available for CpG methylation prediction model building in the ARIES and the Understanding Society data sets. Any SNPs present in a CpG methylation prediction model but not in the validation genotype data are not used for prediction, which could result in reduced prediction accuracy. To investigate this, the proportion of each prediction model's SNPs that were missing from the validation genotype data was calculated. Prediction models were then binned according to this missingness, with prediction models where the greatest proportion of SNPs were missing from the genotype data of the validation set placed into bin 10, and those with the least missingness placed into bin 1. For each bin, the average difference between the prediction accuracy estimate obtained in Chapter 5.3 and the prediction accuracy estimate obtained in the validation step was then calculated.

When considering the prediction models trained using the ARIES data, a slightly greater difference between the prediction accuracy estimates obtained in Chapter 5.3 and those obtained in the validation step was observed in the bins corresponding to the higher validation SNP missingness (Figures 5.14A and 5.14B). Notably, when considering the prediction models trained using the Understanding Society data, the relationship between SNP missingness and difference between prediction accuracy estimates from Chapter 5.3 and the prediction model validation step was even stronger (Figures 5.15A and 5.15B), indicating that greater prediction accuracy estimates may have been obtained in the prediction model validation stage had more complete SNP genotype information been available.

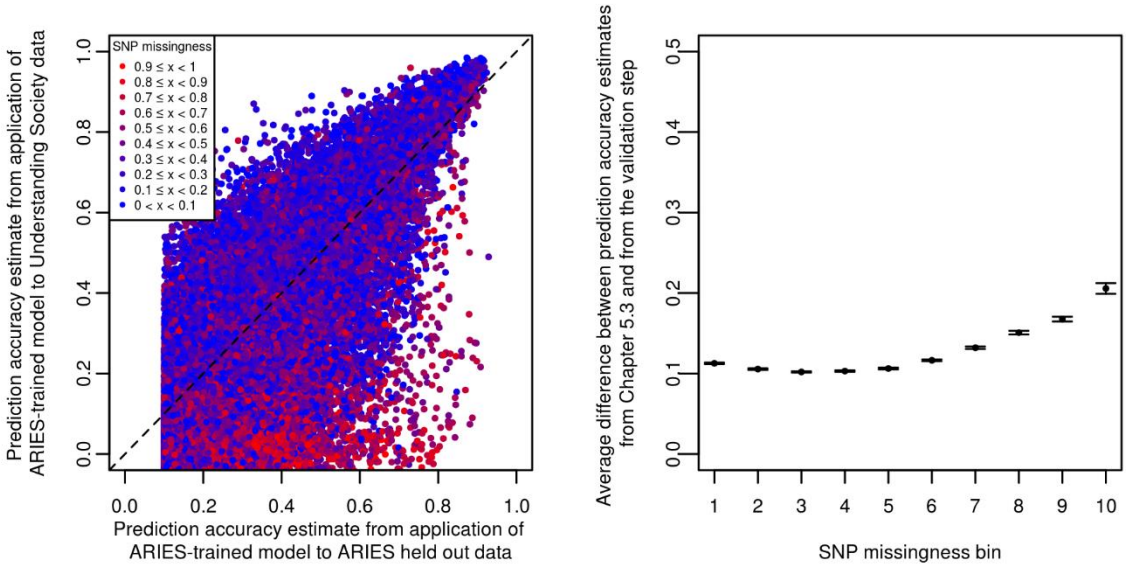

Figure 5.14. The effect of SNP missingness on validation of ARIES-trained models. (A) A comparison between prediction accuracy estimates from application prediction models trained using 70% of ARIES data to the remaining 30% of ARIES data (x axis) and from application of prediction models trained using 100% of ARIES data to 100% of the Understanding Society data (y axis). On each plot, each point represents a CpG site, and is coloured according to the percentage of SNPs required by the ARIES model to predict methylation that were missing in the Understanding Society data, where red represents a high degree of SNP missingness, and blue represents a low degree of missingness. The dashed line represents the line of equality, and the solid red line is a best fit line.

(B) CpG sites were grouped into 10 bins according to the percentage of SNPs required for prediction that were missing, with the CpG sites showing the most missingness placed in bin 10, and those with the least missingness in bin 1. The mean difference (+/- 1 standard error) between the prediction accuracy estimates obtained from application of models to ARIES held out data and to the Understanding Society data is shown for each bin on the y axis.

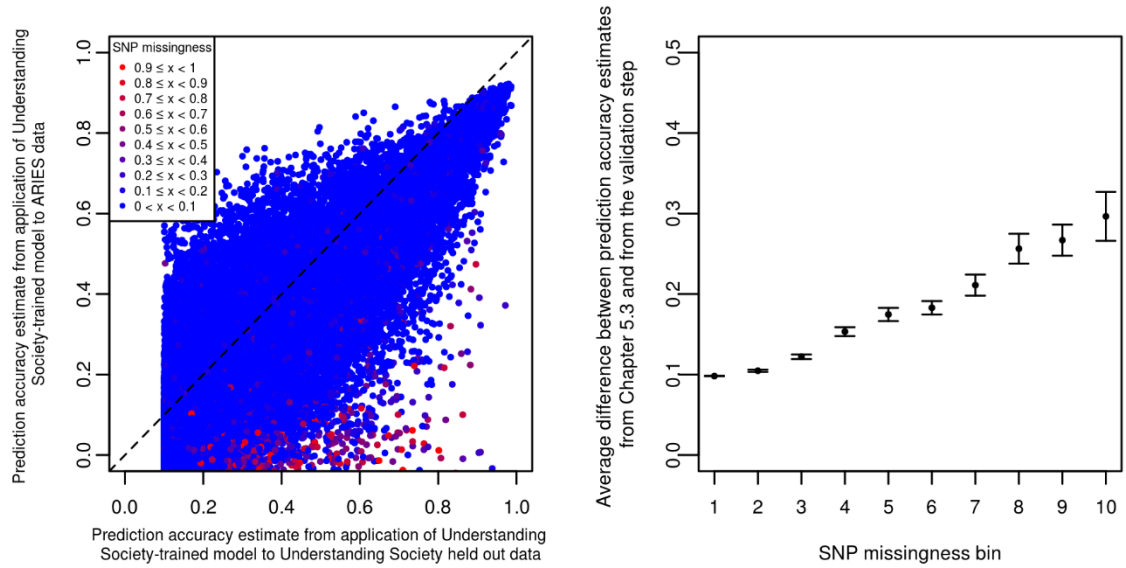

Figure 5.15. The effect of SNP missingness on validation of Understanding Society-trained models. (A) A comparison between prediction accuracy estimates from application prediction models

trained using 70% of Understanding Society data to the remaining 30% of Understanding Society data (x axis) and from application of prediction models trained using 100% of Understanding Society data to 100% of the ARIES data (y axis). On each plot, each point represents a CpG site, and is coloured according to the percentage of SNPs required by the Understanding Society model to predict methylation that were missing in the ARIES data, where red represents a high degree of SNP missingness, and blue represents a low degree of missingness. The dashed line represents the line of equality, and the solid red line is a best fit line.

(B) CpG sites were grouped into 10 bins according to the percentage of SNPs required for prediction that were missing, with the CpG sites showing the most missingness placed in bin 10, and those with the least missingness in bin 1. The mean difference (+/- 1standard error) between the prediction accuracy estimates obtained from application of models to the Understanding Society held out data and to the ARIES data is shown for each bin on the y axis.

5.6 Discussion

In this chapter, a comprehensive analysis of prediction of CpG methylation from SNP genotypes was conducted. Prediction models were first trained and tested using a range of penalised regression methods and range of window sizes. After identifying the optimal method and window size, CpG methylation prediction models were trained using data from the ARIES and Understanding Society projects, and prediction models that could predict CpG methylation with sufficient accuracy were identified to be taken forward for an MWAS analysis.

Comparison of the performance of three different penalised regression approaches showed that the approaches that assumed sparsity (LASSO and elastic net) outperformed the method that assumed polygenicity (ridge regression). This suggests that the underlying genetic architecture of CpG methylation at SNPs proximal to the CpG sites is sparse. While there has been no formal investigation of the genetic architecture of CpG methylation at proximal SNPs, mQTL studies have found that most CpG sites have few mQTLs (if any), each with a large effect size (Gaunt *et al.*, 2016), which is indicative of a sparse local architecture. Interestingly, this finding also matches the findings from the investigation of gene expression prediction in Chapter 4, and also from the literature (Wheeler *et al.*, 2016), indicating that this local sparsity may be a feature shared by multiple omics traits.

Following this, a comparison of prediction accuracy estimates when training CpG methylation prediction models using a range of different window sizes showed that while for many CpG sites changing the window size had only a small effect on

prediction accuracy, there was a group of CpG sites where there was a clear difference in the prediction accuracy estimates obtained at different window sizes. Interestingly, there was no consistent direction of effect to this, with some CpG sites benefitting from a smaller window size, and others benefitting from a larger window size, suggesting that the optimal window size for the prediction of CpG methylation is a CpG-specific quantity. Prior to this analysis, there had been no investigation into the effect of window size on the accuracy with which CpG methylation can be predicted from SNP genotypes. However, given that mQTL studies have shown that methylation at a small number of CpGs is regulated by SNPs distal to the CpG sites (Gaunt *et al.*, 2016), it is perhaps unsurprising that increasing the window size to the point where some of these more distal regulatory SNPs can be included in the prediction models could improve prediction accuracy for some CpG sites. In contrast to this, increasing the window size for the CpG sites where methylation is not known to be regulated by distal SNPs could lead to increased noise in the CpG methylation prediction model fitting procedure, leading to poorer estimation of the prediction model coefficients, and subsequently poorer prediction accuracy. Given that these distal regulatory SNPs are only known to exist for some, not all, CpG sites, this may explain why the average prediction accuracy across all CpG sites examined here fell slightly as the window size increased.

Having identified an optimal method and window size, CpG methylation prediction models were trained and tested using data from ARIES and Understanding Society. A crucial finding from this stage of the analysis is that methylation at most CpG sites cannot be accurately predicted using only proximal SNP genotypes. Despite this, there was a set of CpG sites at which the methylation could be predicted with a high degree of accuracy using only proximal SNPs. This suggests that the statistical power to detect association between predicted methylation and phenotype would be quite poor for most CpG sites, although there would be a subset of CpG sites at which good power could be achieved. However, the power to detect association in an MWAS would rely not only on the accuracy with which CpG methylation could be predicted from SNP genotypes, but also on the sample size of the GWAS data into which CpG methylation is being imputed. Thus, associations between predicted CpG methylation and phenotype may still be observed even at the poorly predicted CpG sites if a GWAS with a sufficient sample size were to be used. Interestingly, these findings also match those found when looking at the prediction accuracy of gene

expression in Chapter 4, suggesting that this may be a feature common to multiple omics traits.

Further investigation and enrichment testing revealed that the CpG sites where methylation could be well predicted from SNP genotypes were depleted at promoter sites and enriched at enhancer sites. Crucially these enrichments/depletions were consistently observed across the two data sets examined here. To date, no other study has looked for common features among CpG sites where methylation could be well predicted. However, it has been shown that CpG sites with strong mQTLs are depleted at promoter sites and enriched at enhancer sites (Gutierrez-Arcelus *et al.*, 2013; Banovich *et al.*, 2014). These CpG sites with strong mQTLs are likely to be the same as the CpG sites where methylation can be predicted well from SNP genotypes, reinforcing the findings shown here. This suggests that the prediction accuracy estimates observed for CpG sites in promoter regions may be low, meaning that the statistical power at the sites would also be low, and GWAS data with a large sample size may be required to detect associations at these sites. This is worth bearing in mind for future MWAS studies. It is worth noting that the chips used to measure methylation focus heavily on CpG sites at known promoters and enhancers, so this analysis is slightly biased. Further work using bisulfite sequencing data that could assay all CpG sites across the genome would be required to test whether these enrichments hold when using CpG sites all across the genome.

Following this, the heritability of CpG methylation was estimated using GCTA. Crucially, the estimates of CpG methylation heritability attributable to only the proximal SNPs were highly correlated with the estimates of prediction accuracy obtained from training and testing CpG methylation prediction models using the same proximal SNPs. This suggests that the prediction procedure was working nearly as well as could be expected, and that the poor average prediction accuracy was due to poor heritability, rather than other factors. However, the average prediction accuracy estimates were still marginally lower than the average heritability estimates, suggesting that there could still be a small improvement to the prediction accuracy estimates seen here before the true cap imposed by trait heritability was reached.

Prior to this analysis, there had been no estimate of the heritability of CpG methylation attributable to only proximal SNPs using data measured in blood. However, (Rowlatt *et al.*, 2016) estimated the heritability attributable to proximal SNPs using data from colorectal tissue, finding an average heritability of 6%, slightly

greater than the estimates of 3% and 4% found here. In addition to a different tissue being used, the slight difference between the estimates achieved here and in (Rowlatt *et al.*, 2016) may have been due to differences in the window sizes. In (Rowlatt *et al.*, 2016), a constant window size of 1Mb was used for all CpG sites, whereas the window size used for heritability estimation here was CpG-specific, and was often smaller than 1Mb. This could have resulted in fewer SNPs, and fewer genotypic effects being used for heritability estimation, resulting in lower estimates.

Despite the relatively poor estimates of CpG methylation heritability and prediction accuracy observed here, it is possible that higher estimates of both of these parameters are achievable. One reason for this is that it is likely that the true heritability is greater than the estimate reported here. This is because the GCTA software used to estimate heritability makes the assumption that all SNPs fed to the software have an effect on CpG methylation, meaning that the GCTA software is polygenic. However, the results from this Chapter suggest that the true underlying genetic architecture of CpG methylation at SNPs proximal to CpG sites is sparse. Thus, it is possible that using a heritability estimation model that is more appropriate for these data (i.e. one that makes assumptions of sparsity rather than polygenicity) could result in a slightly higher estimate of CpG methylation heritability, suggesting that the prediction accuracy of CpG methylation could be improved further. Another reason that higher prediction accuracy estimates may be achievable is that the average estimate of the heritability of CpG methylation found here is much lower than the estimate of 19% reported by (van Dongen *et al.*, 2016), who used SNPs across the entire genome to estimate the heritability, rather than just the SNPs most proximal to CpG sites. This implies that by including more distal SNPs in prediction models, the prediction accuracy achieved may increase.

However, this is in direct contrast to the results of Chapter 5.2, which showed that on average, increasing the window size and using more SNPs to fit prediction models actually led to a reduction in prediction accuracy. This contrast may have occurred because the window sizes used here were not large to incorporate all the distal regulatory SNPs in the prediction models. For example, some trans-mQTLs are known to regulate CpG sites located on a different chromosome to their location (Gaunt *et al.*, 2016), and so using only a window size of 3Mb as done here would miss those. However, increasing the window size also increases the computational burden when fitting prediction models, and increasing the window size to cover the

entire genome would likely resulting the prediction model fitting becoming computationally intractable. Instead, an approach by which the distal SNPs that most strongly influence methylation at a given CpG site could be identified and used to fit prediction models, rather than the “brute force” approach of using all SNPs within a given range of a CpG site as used here, may be useful. This would not only help reduce the computational issues associated with including many distal SNPs in the model fitting procedure, but would also remove the SNPs without an effect, whose inclusion could increase the statistical noise and reduce prediction accuracy. Further work to develop and test such an approach may prove beneficial in improving the prediction accuracy estimates seen here.

Finally, for the CpGs where methylation could be predicted with sufficient accuracy, prediction models were trained using a data set comprising 100% of samples. These prediction models were then validated through application to the other data set. Reassuringly, for the CpG sites where methylation could be predicted with a high degree of accuracy, the prediction accuracy estimates achieved in this validation stage were highly concordant with those achieved earlier. This suggests that for these CpG sites, a similarly high degree of prediction accuracy may be achieved when imputing methylation in GWAS samples, giving confidence that a high degree of statistical power will be available for these CpG sites in downstream analyses. Despite this, there were a number of CpG sites where the estimates observed from prediction model validation were not consistent with those seen earlier. This raises concerns about the reliability of the prediction accuracy estimates observed for those CpG sites, and thus raises doubts about the true power available for these CpG sites in downstream analyses. Furthermore, this shows the importance of having additional data with which to validate the prediction models.

Another key finding from the CpG methylation prediction model testing and validation stage of the analysis was that SNP missingness had an impact on the prediction accuracy estimates observed. This SNP missingness may have been caused by genotyping and imputation differences between the ARIES and Understanding Society data sets. Although both the ARIES and the Understanding Society data sets had genotype data that had been imputed to the 1000 Genomes phase 1 reference panel, there were differences in the genotyping procedures used for the two data sets. The ARIES data were genotyped on the Illumina human660W-quad array, while Understanding Society samples were genotyped on the sparser Illumina Infinium

HumanCoreExome chip. Thus, more SNPs were fed into the genotype imputation for ARIES than Understanding Society, which likely led to better imputation quality for ARIES than Understanding Society. This meant there were fewer poorly imputed variants to remove in post-imputation quality control with ARIES than Understanding Society. The effect of this was that for some of the CpG methylation prediction models trained using data from ARIES, many of the SNPs present in the prediction model were absent from the Understanding Society data, and so could not be used for prediction during the model validation step. By sorting CpG prediction models into bins according to this SNP missingness and comparing the prediction accuracy estimates from model testing and validation, it seemed that the prediction models which showed the largest difference also showed the most missingness, suggesting that missingness reduced prediction accuracy. This suggests that, were the genotyping and imputation procedure better for the Understanding Society data, meaning that SNP missingness was less of an issue, larger prediction accuracy estimates may have been observed, meaning more prediction models may have been taken forward for the MWAS analysis.

Ultimately, this issue would also affect prediction accuracy when applying CpG methylation prediction models to GWAS summary data in an MWAS analysis. If SNPs in the prediction model that were required for prediction were missing from the GWAS data, the analysis here suggests that prediction accuracy would be lost, which would ultimately result in a loss of power to detect association between predicted CpG methylation and phenotype at the CpG site is question. This problem could theoretically be avoided if the prediction model trainer knew which SNPs would be present in the GWAS data. If this were the case, then the CpG methylation prediction models could be trained using only the set of SNPs that were available in the GWAS data to which the prediction models would be applied, this ensuring that there would be no SNP missingness to impact on the analysis. However, given that the CpG methylation prediction models trained here will be applied to a range of publicly available GWAS summary data, each of which comes from a different source and has undergone different genotyping and imputation procedures, this step could not be taken here.

To conclude, some weaknesses of the analysis will be discussed. Some of the weaknesses relate to the methods of analysis chosen in this chapter. One such weakness was not using elastic net (with alpha tuned by cross validation). When

examining the how accurately gene expression could be predicted using only SNP genotypes in Chapter 4, elastic net (with alpha tuned by cross-validation) outperformed all three of the penalised regression approaches tested in this chapter. I chose not to use this method in this chapter as the improvement in prediction accuracy for this method compared to elastic net (with alpha set to 0.5) was small, and tuning the alpha parameter by cross-validation would have been extremely time consuming. However, if the time were available, then tuning this parameter would likely have led to an improvement in the prediction accuracy estimates observed here. This may be worth exploring in future work if time allows.

Another weakness was not using nested cross-validation to evaluate prediction accuracy. Like the decision not to use elastic net (with alpha tuned by cross-validation), the decision not to use nested cross-validation was also taken due to time constraints. By splitting data into testing and training sets, only some of the samples were used for model testing. In future, if time allowed, using nested cross-validation would allow all the samples to be used for model testing, which would give more reliable prediction accuracy estimates that might be considered more trustworthy.

Other drawbacks of the analysis here relate to the data used to conduct the analysis. One of these weaknesses is that the chips used to measure the methylation data here only examine a small proportion of the CpG sites genome wide. For example, the 450K chip, used to measure CpG methylation for the ARIES data, only examined ~2% of CpG sites, while the EPIC chip, used for Understanding Society data, only examines ~4% of CpG sites. This means that there are potentially many more CpG sites that can be predicted well from SNP genotypes that were not able to be examined in this analysis here. Ultimately, to allow for these CpG sites to be modelled well, a data set that has matched genotype and bisulfite sequencing CpG methylation data would be required. Furthermore, the methods used to measure the CpG methylation data mean that you can’t distinguish between methylation and hydroxymethylation. This means that we can’t tell if we’re predicting methylation or hydroxymethylation. Again, bisulfite sequencing data would be required to do this.

In conclusion, this chapter has shown that methylation at most CpG sites cannot be predicted accurately from the genotypes of SNPs proximal to the CpG sites. However, there exists a set of CpG sites at which this methylation can be predicted with a high degree of accuracy, and, reassuringly, these well-predicted CpG sites are consistently well-predicted with multiple different data sets. These well-predicted CpG

sites therefore have the potential to be useful for conducting MWAS using existing GWAS data sets or GWAS summary data.

Chapter 6. Methylome-wide association study elucidates relationships between CpG methylation, gene expression and complex traits

In Chapter 5, a set of CpG methylation prediction models were trained for the CpG sites where methylation could be predicted from SNP genotypes with sufficient accuracy. In this Chapter, I apply those prediction models to GWAS summary data for 30 traits in a methylome-wide association study (MWAS) to learn more about the relationship between CpG methylation and complex traits.

6.1 MWAS of 30 complex traits

First, an MWAS analysis was conducted by applying the set of 232,356 CpG methylation prediction models trained in Chapter 5 to publicly available summary data from GWAS of 30 complex traits, using the MetaXcan software package. In total, 48,382 associations between predicted CpG methylation and complex traits passed the stringent Bonferroni-corrected significance threshold (correcting for 232,356 models times 30 traits) of $p<7.17x10^{-9}$ (Figure 6.1). Significant associations were detected at 22,355 unique CpG sites, and associations were detected for all 30 complex traits. Interestingly, the associations tended to cluster near each other, forming peaks on the Manhattan plots similar to those that would be seen in a GWAS.

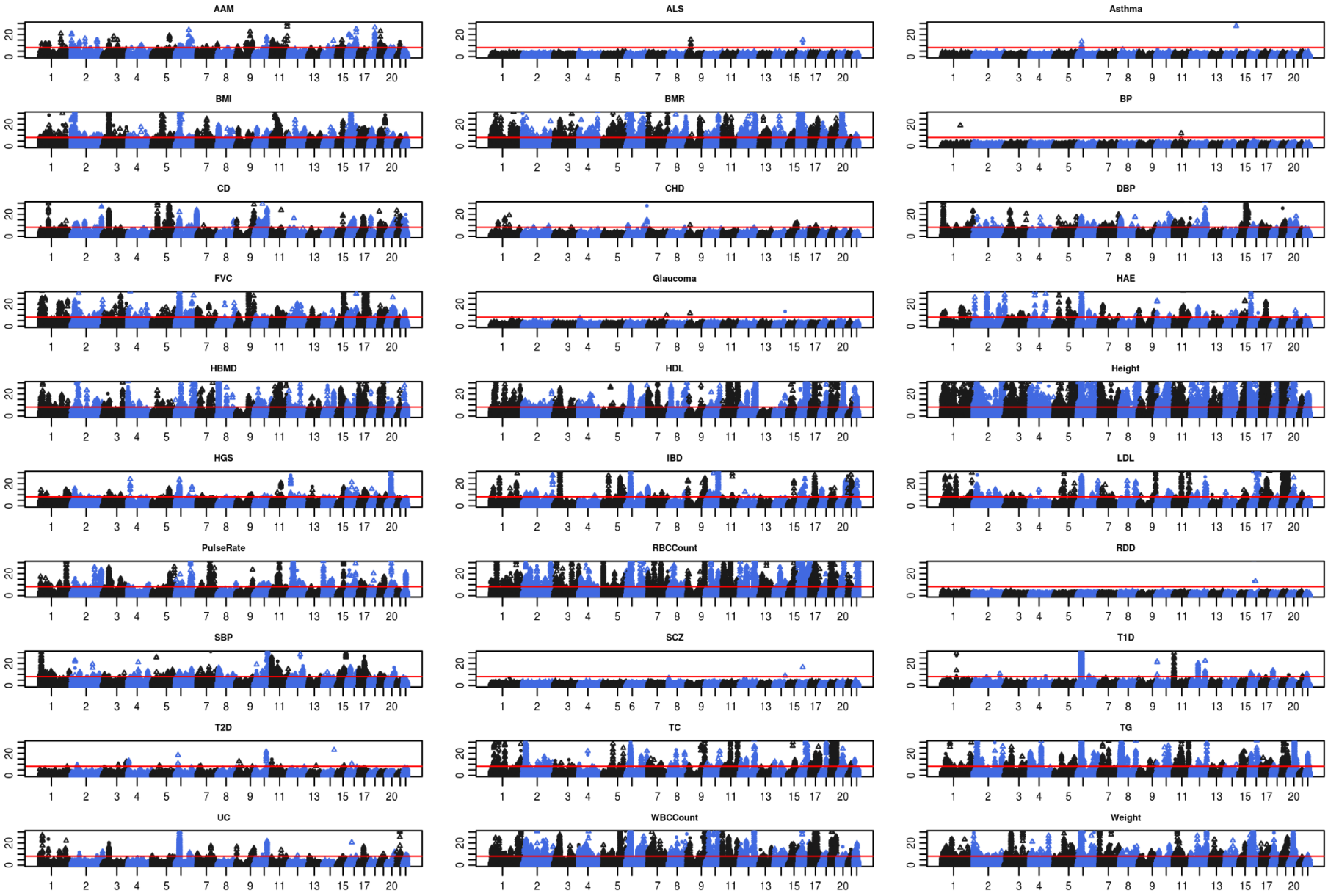
AAM
ALS
Asthma
BMI
BMR
BP
CD
CHD
DBP
FVC
Glaucoma
HAE
HBMD
HDL
Height
HGS
IBD
LDL
PulseRate
RBCCount
RDD
SBP
SCZ
T1D
T2D
TC
TG
UC
WBCCount
Weight
0
20
1 2 3 4 5 7 8 9 11 13 15 17 20
1 2 3 4 5 6 7 8 9 11 13 15 17 20

Figure 6.1. Manhattan plots of results for MWAS on 30 complex traits. Each point represents a CpG site, with its genomic location (as given in the manifest file) shown on the x axis and its p value from MWAS shown on the y axis. Results from ARIES-trained prediction models are shown as circles, and results from Understanding Society-trained prediction models are shown as triangles. Red lines indicate the Bonferroni corrected significance threshold of $p<7.17x10^{-9}$.

Overall, most of the tests conducted in the MWAS were well-informed, with an average of 94.7% of each CpG methylation prediction model's SNPs being present in the set of GWAS summary data to which the model was applied.

For the CpG sites tested using both ARIES-trained and Understanding Society-trained prediction models, the MWAS z scores from the two sets of models were highly correlated (r = 0.902) and concordant (Figure 6.2), indicating a high degree of consistency across the data sets.

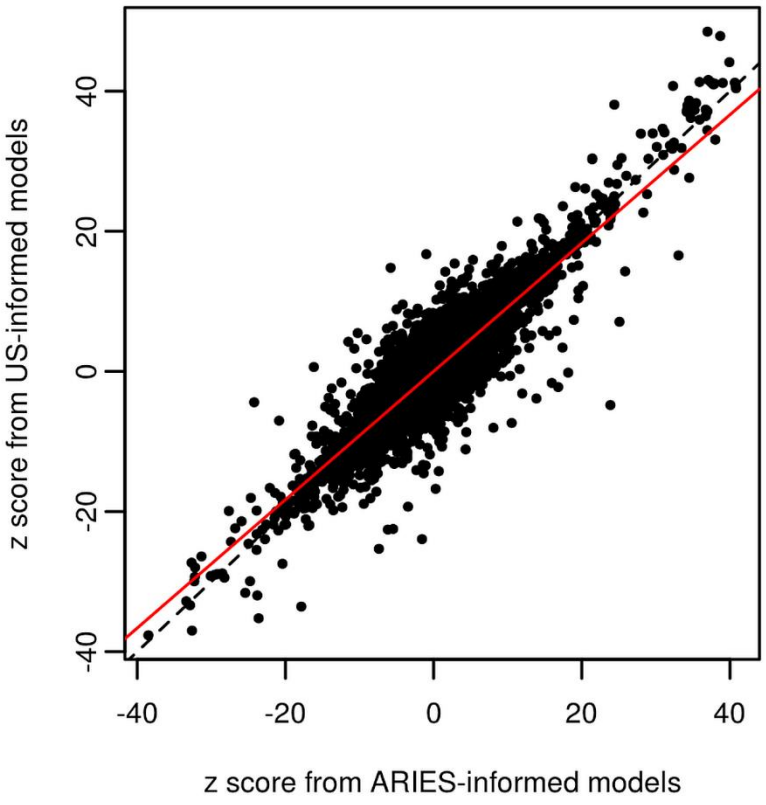

Figure 6.2. Comparison of MWAS results from ARIES-trained models and Understanding Society-trained models. Each point represents an association test between predicted methylation at a CpG site and one of the 30 complex traits. Shown are the z scores from this test when conducted using the ARIES-trained model (x axis) and the Understanding Society-trained model (y axis). The dashed line is the line of equality, and the red line is a best fit line.

According to annotations from the Illumina manifest files, CpG sites where predicted methylation was significantly associated with a complex trait were tagged to effector genes previously suggested in GWAS, including cg14349538 (associated with ulcerative colitis and tagged to *IRF5*) and cg15657055 (associated with BMI and tagged to *ADCY3*). When compared to the background set of the CpG sites that were tested in the MWAS, trait-associated CpGs where the prediction models were trained using data from ARIES were significantly enriched at promoter regions (OR = 1.18, p = 4.08×10^{-6}), CpG islands (OR = 1.36, p = 1.07×10^{-25}) and genes (OR = 1.38, p = 1.07×10^{-24}), implying that trait-associated CpG sites may act on their associated complex traits by regulating the expression of genes. Similar enrichments were observed when considering trait-associated CpG sites where the prediction model was trained using data from Understanding Society (Figure 6.3).

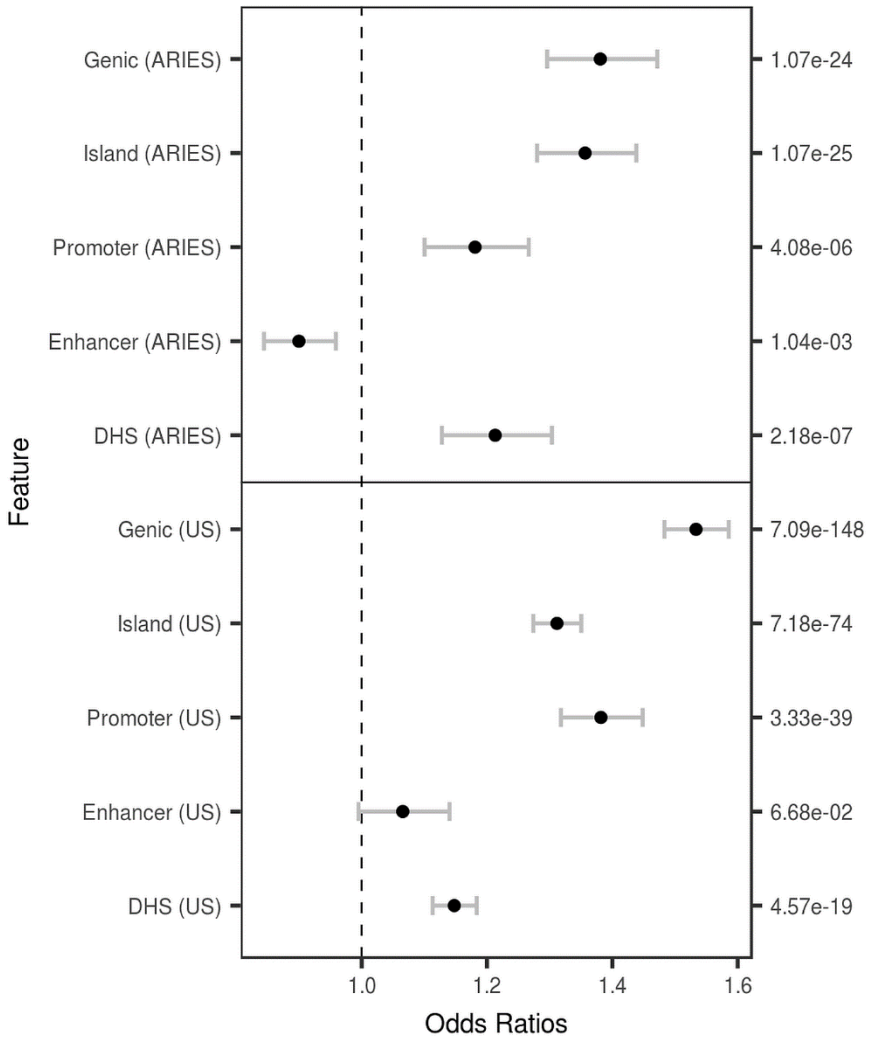

Figure 6.3. Enrichment of CpG annotations among trait-associated CpG sites. CpG sites were annotated using manifest files downloaded from the Illumina website. For each annotation, enrichment of CpG sites significantly associated with traits against the background of all CpGs tested in the MWAS on either the 450k or EPIC chip was tested using a two-sided Fisher's exact test. Odds ratios with 95% confidence intervals are shown on the x axis, with p values shown on the right.

During the course of this research being carried out, the MetaMeth software package was published (Freytag *et al.*, 2018). This package contains a set of CpG methylation prediction models which can be applied to GWAS summary data using a similar approach to MetaXcan. As a sanity check, the results obtained from the MWAS here were compared with those obtained from application of the MetaMeth package to the same 30 sets of GWAS summary data used in the MWAS. Overall, the z scores from MetaMeth were highly correlated and concordant with those from the MWAS when using prediction models trained using ARIES data (Figure 6.4) and when using prediction models trained using Understanding Society data (Figure 6.5), indicating broad agreement and providing further support to the associations detected here.

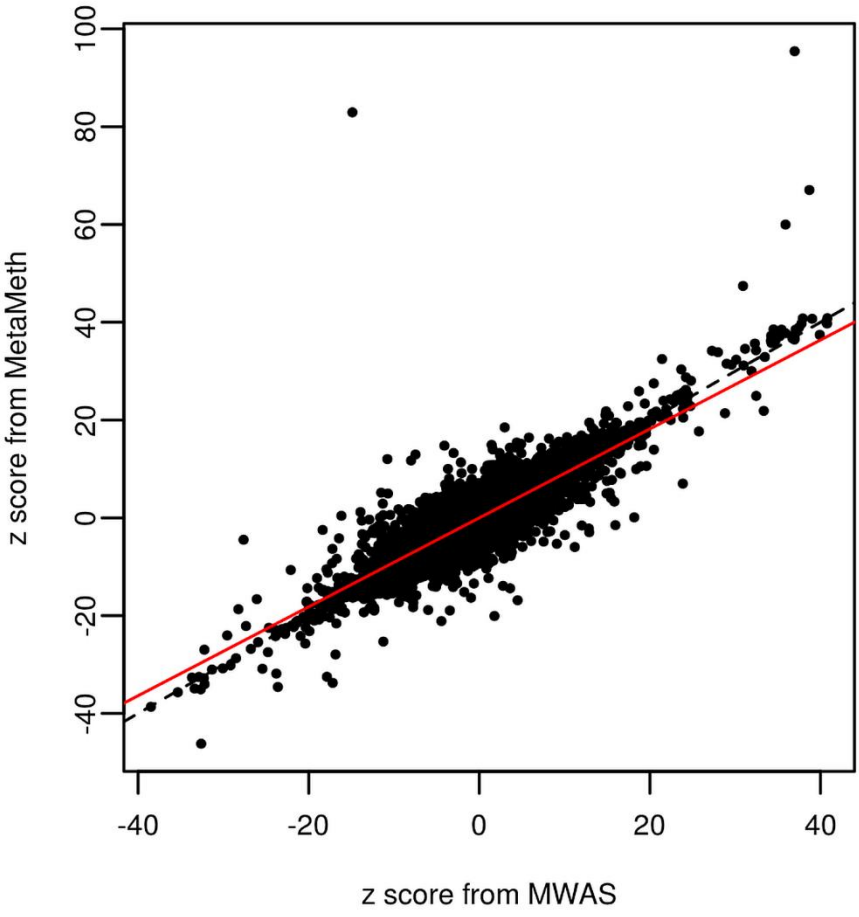

Figure 6.4. Comparison of MWAS results from application of ARIES-trained prediction models with MetaMeth results. Each point represents an association test between predicted methylation at a CpG site and one of the 30 complex traits. Shown are the z scores obtained from using the MWAS approach using prediction models trained using ARIES data (x axis), and the z score obtained from using the MetaMeth approach (y axis). The red line is a best fit line, and the dashed line is the line of equality.

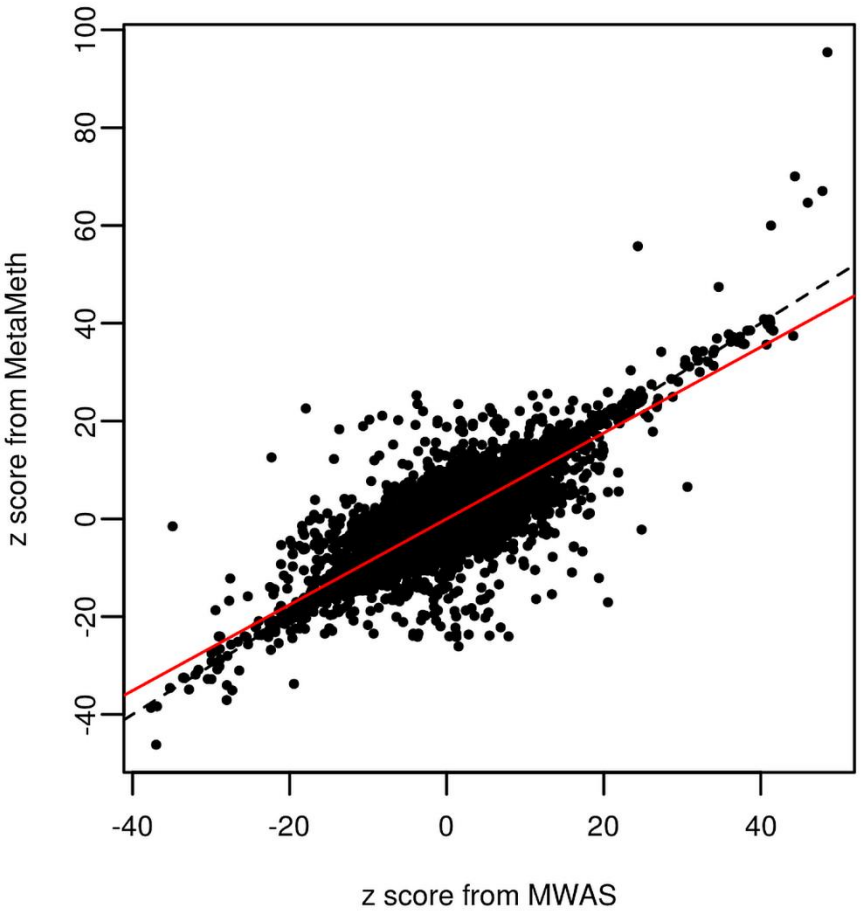

Figure 6.5. Comparison of MWAS results from application of Understanding Society-trained prediction models with MetaMeth results. Each point represents an association test between predicted methylation at a CpG site and one of the 30 complex traits. Shown are the z scores obtained from using the MWAS approach using prediction models trained using Understanding Society data (x axis), and the z score obtained from using the MetaMeth approach (y axis). The red line is a best fit line, and the dashed line is the line of equality.

6.2 TWAS of 30 complex traits

It is thought that CpG methylation acts on complex traits by affecting the expression of genes. To identify genes that may act on the 30 complex traits, a TWAS was conducted. Gene expression prediction models trained using whole blood expression data from version 7 of GTEx were downloaded from predictdb.org and applied to the 30 sets of GWAS summary data using the MetaXcan software. In total, 2,502 associations between predicted gene expression and complex traits passed the Bonferroni corrected significance threshold (correcting for 6,297 models times 30 traits) of $p<2.65x10^{-7}$ (Figure 6.6).

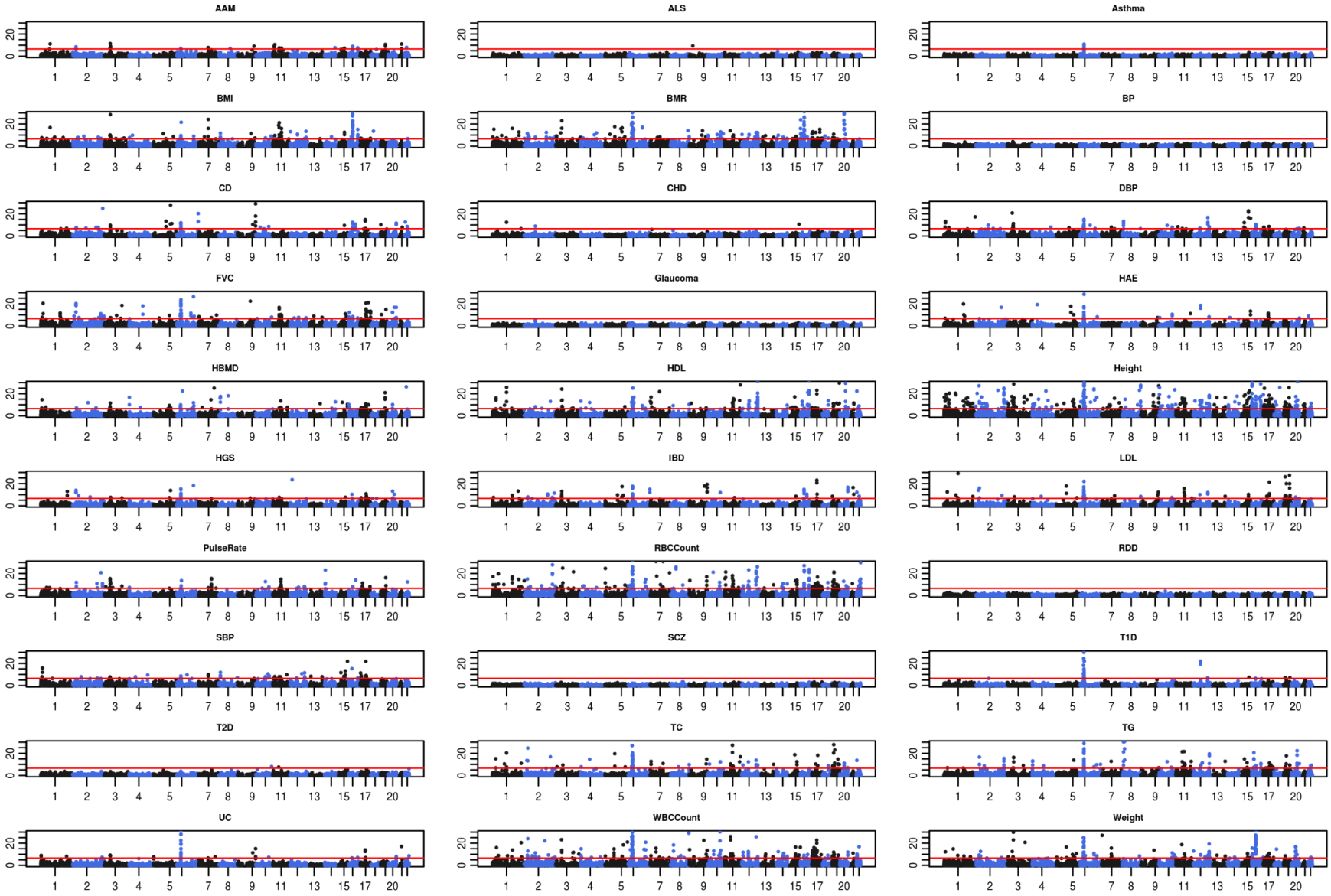
AAM
ALS
Asthma
BMI
BMR
BP
CD
CHD
DBP
FVC
Glaucoma
HAE
HBMD
HDL
Height
HGS
IBD
LDL
PulseRate
RBCCount
RDD
SBP
SCZ
T1D
T2D
TC
TG
UC
WBCCount
Weight
0
20
1 2 3 4 5 7 8 9 11 13 15 17 20

Figure 6.6. Manhattan plots of results for TWAS on 30 complex traits. Each point represents a gene site, with its genomic location shown on the x axis and its p value from TWAS shown on the y axis. Red lines indicate the Bonferroni corrected significance threshold of $p<2.65x10^{-7}$.

Similarly to the MWAS, the TWAS tests were typically well-informed, with an average of 97.2% of each gene expression prediction model's SNPs being present in the GWAS summary data to which the model was applied.

Among the significantly associated genes were some that had been previously suggested to have a role in their phenotype, such as *ADCY3* in BMI, and *CARD9* in IBD, providing confidence in the results found here. Crucially, the significant TWAS hits were often located in the same genomic loci as the significant results from the MWAS analysis, suggesting there may be a relationship between CpG methylation, gene expression and the complex traits.

6.3 Association testing between CpG methylation and gene expression

To investigate whether methylation at the trait-associated CpG sites was also associated with the expression of nearby genes, an MWAS of gene expression was conducted. CpG methylation prediction models for the 22,355 trait-associated CpG sites identified in the MWAS in Chapter 6.1 were applied to summary statistics from cis-eQTL analysis of whole blood gene expression data from the GTEx project (downloaded from the GTEx portal) using the MetaXcan software package.

In total, 10,908 significant associations between predicted CpG methylation and gene expression were identified at a Bonferroni-corrected significance threshold of $p<4.23x10^{-8}$. Significant associations were detected for 6,207 unique CpG sites and for 1,096 unique genes. Reassuringly, the z scores obtained from application of the prediction models trained using ARIES data to the GTEx eQTL data were highly correlated ($r = 0.87$) with those obtained from application of the prediction models trained using the Understanding Society data to the eQTL data.

Unlike the MWAS and TWAS conducted earlier, this MWAS of gene expression was typically less well-informed, with only an average of 54.2% of each CpG methylation prediction model's SNPs being present in the eQTL data.

When looking at the significant associations, the gene annotated to these CpG sites was often not the same as the gene with which the CpG site was associated. Of the 5,146 CpG sites significantly associated with the expression of a gene that were also tagged to a gene in the CpG methylation manifest file, only 1,796 were associated with the expression of the tagged gene, while the other CpG sites were associated with the expression of a different gene. This indicates that CpG methylation may not always act on the expression of the gene that it is suggested to in the manifest file.

When looking at the significant associations, CpG sites were typically located in close proximity to the genes to which they were associated (Figure 6.7), with an average distance between the CpG site and associated gene TSS of 216.5 Kbs. However, a small number of distal relationships were also detected, such as cg22508172, associated with expression of the gene *CD52*, which has a TSS over 2.5 Mbs away from the CpG site.

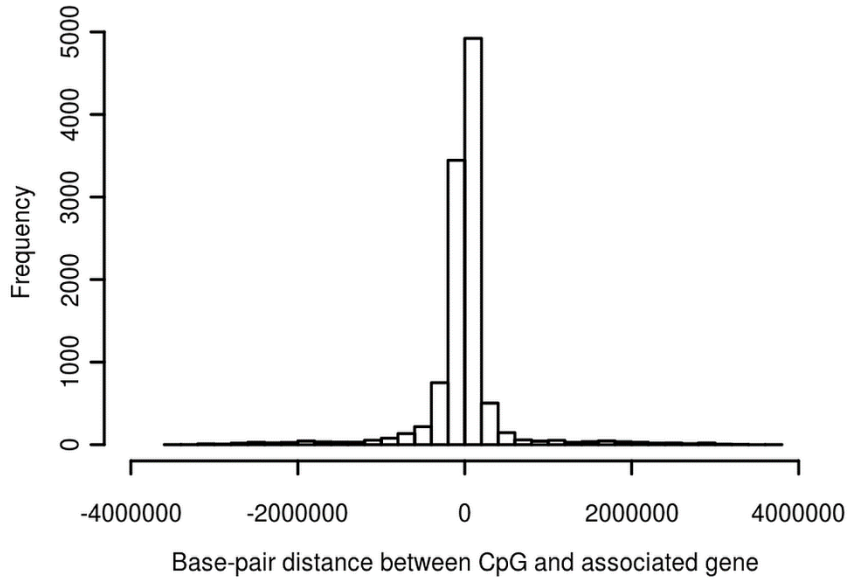

Figure 6.7. Histogram of distances between CpG sites and genes whose expression they are associated with.

6.4 Multi-trait colocalisation analysis of CpG methylation, gene expression and complex traits

Up to this point, associations have been identified between CpG methylation and complex traits (using MWAS), between CpG methylation and gene expression (using

the MWAS of GTEx gene expression data) and between gene expression and complex traits (using TWAS). The significant associations from these three analyses were then combined to generate CpG-gene-trait trios, where:

- Methylation at a CpG site was significantly associated with a complex trait
- Methylation at the same CpG site was significantly associated with the expression of a gene
- Expression of the same gene was significantly associated with the same complex trait as its associated CpG site

In total, 18,641 such trios were identified.

While the MWAS and TWAS approaches enabled identification of these CpG-gene-trait trios, they did not prove that there was a causal relationship between the CpG site, the gene and the phenotype. Indeed, the MWAS and TWAS approaches can detect associations in the presence of three different genetic models: causality (where a SNP causes a change in expression/methylation, which causes a change in the trait), pleiotropy (where a SNP affects expression/methylation and the trait through independent pathways) or linkage disequilibrium, where two SNPs that each affect only one trait are in strong linkage disequilibrium with each other. The CpG-gene-trait trios of most interest are those that are the result of causality between CpG methylation, gene expression and the trait, with those resulting from pleiotropy and linkage disequilibrium being of less interest. To identify (and remove from further consideration) the trios resulting from linkage disequilibrium, multi-trait colocalisation (moloc) analysis was applied. moloc uses summary statistics from association analyses between SNPs and traits to estimate the posterior probability of traits sharing the same causal SNP. For each of the 18,641 trios, moloc was applied using summary statistics from GWAS, GTEx whole blood eQTL data and mQTL data calculated using either ARIES or Understanding Society data (depending on which set of prediction models detected the associations in MWAS).

Overall, most trios showed strong support for a model in which all three traits did not colocalise to the same causal SNP (Table 6.1). 6,161 trios (33.1%) showed strong support (posterior probability ≥ 0.8) for the a.b.c hypothesis, indicating that all three traits were highly likely to have different causal SNPs. Furthermore, 5,790 trios (31.1%) showed strong support for one of the hypotheses in which two traits colocalised to the same causal SNP, but the third trait did not. For these trios, it is

likely that some (or all) of the associations between CpG methylation, gene expression and complex traits detected through MWAS and TWAS were induced by linkage disequilibrium, not pleiotropy or causality. In contrast, 663 trios (3.6%) showed strong support for the colocalisation of CpG methylation, gene expression and the complex trait to the same causal SNP, consistent with either a pleiotropic or causal relationship between CpG methylation, gene expression and the complex trait. Among these 663 trios were some that included genes with clear relevance to their phenotype, including *CARD9*, which colocalised with methylation at a number of CpG sites and with the CD, IBD and UC phenotypes, and *UBE2L3*, which colocalised with methylation at a number of CpG sites and with autoimmune phenotypes including CD, HAE and IBD.

Hypothesis	**Number of trios showing strong support for this hypothesis**
zero	0
a	0
b	0
c	0
ab	6
ac	0
bc	0
a.b	70
a.c	0
b.c	0
ab.c	620
ac.b	1014
a.bc	4156
a.b.c	6161
abc	663

Table 6.1. Results from moloc analysis of CpG methylation, gene expression and complex traits for 18,641 CpG-gene-trait trios. The 15 hypotheses considered were zero (genotypes at SNPs in the genomic region analysed are not associated with CpG methylation, gene expression or the complex trait), a (genotypes at SNPs in the analysis region are only associated with CpG methylation), b (genotypes at SNPs in the analysis region are only associated with gene expression), c (genotypes at SNPs in the analysis region are only associated with the complex trait), ab (CpG methylation and gene expression colocalise to the same causal SNP in the analysis region), ac (CpG methylation and the complex trait colocalise to the same causal SNP in the analysis region), bc (gene expression and the complex trait colocalise to the same causal SNP in the analysis region), a.b (CpG methylation and gene expression are both associated with SNPs in the analysis region, but do not colocalise to the same causal SNP), a.c (CpG methylation and the complex trait are both associated with SNPs in the analysis region, but do not colocalise to the same causal SNP) , b.c (gene expression and the complex trait are both associated with SNPs in the analysis region, but do not colocalise to the same

causal SNP), ab.c (CpG methylation and gene expression colocalise to the same causal SNP in the analysis region, but the complex trait has a different causal SNP in the analysis region), ac.b (CpG methylation and the complex trait colocalise to the same causal SNP in the analysis region, but gene expression has a different causal SNP in the analysis region), a.bc (gene expression and the complex trait colocalise to the same causal SNP in the analysis region, but CpG methylation has a different causal SNP in the analysis region), a.b.c (CpG methylation, gene expression and the complex trait all have different causal SNPs in the analysis region) and abc (CpG methylation, gene expression and the complex trait all share the same causal SNP in the analysis region).

6.5 Two-step Mendelian Randomisation analysis

Having identified trios where CpG methylation, gene expression and phenotype all colocalised to the same SNP, two-step Mendelian Randomisation was next applied to examine a potential causal effect. Independent, valid instruments for both steps of the two-step MR were identified for 342 of the 663 trios that passed moloc analysis.

Of these 342 trios, a Bonferroni-corrected significant ($p<7.31x10^{-5}$) potential causal effect of CpG methylation on gene expression and a subsequent Bonferroni-corrected significant ($p<7.31x10^{-5}$) potential causal effect of gene expression on the complex trait of interest were identified in 102 of the trios (Appendix A). Among the significant findings were *TUFM*, which is known to have a role in the regulation of energy production in mitochondria, which is thought to play a role in both BMI and weight.

6.6 Discussion

In this chapter, an MWAS of 30 complex traits was conducted using CpG methylation prediction models trained using data from ARIES and Understanding Society. This approach identified thousands of significant associations between CpG methylation and complex traits, indicating that it is a powerful approach, and that further application to more GWAS summary data may help to improve understanding of more biological mechanisms underlying GWAS hits. Interestingly, most of the significant associations tended to cluster near each other, often appearing as towers on the Manhattan plots, similar to those that would be found on a Manhattan plot of GWAS results. This likely occurred because the predicted methylation states of nearby CpG sites were highly correlated with each other. This could make it harder to identify the truly causal CpG site(s) underlying the MWAS associations. However, recently developed methods of fine-mapping analysis for TWAS (Mancuso *et al.*,

2019; Wu and Pan, 2020) could theoretically be adapted for use on MWAS results, and could help identify which trait-associated CpG sites are most likely causal, helping to solve this problem. Further work will be required to adapt this fine-mapping approach and to apply it to MWAS results.

As a sanity check, the results from the MWAS were compared with results from similar experiments. First, the results of the MWAS conducted using prediction models trained with ARIES data were compared with those obtained from MWAS using prediction models trained with Understanding Society data, and were found to be highly concordant. This is reassuring, and gives confidence that the results detected here are correct. In addition, the MWAS results here were compared with those obtained from application of the MetaMeth (Freytag *et al.*, 2018) software to the same 30 sets of GWAS summary data. Again, the results were highly concordant, giving further confidence in the results detected here.

Following this, a TWAS was conducted by applying gene expression prediction models trained using GTEx whole blood expression data to the same 30 sets of GWAS summary data used in the MWAS. This approach returned hits at the expected genes, such as *CARD9* in the IBD-related traits and *ADCY3* in BMI. Notably, many of the associations detected in the TWAS here were also detected in the PhenomeXcan TWAS analysis (Pividori *et al.*, 2019; Barbeira *et al.*, 2020), corroborating the results found here.

Following this, the MWAS approach was used again to investigate the relationship between CpG methylation and gene expression through application of CpG methylation prediction models to whole blood eQTL summary statistics from GTEx. Overall, this found that many CpG sites associated with complex traits were also associated with the expression of genes. This is potentially more informative than relying on the gene annotations provided in the manifest files, which are based on proximity alone, and allowed the identification of potential effector genes for CpGs without gene-level annotation in the manifest files, indicating that this may be a useful way of annotating CpG sites in future MWAS. Most methylation-expression associations identified here were for CpG sites and genes in close proximity, reinforcing previous findings (Bonder *et al.*, 2017), although some relationships were identified between more distal CpGs and genes. While CpG methylation is typically thought to affect the expression of proximal genes, there is some evidence that CpG methylation at enhancer regions can affect the expression of more distal genes

(Fleischer *et al.*, 2017). This may explain some of the distal associations between CpG methylation and gene expression observed here. However, a study in healthy skeletal muscle tissue found no associations between measured CpG methylation and measured gene expression where the CpG site was located more than 1 Mb away from the gene TSS after correcting for standard covariates (Taylor *et al.*, 2019), contradicting the results found here. It is possible that the associations found here between distal CpG sites and genes may not be true causal relationships between CpG methylation and gene expression, but may have been induced by linkage disequilibrium or pleiotropy. Further work will be required to investigate this possibility.

While the MWAS of gene expression enabled the detection of many relationships between CpG methylation and gene expression, it suffered from poorly-informed prediction models. This was likely due to the nature of data used to perform this analysis. This analysis was conducted using whole blood eQTL data from GTEX. Unfortunately, these data only contained summary statistics from cis-eQTL tests (where the SNP and gene tested are in close proximity), and did not contain any summary data for tests where the SNP and gene were greater than 1 Mb apart. This meant that many SNPs in the CpG methylation prediction model were not present in the eQTL summary data unless the CpG site and gene were highly proximal, and the window size used to train the CpG methylation prediction model was 1 Mb or less. This resulted in many poorly-informed association tests. As shown in Chapter 5, when a CpG methylation prediction model is more poorly-informed, it tends to predict CpG methylation with less accuracy then when it is well-informed. In addition to this, it has been shown that in some contexts a poorly-informed prediction model has a greater chance of giving a spurious result than a perfectly-informed prediction model (Barbeira *et al.*, 2019). In fact, most (but not all) of the significant associations between distal CpG sites and genes were the result of poorly-informed prediction models, meaning that some of these more distal associations between CpG methylation and gene expression may have been spurious. Overall, caution should be taken when interpreting the results of the poorly-informed tests conducted here.

One way to avoid these problems related to poor overlap of SNPs in the prediction models and SNPs in the summary statistics would be to use eQTL summary statistics from tests of every SNP against every gene, genome-wide. However, due to the large multiple testing burden that would come with testing all genes against all SNPs,

most studies of trans-eQTLs often restrict analysis to a small number of SNPs. For example, the (Westra *et al.*, 2013) study performed trans-eQTL analysis only for those SNPs that had been shown to be associated with a phenotype through GWAS. As a result of this, there are no large scale, publicly available data sets with genome wide summary statistics for eQTL testing. Indeed, the cis-eQTL GTEx data used here were used as they were the only eQTL data freely available for download from the GTEx portal. Given this large multiple testing burden, a solution to this problem involving genome-wide eQTL summary data is unlikely to be developed any time soon.

Another potential way to avoid these problems would be to impute the eQTL summary data for those SNPs in the CpG methylation prediction models but not present in the eQTL summary data, for example using the ImpG-summary algorithm (Pasaniuc *et al.*, 2014). This summary statistics imputation is an approach used to deal with missing data by the FUSION package for TWAS (Gusev *et al.*, 2016), and has been suggested as a potential addition to the PrediXcan software package in future updates. However, given that the GTEx eQTL summary data contain only a relatively small number of SNPs for each gene, only a small number of SNPs would be fed into the summary statistics imputation procedure. Thus, it is likely that much of the imputation would be inaccurate. Furthermore, this could also result in the imputation of SNPs that were removed from the data purposefully, for example due to low minor allele frequency, or due to failing GWAS data quality control thresholds. Thus, this method would not seem to be an appropriate solution to the problem either. It is clear that much further work will be required to develop a robust solution to the problem.

While the TWAS and MWAS approaches described above were powerful at identifying associations, they were all vulnerable to the detection of associations induced by LD between SNPs (where a SNP associated with gene expression or CpG methylation was in high LD with a separate SNP that was associated with the phenotype, resulting in a “false” TWAS/MWAS association). To identify instances where associations may have been induced by LD between SNPs, a colocalisation approach was taken, similar to the approach recommended in (Barbeira *et al.*, 2018). Here, as three traits were considered in each instance, a multi-trait colocalisation (Giambartolomei *et al.*, 2018) approach was used. Crucially, this analysis revealed that many of the trios identified through MWAS and TWAS were induced by LD, and

thus did not represent causal or pleiotropic relationships between CpG methylation, gene expression and complex traits. This highlights a key weakness of the MWAS (and TWAS) approach, and shows the importance of using a colocalisation (or similar) approach as a follow-up analysis, as without this type of analysis, many spurious findings would be taken forward. Despite this, the moloc approach did detect a number of findings that were not the result of linkage disequilibrium, but instead were due to either causality or pleiotropy. Among these findings were some interesting genes, including *UBE2L3*, a ubiquitin conjugating enzyme that is known to be important for the activation of NF-kB, and its subsequent role in downstream immune processes (Lewis *et al.*, 2015). Additionally, *UBE2L3* has been previously reported as a potential risk gene for a number of autoimmune conditions including CD (Fransen *et al.*, 2010) and SLE (Harley *et al.*, 2008), providing further evidence of its relevance to autoimmune conditions. This shows that with proper follow-up analysis, the TWAS and MWAS approaches can be powerful tools for the investigation of GWAS findings and can identify potentially disease-relevant CpG sites and genes.

Finally, to test for a potential causal relationship between CpG methylation, gene expression and complex traits for those CpG sites, genes and complex traits where colocalisation was found to occur, a two-step Mendelian Randomisation approach was used. This identified a number of potentially causal relationships between CpG methylation, gene expression and complex traits. Among these were significant potentially causal effects of methylation at cg04414917, cg01499465 and cg05684748 on *TUFM* expression, and of *TUFM* expression on BMI. *TUFM* is a translation elongation factor found in the mitochondria, and is thought to be crucial in oxidative phosphorylation, suggesting that it plays a role in the regulation of energy metabolism, and indicating that it may be of interest in both BMI and weight.

Interestingly, some of the other genes at which a potentially causal effect was detected were HLA genes, with effects found on a number of autoimmune conditions. While the HLA genes are known to be important in these autoimmune conditions, it is thought that coding variation that alters the amino acids of a protein, rather than non-coding variation that regulates gene expression, is responsible for genetic associations between HLA genes and autoimmune conditions (Darlay *et al.*, 2018). Thus, it is surprising to see potentially causal effects of HLA gene expression on the complex traits here. One potential reason for these surprising results is that there is

extremely high linkage disequilibrium between genetic variants located in the MHC. It is known that colocalisation methods are more likely to give spurious results under conditions of extreme linkage disequilibrium (Giambartolomei *et al.*, 2014), and so it is possible that the results for the HLA genes found here are spurious. Thus, caution should be taken when interpreting the results related to the HLA genes.

Many of the other genes found here to have a potentially causal effect on their phenotype are genes about which little is known, including *ZFP57* and *RNASET2*. These genes represent exciting new avenues for further exploration, and may provide further insight into complex trait biology.

Note that care is taken with the use of the word “causal”. This is because Mendelian Randomisation requires three assumptions to be met for true causality to be claimed. First, the instrument must be strongly associated with the exposure of interest. Second, the instrument must not be associated with the outcome via any mechanism other than the exposure (the no pleiotropy assumption). Third, there must be no unobserved confounders of the relationship between exposure and outcome. Here, the first assumption is met by only selecting instruments that reach a strong significance threshold in mQTL/eQTL analysis. However, neither the second nor the third assumption can be tested via a simple Mendelian Randomisation approach, and so it is not known whether they are met. For this reason, the findings are referred to as “potentially causal”.

However, in some of the more recent Mendelian Randomisation approaches, methods have been developed that either enable the testing of these assumptions, or that relax the assumptions slightly, providing more confidence in the results of Mendelian Randomisation analyses. For example, the MR-Egger (Bowden *et al.*, 2015) approach enables the user to identify whether the genetic instruments used in Mendelian randomisation show a consistent direction of pleiotropy and allows for the no-pleiotropy assumption to be relaxed slightly. Similarly, the MR-PRESSO approach (Verbanck *et al.*, 2018) also enables the identification of genetic instruments that violate the no pleiotropy assumption, and corrects for the presence of these problematic instruments. In addition to these adapted Mendelian Randomisation approaches, more recent causal inference approaches such as Bayesian Networks (Howey *et al.*, 2020), which are not subject to the same stringent assumptions as Mendelian Randomisation, may also be of use. The further use of two-step Mendelian Randomisation, as well as use of the more sophisticated causal inference

techniques as further sensitivity analyses, would help to further refine the set of potentially causal associations found here, and represents a clear opportunity for further study.

To conclude, a number of limitations of this study are discussed. Here, CpG methylation prediction models were trained using data measured in blood. However, blood is unlikely to be the true causal tissue of interest for some of the complex traits studied in here. For example, for IBD it has been suggested that immune cell subtypes are likely to be causal for the disease (Momozawa *et al.*, 2018), while for the depression phenotype, it has been suggested that brain tissues are the most likely to play a causal role in the disease (Wray *et al.*, 2018). While some studies have found high overall correlation between methylation states in different tissues (Hewitt *et al.*, 2017; Braun *et al.*, 2019), these have often been conducted using quite small sample sizes, and have typically only examined a very small number of tissues, mostly focussing on brain tissues.

Perhaps of more relevance to this project is the concordance of SNP effects on CpG methylation across different tissues. (Hannon *et al.*, 2017) compared the results of mQTL analysis using data from whole blood with the results of mQTL analysis using fetal brain data, finding that over 75% of relationships between SNP genotypes and CpG methylation identified using the whole blood CpG methylation data were also found when using the fetal brain methylation, with a consistent direction of effect seen across the two tissues. Similarly, (Shi *et al.*, 2014) reported strong concordance of mQTL effects across lung, breast and kidney tissues, while (Lin *et al.*, 2018) reported strong concordance of mQTL effects across brain, blood and saliva. Together, these studies suggest that similar results to those obtained in the MWAS here could be obtained if CpG methylation prediction models were trained using data from another tissue. However, there are a large number of tissues where the concordance between their mQTL effects and those from blood are unknown. For these tissues, it is difficult to say how well the CpG methylation prediction models or association results generated here translate across. To establish this, more matched genotype and methylation data from a wide range of tissues would be required.

Finally, it is likely that the significance thresholds used in the MWAS and TWAS analyses here were overly stringent, meaning that some associations were potentially missed. Here, in the MWAS of 30 complex traits, a strict Bonferroni-corrected significance threshold used to correct for 232,356 tests for each of 30

phenotypes, giving a threshold of $p<7.17x10^{-9}$. However, many of the phenotypes used here were highly correlated with each other, and the MWAS of the 30 traits were not truly independent from one another. For example, the IBD summary data used here were generated by combining data from the CD and the UC GWAS, meaning that the CD, IBD and UC phenotypes are all highly correlated with each other. Similarly, height, BMI and weight were all examined as separate phenotypes here, but are all known to be correlated with each other. Instead of stringently correcting for the 30 different phenotypes, an approach such as PhenoSpD (Zheng *et al.*, 2018) could have been used to identify the correlations between the phenotypes and to adjust the multiple testing correction thresholds used here, potentially allowing for detection of more associations.

Overall, this chapter demonstrates that MWAS is a powerful approach for the integration of CpG methylation and GWAS data and the identification of associations between the two. This MWAS approach still suffers from the same vulnerabilities as TWAS, including the detection of many relationships induced by linkage disequilibrium. However, further follow up analyses including TWAS, colocalisation and causal inference can allow the identification of potentially causal relationships between CpG methylation, gene expression and complex traits. Further application of MWAS to more phenotypes and follow up with more sophisticated causal inference analysis may help to elucidate the true causal biological mechanisms underlying significant GWAS findings.

Chapter 7. Post-GWAS analysis of PBC

Primary biliary cholangitis (PBC) is an autoimmune liver disease characterised by cholestasis, destruction of bile ducts and in many cases liver failure (Hirschfield *et al.*, 2018). A recent genome-wide meta-analysis of PBC conducted by Heather Cordell (manuscript in preparation) identified over 50 significant PBC risk loci across the genome. To investigate the role of gene expression, CpG methylation and protein levels in the hits found in this recent meta-analysis, TWAS, MWAS and a proteome-wide association study (PWAS) were conducted using summary statistics from this genome-wide meta-analysis.

7.1 TWAS of PBC using gene expression prediction models trained using GTEx data

To investigate the role of gene expression in PBC, a TWAS was conducted by applying gene expression prediction models to PBC GWAS summary data using the MetaXcan package. To maximise the number of genes examined in the TWAS, all 48 sets of gene expression prediction models, each trained using data from a different GTEx tissue, were downloaded from predictdb.org and applied to the PBC GWAS summary data.

In total, 1,758 associations between predicted gene expression and PBC were detected at a transcriptome-wide Bonferroni-corrected significance threshold of $p<2.15e^{-07}$ (Figure 7.1). These significant associations covered 347 unique genes, and significant associations were detected in each of the 48 tissues tested.

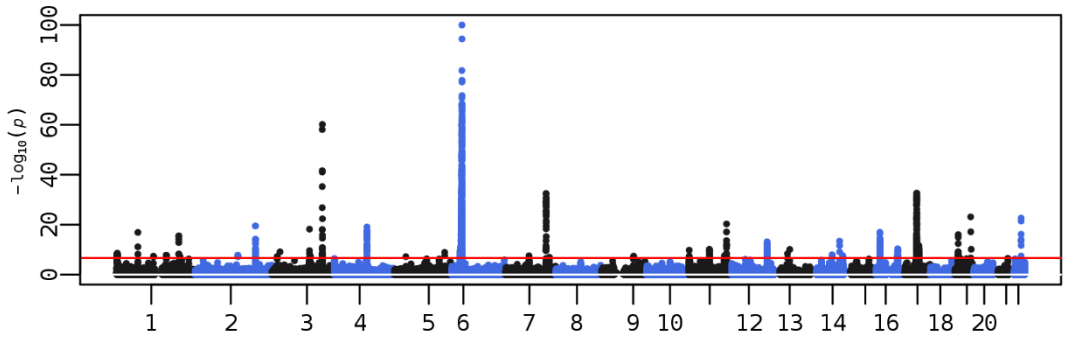

Figure 7.1. TWAS of PBC using gene expression prediction models from 48 GTEx tissues. Shown are the results from the TWAS for all 48 GTEx tissues (superimposed on top of one another). Each point represents a test of association between the predicted tissue expression of a gene, with the genomic location of the gene shown on the x axis and the p value from the TWAS shown on the y axis. The red line indicates the Bonferroni-corrected significance threshold at $p<2.15x10^{-7}$.

The significant TWAS associations mapped to 38 unique genomic loci. The locus containing the strongest associations was the MHC on chromosome 6, where 876 significant associations between predicted gene expression and PBC were detected, representing 49.8% of all associations detected in the TWAS. Multiple significant associations were identified at 31 of the 38 loci, while the TWAS prioritised only a single gene at seven of the 38 loci.

35 of the 38 detected genomic loci contained at least one SNP that reached a genome-wide significance threshold of $p<5x10^{-8}$ in the PBC meta-analysis. The remaining three loci (which contained a total of five significant associations between predicted gene expression and PBC) did not contain any genome-wide significant SNPs, so would not be found through traditional GWAS, indicating that using TWAS enabled these extra associations to be found.

Among the significant genes identified through TWAS were those involved in known PBC pathways, such as *IL12RB2* and *IL12A*, both part of the IL-12 signalling pathway. In addition to this, a number of good candidate genes that had not been previously suggested to have an effect on PBC were identified through TWAS. This included *FCRL3*, a gene known to play a role in the proliferation and activation of B cells, and *NDFIP1*, which is thought to have a role in regulating tolerance in peripheral T cells.

7.2 MWAS of PBC using CpG methylation prediction models trained using ARIES data

To investigate the role of CpG methylation in PBC, an MWAS was conducted by applying the set of CpG methylation prediction models trained using 100% of the ARIES data in Chapter 5 to the PBC meta-analysis summary data.

In total, 148 CpGs showed a significant association between predicted CpG methylation and PBC at a Bonferroni-corrected significance threshold of $p<1.03x10^{-6}$ (Figure 6.2). Significant associations were located at 27 unique genomic loci, of which 23 contained SNPs that reached genome-wide significance in the PBC meta-analysis, while 4 loci had no genome-wide significant SNPs and thus represented new "discoveries". These new discoveries included 4 CpGs on chromosome 19, annotated to *MAST3* (according to the Illumina manifest file), which is thought to have a role in immunity and has previously been implicated in IBD (Labbé *et al.*, 2008).

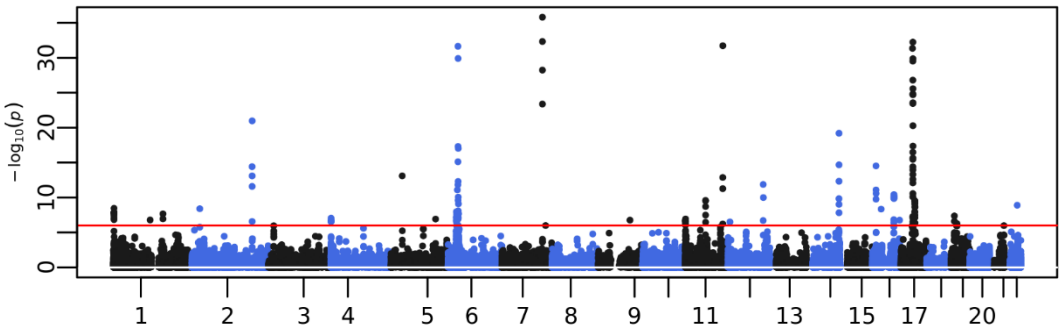

Figure 7.2. MWAS of PBC using CpG methylation prediction models trained using data from ARIES. Shown are the results from the MWAS. Each point represents a test of association between the predicted methylation of a CpG site, with the genomic location of the CpG site shown on the x axis and the p value from the MWAS shown on the y axis. The red line indicates the Bonferroni-corrected significance threshold at $p<1.03x10^{-6}$.

There was a clear overlap between the TWAS and MWAS results, with many genomic loci showing significant genes in the TWAS and also showing significant CpGs in the MWAS, suggesting a link between methylation and expression in PBC. Furthermore, some of the significant CpG sites were tagged to genes found to be significant in the TWAS and known to play a role in immunity, including cg04864179, tagged to *IRF5*, a transcription factor thought to regulate immune system activity.

7.3. Prediction of serum protein levels from SNP genotypes

Prior to conducting a PWAS, the ability of prediction models to predict protein levels from SNP genotypes was assessed using an approach similar to that used for gene expression in Chapter 4 and for methylation in Chapter 5. Using Olink proteomics data measured in the PBC cases (see Chapter 2), a 10-fold nested cross-validation experiment was performed using elastic net, with alpha = 0.5, lambda determined by cross-validation and a window size of 1Mb. SNPs not present in the PBC meta-analysis summary statistics were removed from the PBC cases data prior to analysis. As this data set had a lot of missing proteomics data, the 10 folds were generated on a protein-by-protein basis. For each of the 341 proteins, the samples with missing data were removed, and remaining samples were randomly split into 10 groups. In each fold of the cross-validation, 1 group was held out for model testing, with the other 9 groups combined and used for model training.

Overall, prediction accuracy for most proteins was near zero (Figure 7.3), with the mean prediction accuracy across all 341 proteins = 0.086, although there was a small set of proteins for which the levels could be predicted with a high degree of accuracy. Proteins whose levels were well predicted from SNPs include *TLR3* (accuracy = 0.852), *FR-gamma* (accuracy = 0.815) and *IL6R* (accuracy = 0.743). In total, 126 proteins achieved prediction accuracy estimates ≥ 0.1 in the 10-fold cross-validation. For these proteins, prediction models were trained on all samples without missing data, and were taken forward for PWAS.

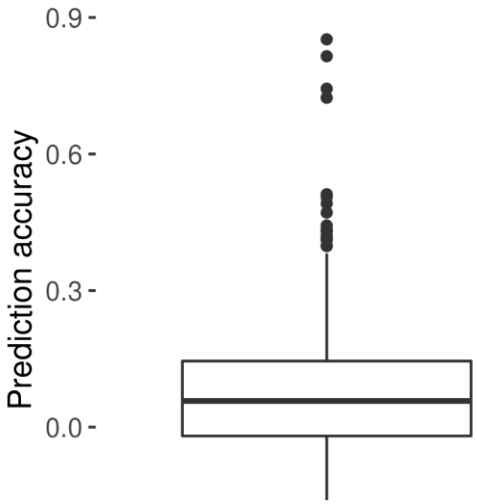

Figure 7.3. Prediction accuracy estimates for the prediction of plasma protein levels from local SNP genotypes using the PBC cases data set.

Following this, a 10-fold nested cross-validation was also conducted using the INTERVAL proteomics data, again using elastic net with alpha=0.5, lambda determined by cross-validation and a 1Mb window size. As with the previous analysis, SNPs not present in the PBC meta-analysis were removed from the INTERVAL data prior to analysis. Overall, prediction accuracy estimates were near zero for most proteins (Figure 7.4), with a mean cross-validation prediction accuracy estimate across all 3106 SOMAmers tested = 0.056, although again there existed a set of proteins where expression could be predicted accurately. In total, 549 SOMAmers achieved a prediction accuracy estimate ≥ 0.1 in the 10-fold nested cross-validation analysis. For these 549 proteins, prediction models were then trained using all INTERVAL samples, and were taken forward for PWAS.

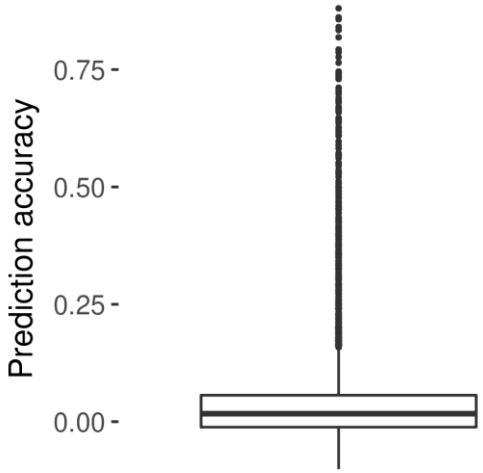

Figure 7.4. Prediction accuracy estimates for the prediction of plasma protein levels from local SNP genotypes using the INTERVAL data set.

A total of 233 proteins were tested with 10-fold nested cross-validation in the PBC cases analysis and the INTERVAL analysis. For these proteins, prediction accuracy estimates from the two cross-validation analyses were mildly positively correlated (r=0.554), although not always concordant (Figure 7.5). For example, CCL24 showed prediction accuracy = 0.505 from the PBC cases cross-validation, but only prediction accuracy = -0.041 from the INTERVAL cross-validation. The differences in prediction accuracy estimates for proteins such as this may be reflective of differences in data measurement and processing, or may be reflective of PBC-specific effects that would not be seen in a population cohort such as INTERVAL. On average, prediction accuracy estimates were slightly greater from the INTERVAL cross-validation (mean accuracy across the 233 proteins = 0.102) than from the PBC cases cross-validation (mean accuracy across the 233 proteins = 0.089), likely due to the much greater sample size of the INTERVAL data.

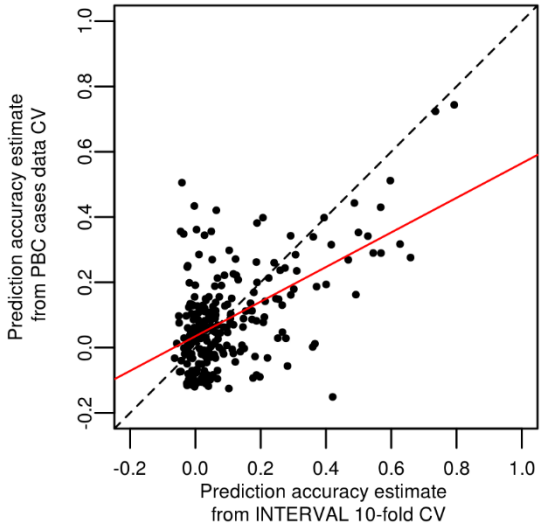

Figure 7.5. Comparison of prediction accuracy estimates from 10-fold nested cross-validation using the INTERVAL and PBC cases data sets. Each point represents a protein, with the prediction accuracy estimate achieved for the protein in the cross-validation using the INTERVAL data shown on the x axis, and the prediction accuracy estimate achieved using the PBC cases data shown on the y axis. The dashed line is the line of equality, and the red line is a best fit line.

7.4. PWAS of PBC using protein level prediction models trained using the PBC cases data

To investigate the role of genetically-regulated protein levels in PBC, a PWAS was conducted by applying the 126 protein level prediction models trained using PBC cases data to PBC meta-analysis summary statistics using MetaXcan. 3 proteins reached a Bonferroni-corrected significance threshold of $p < 3.97 \times 10^{-4}$, with 8 more proteins reaching nominal significance of $p<0.05$ (Table 7.1). Among these nominally significant proteins were a number of good candidates for further exploration in PBC, a number of interleukins (IL-10RB and IL-12B) which have roles in immune signalling, and the TNFB protein, which also plays a role in immunity.

Protein	Z score	Effect size	P value
TNFB	-8.524	-0.765	1.53E-17
IDUA	4.449	0.733	8.62E-06
ICOSLG	-3.663	-1.506	2.50E-04
RAGE	-3.186	-2.272	1.44E-03
CD40	-3.145	-0.430	1.66E-03
IL-10RB	3.003	0.880	2.67E-03
BLM hydrolase	-2.983	-0.419	2.86E-03
IL-12B	-2.818	-0.325	4.83E-03
GPC1	-2.314	-0.647	0.021
CD207	-2.265	-0.418	0.023
RARRES2	-2.127	-1.211	0.033

Table 7.1. Nominally significant results from the PBC PWAS analysis using PBC cases prediction models

7.5. PWAS of PBC using protein level prediction models trained INTERVAL data

In addition, a PWAS was conducted using the 549 INTERVAL-informed models. 11 SOMAmers, covering 10 unique proteins, reached a Bonferroni-corrected significance threshold of $p<9.11x10^{-5}$ (Table 7.2). Among the significant results were IL-12 RB2, a receptor in IL-12 signalling, which has previously been implicated in PBC. In addition, a number of other proteins involved in immunity reach significance, including ICAM proteins involved in cell-cell adhesion and movement, which have been implicated in other autoimmune diseases such as Crohn's disease, and FCRL proteins, which are involved in a range of immune processes.

SOMAmer	Protein	Z score	Effect size	P value
HLA.DQA2.7757.5.3	HLA DQA2	9.092	0.441	9.72E-20
MANBA.6382.17.3	MANBA	6.521	0.286	6.97E-11
COL11A2.11278.4.3	COL11A2	6.443	0.397	1.17E-10
IL12RB2.3815.14.1	IL-12 RB2	-6.059	-0.813	1.37E-09
FCRL1.5728.60.3	FCRL1	-5.647	-1.113	1.63E-08
FCRL3.4440.15.2	FCRL3	-5.517	-0.249	3.45E-08
ICAM5.5124.62.3	sICAM-5	4.692	0.182	2.70E-06
ICAM5.8245.27.3	sICAM-5	4.598	0.145	4.27E-06
FAM177A1.8039.41.3	F177A	4.159	0.162	3.20E-05
ERAP1.4964.67.1	ARTS1	-3.975	-0.085	7.05E-05
ICAM1.4342.10.3	sICAM-1	-3.956	-0.071	7.62E-05

Table 7.2. Bonferroni-corrected significant results from the PBC PWAS analysis using INTERVAL prediction models

In total, 50 proteins were tested in both the INTERVAL-informed PWAS and the PBC cases-informed PWAS. The z scores achieved in the two PWAS analyses were mildly positively correlated (r=0.440) for these 50 proteins, although for a small number of proteins, discordant results were achieved.

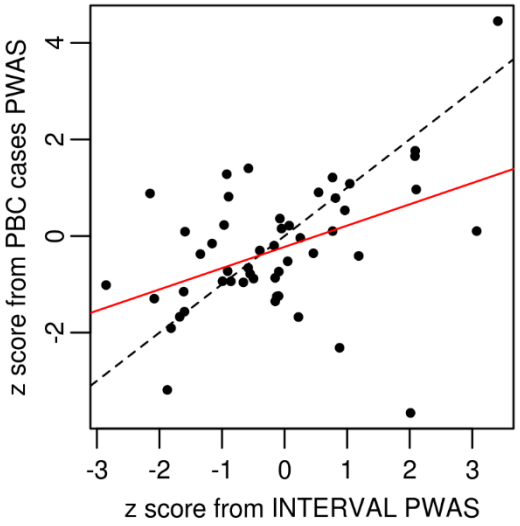

Figure 7.6. Comparison of results from PWAS using protein level prediction models trained using the INTERVAL and PBC cases data sets. Each point represents a protein, with the z score achieved for the protein in the PWAS using the INTERVAL prediction models shown on the x axis, and the z score achieved in the PWAS using the PBC cases prediction models shown on the y axis. The dashed line is the line of equality, and the red line is a best fit line.

Reassuringly, a number of the significant proteins identified through PWAS are the products of genes whose expression was found to be significantly associated with PBC through TWAS. One such protein was FCRL3, which achieved a PWAS z score of -5.517 in the INTERVAL PWAS, and a TWAS z score of -5.638 when tested using prediction models trained with GTEx whole blood data. Among the other proteins which showed strong concordance between PWAS and whole blood TWAS results HLA-DQA2 (INTERVAL PWAS z = 9.092; GTEx whole blood TWAS z = 9.792) and FAM177A1 (INTERVAL PWAS z = 4.159; GTEx whole blood TWAS z = 3.630).

7.6 Multi-trait colocalisation analysis of CpG methylation, gene expression, protein levels and PBC

While the TWAS, MWAS and PWAS approaches are powerful for the detection of associations between the intermediate traits and PBC, they are vulnerable to detecting associations between the intermediate trait and PBC that are not causal or pleiotropic, but are induced by linkage disequilibrium. To attempt to find relationships between the intermediate traits and PBC that were not induced by linkage disequilibrium, Bayesian multi-trait colocalisation was applied to PBC, expression, methylation and INTERVAL protein summary statistics using the R package *moloc*.

First, 3 sets of pairwise colocalisation analyses were conducted, each between PBC and one of the omics traits (i.e. PBC and gene expression, PBC and methylation, PBC and proteomics). The results for which at least 50 SNPs were used for colocalisation analysis, and which showed a posterior probability >= 0.8 for the hypothesis in which PBC colocalised with the omic trait were considered "significant". Following this, each possible multi-trait colocalisation involving PBC and two intermediate traits (PBC-expression-methylation, PBC-expression-protein, PBC-methylation-protein) was tested using moloc. Results in which at least 50 SNPs were used, which showed strong evidence of colocalisation between all 3 traits (PPA >= 0.8), and in which each of the two intermediate traits showed strong evidence of colocalisation in the previous pairwise test were considered "significant". Finally, multi-trait colocalisation involving PBC and all 3 intermediate traits was tested using moloc. Results in which at least 50 SNPs were used, which showed strong evidence of colocalisation between all 4 traits (PPA >= 0.8), and in which each of the three intermediate traits showed strong evidence of colocalisation in the previous pairwise tests were considered "significant".

In total, 251 colocalisation tests met these stringent conditions for "significance". After removing redundant models (for example, PBC-expression models contained entirely within PBC-expression-methylation models), 197 different models showing strong support for colocalisation between PBC and 1,2 or 3 intermediate traits remained (Figure 7.7). These 197 models covered expression of 61 unique genes, methylation at 140 unique CpGs and plasma levels of 6 unique proteins.

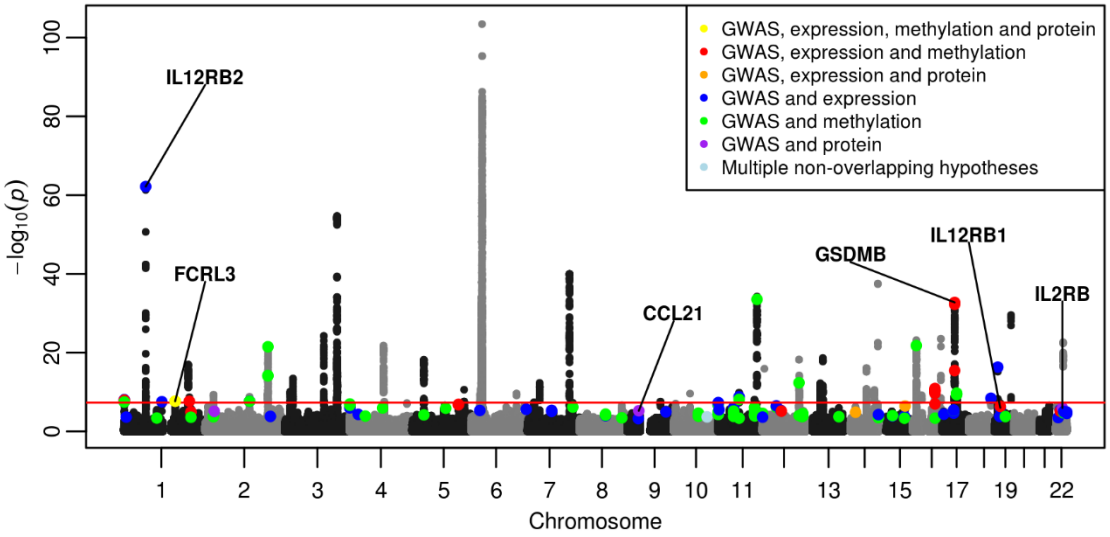

Figure 7.7. Results from moloc analysis. Each point represents a SNP, and is plotted according to its P value in the PBC meta-analysis (y axis) and its genomic location (x axis). SNPs which were identified as the potential causal SNPs by moloc are coloured, with the hypothesis they supported shown in the legend. The "multiple non-overlapping hypotheses" refers to one result where a colocalisation between PBC and the gene expression of CWF19L1 was observed at the same causal SNP as a separate colocalisation between PBC and the methylation of cg11888571, but the three-way test of colocalisation between PBC, CWF19L1 gene expression and cg11888571 methylation just missed the PPA ≥ 0.8 threshold used to define significance.

Of the 197 non-redundant models that were considered "significant", 2 showed evidence for colocalisation of all 4 traits to the same causal SNP: PBC – FCRL3 (expression) – cg17134153 – FCRL3 (protein level) and PBC – FCRL3 (expression) – cg25259754 – FCRL3 (protein level). In addition to the support from colocalisation, all variables in these models reached Bonferroni significance in their respective TWAS, MWAS and PWAS tests. Overall, evidence from multiple analyses suggest that FCRL3 and these CpGs represent good candidates in PBC.

In addition to the models containing FCRL3, a further 37 models supported colocalisation between PBC and 2 intermediate traits. Of these, 2 showed colocalisation between PBC, gene expression and protein levels: PBC – CTSH (expression) – CTSH (protein levels), and PBC – FAM177A1 (expression) – FAM177A1 (protein levels). While CTSH did not reach a Bonferroni-corrected significance level in either the TWAS or PWAS, it reached at least nominal significance in both analyses (TWAS p = 0.029, PWAS p = 1.73×10^{-4}), with concordant directions of effect. As CTSH is known to have a role in immunity, it represents a good candidate for PBC. Similarly, FAM177A1 achieved at least

nominal significance in both the PWAS and TWAS (TWAS p =2.83x10^{-4}, PWAS p = 3.20x10^{-5}), although little is known about the function of this protein, meaning much more work is required to establish its potential role in PBC.

Overall, there was some concordance between the results of the moloc analyses and the TWAS, MWAS and PWAS results, with a number of other genes, CpGs and proteins implicated in PBC by moloc also implicated through the TWAS/MWAS/PWAS analyses. These included good candidates for PBC, such as *GSDMB*, which has been implicated in other autoimmune conditions such as IBD (Chao *et al.*, 2017), and *CCR6*, a chemokine important for the chemotaxis of B cells and T cells. However, the moloc approach also enabled identification of potential candidates that were not discovered through the TWAS/MWAS/PWAS approach. For example, moloc identified colocalisation between PBC and plasma protein levels of IL2RB, yet this protein was not tested in the PWAS as its prediction model did not meet the prediction accuracy threshold.

7.7 Discussion

Through application of TWAS, MWAS, a newly-developed PWAS approach and multi-trait colocalisation analysis to summary statistics from a recent meta-analysis of PBC, many genes, CpG sites and proteins across the genome were identified as potential candidates for involvement in PBC. Reassuringly, these approaches identified many candidates that are either known to have a role in PBC, such as those in the IL-12 pathway, or that have been suggested in previous GWAS of PBC, including *GSDMB* and *IRF5*. In addition to these previously known candidates, this approach was also able to identify some candidates previously unknown in PBC. The new candidate with the strongest evidence was *FCRL3*, which has a role in the proliferation and activation of B cells. Both TWAS and PWAS supported an association between decreased expression (at the RNA and protein level) of this gene with increased PBC risk, while the moloc approach identified colocalisation of expression of this gene (at the RNA and protein level) with PBC. Furthermore, genetic polymorphisms in and near this gene have been associated with a range of other autoimmune conditions, including rheumatoid arthritis, autoimmune thyroid disease (Kochi *et al.*, 2005) and Graves' disease (Chu *et al.*, 2011), giving further confidence that this gene may play a role in PBC.

In addition to the identification of potential candidate genes for PBC, the accuracy with which plasma protein levels can be predicted from proximal SNP genotypes was also investigated in this chapter, using data from INTERVAL and from PBC cases. A 10-fold nested cross-validation experiment revealed that the plasma levels of most proteins could not be accurately predicted from SNP genotypes, with most proteins showing cross-validation prediction accuracy estimates close to zero, although there existed a set of proteins that could be predicted with a high degree of accuracy. Reassuringly, both cross-validation analyses demonstrated this, giving confidence in this conclusion. While there was some agreement between the prediction accuracy estimates achieved using the two data sets, it is worth noting that there was a small number of proteins for which the prediction accuracy estimates obtained from the two cross-validations were not concordant. The reason for this was unclear – it is possible that this was due to differences in data measurement and processing, or it could have been due to PBC-specific effects that would not be seen in a population-based cohort such as INTERVAL. Further investigation is required to determine this.

As this is the first investigation into how accurately plasma protein levels can be predicted from proximal SNP genotypes, there are no other known estimates of prediction accuracy to compare with. However, given that the expression of most genes cannot be predicted accurately from SNP genotypes (as demonstrated in Chapter 4 of this thesis), and given that there is some (but not complete) overlap between the genetic regulation of gene expression and plasma protein levels (Sun *et al.*, 2018; Yao *et al.*, 2018; Ruffieux *et al.*, 2020), the results found here are not surprising.

While there is a lack of other estimates of protein level prediction accuracy, there have been some attempts to estimate the heritability of protein levels, which represents the theoretical maximum prediction accuracy that could be achieved when predicting using SNP genotypes. Most of these heritability estimates have been obtained using a family-based approach. For example, (Liu *et al.*, 2015b) used a twin study design to examine the heritability of 342 proteins, finding an average heritability of 13.6%. While this is slightly greater than the average prediction accuracy estimates observed here, it is worth noting that (Liu *et al.*, 2015b) examined a far smaller set of proteins than examined here, so their average heritability estimate is not directly comparable with the average prediction accuracy found here. Similarly, (Wu *et al.*, 2013) estimated heritability of ~6000 proteins using parent-offspring

regression, finding an average heritability of 0.06, which is more similar to the average prediction accuracy estimate found here. However, this heritability estimation was done using data measured in LCLs, not plasma, so again the heritability estimate is not directly comparable to the prediction accuracy estimate achieved here. Additionally, these family-based approaches inherently take into account all additive genetic effects, not just those at the SNPs most proximal to the protein's genes, and so the estimates from these family-based approaches are likely to be higher than heritability estimates that only use the set of proximal SNPs.

In contrast to the family-based approaches, (Johansson *et al.*, 2013) estimated the heritability of ~160 proteins using a SNP based approach. Interestingly, they found that most proteins had poor heritability, although there was a small group of proteins with a much higher estimate, which supports the findings of the 10-fold cross-validation analyses performed here. However, this analysis also examined only a small number of proteins, and used SNPs across the genome, not just the SNPs proximal to genes, so again the heritability estimates are not directly comparable to the prediction accuracy estimates found here. In the future, estimating the heritability of protein levels that is attributable to only the SNPs proximal to the genes would provide heritability estimates that are directly comparable to the prediction accuracy estimates achieved here, and would show how close to the theoretical maximum those prediction accuracy estimate are.

The poor prediction levels observed here mean that most PWAS tests have low power to detect an association between protein levels and PBC. As the prediction models were fitted using a sparse method (elastic net), and as the INTERVAL prediction models showed a poor prediction accuracy despite being fitted using over 3,300 samples, drastic increases to prediction accuracy are unlikely to be seen by changing the method or increasing the sample size. However, one way in which prediction accuracy may be improved is to alter the window size used when fitting prediction models. Although this did not have much impact on the prediction accuracy of methylation at most CpG sites in Chapter 5, it may have a stronger effect on the prediction accuracy of protein levels. This is because the number of known trans-mQTLs is relatively small, however many trans-chromosomal effects of SNPs on protein levels have been observed, often with large effect sizes (Sun *et al.*, 2018; Yao *et al.*, 2018). However, it is worth considering that modelling the effects of SNPs across the whole genome on levels of thousands of proteins is likely to be

computationally taxing, and so it is unclear whether fitting prediction models using SNPs genome-wide would even be possible. Much further work will be needed to examine both the feasibility of increasing the window size, and to determine the impact of window size on prediction accuracy of protein levels.

To conclude, a number of caveats will be discussed. First, the TWAS, MWAS and PWAS approaches used here are all subject to the same drawbacks. These methods are all correlational in their nature, and do not prove a causal link between expression/methylation/protein level of any effectors identified here and PBC. The significant findings here may have been due to pleiotropy (where a SNP has independent effects on multiple traits), or may have been induced by linkage disequilibrium between two separate SNPs, one affecting the expression/methylation/protein level, and one affecting PBC. Causal inference methods including Mendelian Randomisation and Bayesian networks (Howey *et al.*, 2020), and functional follow-up studies will be required to identify whether the associations identified here are truly causal.

As previously established, prediction of expression/methylation/protein levels is typically poor for most genes/CpGs, and was likely even poorer in this analysis due to a lack of SNPs required for prediction. This makes interpretation of some of the results tricky, as there is a large component of expression/methylation/protein levels, regulated by either environmental factors or other genetic factors that are not accounted for. Thus, it remains to be seen to what extent these results can be translated into therapies.

Finally, PBC is an autoimmune disease of the liver, and it is thought that a number of immune cell types, such as Th1 and Th17 cells, are of particular relevance to PBC aetiology (Yang *et al.*, 2014). However, none of these cell types were analysed here, with all the data used in this analysis measured in blood, and it is unclear how well the results found in blood will translate to the more likely causal tissues and cells of PBC. There is strong evidence that the genetic architecture of gene expression at SNPs proximal to genes is shared across different tissues (Liu *et al.*, 2017), suggesting that TWAS results would translate well across different tissues. However, the (Liu *et al.*, 2017) study focussed mostly on bulk tissue, and did not examine any of the immune cell types thought to be involved in PBC. There is also some evidence of shared genetic effects on CpG methylation across a number of tissues (Shi *et al.*, 2014; Hannon *et al.*, 2018; Lin *et al.*, 2018), although again there has been no study

of the concordance of genetic effects on methylation between blood and immune cell types. In addition to this, there has been no large scale study of the concordance of genetic effects on protein levels across tissues, so the relevance of the PWAS results found here to immune cell types is once again unknown. Furthermore, there is a lack of large-scale publicly available data sets with measures of genotype and expression/methylation/protein levels in these immune cell types. Until these data sets become more widely available, the ability of TWAS/MWAS/PWAS to investigate the likely causal tissues of PBC may be limited.

Chapter 8. Conclusions and future work

In this chapter, the key findings of the thesis and opportunities for future exploration will be discussed.

Over the last few years, TWAS has become a popular post-GWAS method for investigating the role of gene expression in complex traits. Many different software packages for conducting TWAS, each of which uses a slightly different method, have been released (Gamazon *et al.*, 2015; Gusev *et al.*, 2016; Barbeira *et al.*, 2018; Barbeira *et al.*, 2019). However, prior to the work carried out in this thesis, there had been no comparison of how these different software packages performed, meaning that the effect of these methodological differences on TWAS results had not been explored. Thus, the first aim of this thesis was to compare these software packages through application to the same data.

Through application of TWAS software packages to data from Geuvadis and from WTCCC1 in Chapter 3, I showed that all TWAS software packages performed similarly to one another, both in terms of the accuracy with which they predicted gene expression and the associations they detected between gene expression and complex traits. This corroborated the results of a similar comparison of TWAS software packages (Barbeira *et al.*, 2018) that was published at a similar time to my manuscript, and suggests that when conducting a TWAS, the decision of which software package to use is unlikely to affect the results much. As discussed in Chapter 3, the main difference between the software packages was the set of genes they tested, with the PrediXcan and MetaXcan packages testing the most. Given that the aim of TWAS is often to detect as many associations between gene expression and phenotype as possible, it would seem desirable to test as many genes as possible. Thus, I consider that the PrediXcan and MetaXcan packages are the best (of those compared in this thesis) and would recommend the use of these packages going forward.

However, it is worth noting that a number of the software packages used here have since been updated and have changed their methodology slightly. For example, in the most recent release of the PrediXcan software, there is an option to impute genotypes at SNPs that are in the gene expression prediction models, but not in the

GWAS data. Furthermore, updated versions of gene expression prediction models, trained using different methods and different data, have also been released. For example, the most recent version of the PrediXcan prediction models was trained using the MASHR method (Urbut *et al.*, 2019), which uses fine-mapping data along with functional annotations to fit gene expression prediction models. These factors are likely to impact the results obtained from TWAS analysis, and so it would be valuable to repeat the analysis using these updated software and prediction models to see how well the conclusions drawn here still hold.

In Chapter 3 I also found that for many genes, the expression could not be predicted with a high degree of accuracy. Indeed, this was also observed in some earlier attempts to predict gene expression from SNP genotypes (Manor and Segal, 2013; Gamazon *et al.*, 2015), however the Manor and Segal attempt used a smaller sample size and window size than is typically used now, and the Gamazon *et al.* effort suffered from methodological issues. Furthermore, the issue of prediction accuracy had received little attention in the TWAS literature, which was surprising, as it is expected that the accuracy with which gene expression can be predicted from SNP genotypes will affect the power of the test of association between gene expression and the phenotype. This led to the second aim of my thesis, which was to investigate how accurately gene expression could be predicted from proximal SNP genotypes.

In Chapter 4, through a 10-fold nested cross-validation experiment I showed that for most genes, expression could not be predicted accurately, however there existed a set of genes where expression could be predicted with a high degree of accuracy, corroborating the results of previous studies (Manor and Segal, 2013; Gamazon *et al.*, 2015). This suggests that for most genes, the power to detect associations through TWAS is quite poor, indicating that large GWAS sample sizes may be needed to detect associations at these genes.

Following this, by using the GCTA software to estimate heritability, I found that heritability estimates were concordant with, but on average slightly smaller than, the prediction accuracy estimates observed previously. As the heritability is the upper bound on the prediction accuracy achievable when predicting using SNP genotypes, this suggests that prediction accuracy may only be improved slightly before reaching its limit. This indicates that large increases in the power of TWAS are unlikely to be achieved only by improving prediction accuracy, but that increasing the sample size

of the GWAS data to which the prediction models are applied is also likely to be required.

When investigating how accurately gene expression could be predicted from proximal SNP genotypes, the sparse modelling methods tended to outperform those with more polygenic assumptions, corroborating the findings of previous studies (Wheeler *et al.*, 2016), and indicating that the underlying proximal genetic architecture of gene expression is likely sparse. Unsurprisingly, greater average prediction accuracy was also achieved by increasing the sample size of the data used to fit the prediction models, and by carefully matching the prediction model training and testing data in terms of both ancestry and tissue. This suggests that both the data used to fit the prediction models and the data to which the prediction models are applied are both crucial in determining the prediction accuracy estimates that are achievable. Notably, most of the GTEx reference panels used to fit gene expression prediction models have small sample sizes, so there is an opportunity to improve prediction accuracy by increasing the size of those data. Furthermore, the majority of samples gathered by GTEx are of white European ancestry, and while some work has gone into gathering data from and fitting prediction models for some non-European populations (Mogil *et al.*, 2018), there is still much room for improvement here. Finally, while data has been gathered from a range of tissues by the GTEx project, the sample sizes for many of these tissues is still small. In addition, data from individual cell types, which may be more informative than data from bulk tissue, are still widely unavailable, so there is further room for improvement there. However, it is worth remembering that while the above steps will help to somewhat improve the prediction accuracy estimates, these efforts will ultimately be limited as the ceiling on prediction accuracy imposed by heritability is still poor for many genes.

Finally, the third aim of this thesis was to investigate how well the TWAS approach could be adapted to investigate the relationship between intermediate traits other than gene expression and complex phenotypes. Here, I chose to focus on CpG methylation in Chapters 5 and 6, and on serum proteomics in Chapter 7. When training and testing prediction models, poor prediction accuracy was observed for methylation at most CpG sites and for the levels of most proteins. Despite this, I found that there was a set of CpG sites and proteins where high prediction accuracy could be achieved. Interestingly, this was the same conclusion that was reached when examining the prediction of gene expression, indicating that this may be a

feature common to many different omics measurements, and suggesting that similar findings may be observed if the TWAS method were extended to other omics measurements not considered here. Furthermore, this also indicates that the MWAS and PWAS methods also suffer from one of the same weaknesses as TWAS – that, for most of the tests performed, the power to detect associations is quite poor.

For CpG methylation, I then followed this up by comparing prediction accuracy estimates to estimates of heritability obtained from application of GCTA. Once again, I found that heritability estimates were concordant with prediction accuracy estimates, suggesting that only slight improvements to prediction accuracy may be achievable before the upper bound is reached. Although I did not have time to examine the heritability of proteomics data here, it would be useful to investigate this in the future, as this would help evaluate how much prediction accuracy could be improved before reaching its upper limit. However, as that the upper bound was so close to the observed prediction accuracy estimates for both gene expression and CpG methylation, it seems likely that a similar conclusion would be reached for serum proteomics.

Despite the poor prediction accuracy estimates achieved for most CpG sites and proteins, both the MWAS and PWAS methods were able to identify many associations between predicted methylation/proteomics and complex traits in Chapters 6 and 7, including potentially trait-relevant CpG sites and proteins, indicating that MWAS and PWAS can be powerful approaches for improving understanding of mechanisms underlying GWAS hits. Following up on these analyses with colocalisation analysis revealed that many of the associations found through the TWAS, MWAS and PWAS approaches were induced by linkage disequilibrium, and were not indicative of causality. Despite this, some associations that were clearly of great relevance to their phenotype, and which could have been the result of a causal relationship between the omics trait and the complex trait were still found, such as *UBE2L3* in Chapter 6 and *FCRL3* in Chapter 7. This indicates that, although the extensions of the TWAS approach developed here are still vulnerable to the same limitations as TWAS itself, by using appropriate follow up techniques such as colocalisation, some of these limitations can be overcome.

Here, there is much opportunity for future work. Thanks to resources such as UK Biobank, GWAS data are available for thousands of complex traits. The application of TWAS, MWAS, PWAS and suitable follow-up analyses to these data, similar to the

PhenomeXcan approach (Pividori *et al.*, 2019), would provide a valuable resource to the community. Another direction that may be interesting to explore is the extension of the TWAS/MWAS/PWAS approach to other omics data. One such form of omics data that may be interesting to explore is metabolomics. There is much interest in metabolites and small molecules as these are often targetable through drugs. Using a TWAS-like approach to fit metabolite level prediction models and then to identify a set of metabolites or small molecules that are associated with disease may enable better repurposing of existing drugs to treat other complex diseases. Additionally, data sets of a good sample size that would be suitable for the fitting of metabolite level prediction models already exist (Shin *et al.*, 2014), so carrying out this type of analysis would be relatively easy.

In summary, in this thesis I have shown that when conducting a TWAS using publicly available TWAS software packages, the PrediXcan and MetaXcan packages should be chosen as they allow examination of more genes. However, for most genes examined in a TWAS, the accuracy with which the expression can be predicted is poor, meaning that there is little power to detect associations for many genes. Although this prediction accuracy can be improved slightly by using sparse modelling methods, increasing sample size, and matching training and testing data in terms of tissue and ancestry, it is unlikely to increase very much given the limit imposed by heritability. Finally, the TWAS methods can be extended to examine other omics including CpG methylation and proteomics, although, as for gene expression, the prediction accuracy for most CpG sites and proteins is quite poor. Finally, application of the MWAS and PWAS methods in addition to appropriate follow up analysis can be a powerful tool to improve the understanding of biological mechanisms at GWAS risk loci.

References

Amberger, J.S., Bocchini, C.A., Scott, A.F. and Hamosh, A. (2019) 'OMIM.org: leveraging knowledge across phenotype-gene relationships', *Nucleic Acids Res*, 47(D1), pp. D1038-D1043.

Astle, W.J., Elding, H., Jiang, T., Allen, D., Ruklisa, D., Mann, A.L., Mead, D., Bouman, H., Riveros-Mckay, F., Kostadima, M.A., Lambourne, J.J., Sivapalaratnam, S., Downes, K., Kundu, K., Bomba, L., Berentsen, K., Bradley, J.R., Daugherty, L.C., Delaneau, O., Freson, K., Garner, S.F., Grassi, L., Guerrero, J., Haimel, M., Janssen-Megens, E.M., Kaan, A., Kamat, M., Kim, B., Mandoli, A., Marchini, J., Martens, J.H.A., Meacham, S., Megy, K., O'Connell, J., Petersen, R., Sharifi, N., Sheard, S.M., Staley, J.R., Tuna, S., van der Ent, M., Walter, K., Wang, S.Y., Wheeler, E., Wilder, S.P., Iotchkova, V., Moore, C., Sambrook, J., Stunnenberg, H.G., Di Angelantonio, E., Kaptoge, S., Kuijpers, T.W., Carrillo-de-Santa-Pau, E., Juan, D., Rico, D., Valencia, A., Chen, L., Ge, B., Vasquez, L., Kwan, T., Garrido-Martin, D., Watt, S., Yang, Y., Guigo, R., Beck, S., Paul, D.S., Pastinen, T., Bujold, D., Bourque, G., Frontini, M., Danesh, J., Roberts, D.J., Ouwehand, W.H., Butterworth, A.S. and Soranzo, N. (2016) 'The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease', *Cell*, 167(5), pp. 1415-1429.e19.

Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A. and Abecasis, G.R. (2015) 'A global reference for human genetic variation', *Nature*, 526(7571), pp. 68-74.

Banovich, N.E., Lan, X., McVicker, G., van de Geijn, B., Degner, J.F., Blischak, J.D., Roux, J., Pritchard, J.K. and Gilad, Y. (2014) 'Methylation QTLs are associated with coordinated changes in transcription factor binding, histone modifications, and gene expression levels', *PLoS Genet*, 10(9), p. e1004663.

Barbeira, A.N., Bonazzola, R., Gamazon, E.R., Liang, Y., Park, Y., Kim-Hellmuth, S., Wang, G., Jiang, Z., Zhou, D., Hormozdiari, F., Liu, B., Rao, A., Hamel, A.R., Pividori, M.D., Aguet, F., Bastarache, L., Jordan, D.M., Verbanck, M., Do, R., Stephens, M., Ardlie, K., McCarthy, M., Montgomery, S.B., Segrè, A.V., Brown, C.D., Lappalainen, T., Wen, X. and Im, H.K. (2020) 'Exploiting the GTEx resources to decipher the mechanisms at GWAS loci', *bioRxiv*, p. 814350.

Barbeira, A.N., Dickinson, S.P., Bonazzola, R., Zheng, J., Wheeler, H.E., Torres, J.M., Torstenson, E.S., Shah, K.P., Garcia, T., Edwards, T.L., Stahl, E.A., Huckins, L.M., Nicolae, D.L., Cox, N.J. and Im, H.K. (2018) 'Exploring the phenotypic

consequences of tissue specific gene expression variation inferred from GWAS summary statistics', *Nat Commun*, 9(1), p. 1825.
Barbeira, A.N., Pividori, M., Zheng, J., Wheeler, H.E., Nicolae, D.L. and Im, H.K. (2019) 'Integrating predicted transcriptome from multiple tissues improves association detection', *PLoS Genet*, 15(1), p. e1007889.
Battle, A., Brown, C.D., Engelhardt, B.E. and Montgomery, S.B. (2017) 'Genetic effects on gene expression across human tissues', *Nature*, 550(7675), pp. 204-213.
Battle, A., Khan, Z., Wang, S.H., Mitrano, A., Ford, M.J., Pritchard, J.K. and Gilad, Y. (2015) 'Genomic variation. Impact of regulatory variation from RNA to protein', *Science*, 347(6222), pp. 664-7.
Battle, A., Mostafavi, S., Zhu, X., Potash, J.B., Weissman, M.M., McCormick, C., Haudenschild, C.D., Beckman, K.B., Shi, J., Mei, R., Urban, A.E., Montgomery, S.B., Levinson, D.F. and Koller, D. (2014) 'Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals', *Genome Res*, 24(1), pp. 14-24.
Bell, J.T., Tsai, P.C., Yang, T.P., Pidsley, R., Nisbet, J., Glass, D., Mangino, M., Zhai, G., Zhang, F., Valdes, A., Shin, S.Y., Dempster, E.L., Murray, R.M., Grundberg, E., Hedman, A.K., Nica, A., Small, K.S., Dermitzakis, E.T., McCarthy, M.I., Mill, J., Spector, T.D. and Deloukas, P. (2012) 'Epigenome-wide scans identify differentially methylated regions for age and age-related phenotypes in a healthy ageing population', *PLoS Genet*, 8(4), p. e1002629.
Bonder, M.J., Luijk, R., Zhernakova, D.V., Moed, M., Deelen, P., Vermaat, M., van Iterson, M., van Dijk, F., van Galen, M., Bot, J., Slieker, R.C., Jhamai, P.M., Verbiest, M., Suchiman, H.E., Verkerk, M., van der Breggen, R., van Rooij, J., Lakenberg, N., Arindrarto, W., Kielbasa, S.M., Jonkers, I., van 't Hof, P., Nooren, I., Beekman, M., Deelen, J., van Heemst, D., Zhernakova, A., Tigchelaar, E.F., Swertz, M.A., Hofman, A., Uitterlinden, A.G., Pool, R., van Dongen, J., Hottenga, J.J., Stehouwer, C.D., van der Kallen, C.J., Schalkwijk, C.G., van den Berg, L.H., van Zwet, E.W., Mei, H., Li, Y., Lemire, M., Hudson, T.J., Slagboom, P.E., Wijmenga, C., Veldink, J.H., van Greevenbroek, M.M., van Duijn, C.M., Boomsma, D.I., Isaacs, A., Jansen, R., van Meurs, J.B., t Hoen, P.A., Franke, L. and Heijmans, B.T. (2017) 'Disease variants alter transcription factor levels and methylation of their binding sites', *Nat Genet*, 49(1), pp. 131-138.

Bowden, J., Davey Smith, G. and Burgess, S. (2015) 'Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression', *Int J Epidemiol*, 44(2), pp. 512-25.
Braun, P.R., Han, S., Hing, B., Nagahama, Y., Gaul, L.N., Heinzman, J.T., Grossbach, A.J., Close, L., Dlouhy, B.J., Howard, M.A., 3rd, Kawasaki, H., Potash, J.B. and Shinozaki, G. (2019) 'Genome-wide DNA methylation comparison between live human brain and peripheral tissues within individuals', *Transl Psychiatry*, 9(1), p. 47.
Breiman, L. (2001) 'Random Forests', *Machine Learning*, 45(1), pp. 5-32.
Brynedal, B., Choi, J., Raj, T., Bjornson, R., Stranger, B.E., Neale, B.M., Voight, B.F. and Cotsapas, C. (2017) 'Large-Scale trans-eQTLs Affect Hundreds of Transcripts and Mediate Patterns of Transcriptional Co-regulation', *Am J Hum Genet*, 100(4), pp. 581-591.
Buniello, A., MacArthur, J.A.L., Cerezo, M., Harris, L.W., Hayhurst, J., Malangone, C., McMahon, A., Morales, J., Mountjoy, E., Sollis, E., Suveges, D., Vrousgou, O., Whetzel, P.L., Amode, R., Guillen, J.A., Riat, H.S., Trevanion, S.J., Hall, P., Junkins, H., Flicek, P., Burdett, T., Hindorff, L.A., Cunningham, F. and Parkinson, H. (2019) 'The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019', *Nucleic Acids Res*, 47(D1), pp. D1005-D1012.
Burgess, S., Small, D.S. and Thompson, S.G. (2017) 'A review of instrumental variable estimators for Mendelian randomization', *Stat Methods Med Res*, 26(5), pp. 2333-2355.
Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L.T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., Cortes, A., Welsh, S., Young, A., Effingham, M., McVean, G., Leslie, S., Allen, N., Donnelly, P. and Marchini, J. (2018) 'The UK Biobank resource with deep phenotyping and genomic data', *Nature*, 562(7726), pp. 203-209.
Chao, K.L., Kulakova, L. and Herzberg, O. (2017) 'Gene polymorphism linked to increased asthma and IBD risk alters gasdermin-B structure, a sulfatide and phosphoinositide binding protein', *Proceedings of the National Academy of Sciences*, 114(7), pp. E1128-E1137.
Chu, X., Pan, C.M., Zhao, S.X., Liang, J., Gao, G.Q., Zhang, X.M., Yuan, G.Y., Li, C.G., Xue, L.Q., Shen, M., Liu, W., Xie, F., Yang, S.Y., Wang, H.F., Shi, J.Y., Sun, W.W., Du, W.H., Zuo, C.L., Shi, J.X., Liu, B.L., Guo, C.C., Zhan, M., Gu, Z.H., Zhang,

X.N., Sun, F., Wang, Z.Q., Song, Z.Y., Zou, C.Y., Sun, W.H., Guo, T., Cao, H.M., Ma, J.H., Han, B., Li, P., Jiang, H., Huang, Q.H., Liang, L., Liu, L.B., Chen, G., Su, Q., Peng, Y.D., Zhao, J.J., Ning, G., Chen, Z., Chen, J.L., Chen, S.J., Huang, W. and Song, H.D. (2011) 'A genome-wide association study identifies two new risk loci for Graves' disease', *Nat Genet*, 43(9), pp. 897-901.

Cooper, N.J., Wallace, C., Burren, O.S., Cutler, A., Walker, N. and Todd, J.A. (2017) 'Type 1 diabetes genome-wide association analysis with imputation identifies five new risk regions', *bioRxiv*.

Darlay, R., Ayers, K.L., Mells, G.F., Hall, L.S., Liu, J.Z., Almarri, M.A., Alexander, G.J., Jones, D.E., Sandford, R.N., Anderson, C.A. and Cordell, H.J. (2018) 'Amino acid residues in five separate HLA genes can explain most of the known associations between the MHC and primary biliary cholangitis', *PLoS Genet*, 14(12), p. e1007833.

Das, S., Abecasis, G.R. and Browning, B.L. (2018) 'Genotype Imputation from Large Reference Panels', *Annu Rev Genomics Hum Genet*, 19, pp. 73-96.

Das, S., Forer, L., Schonherr, S., Sidore, C., Locke, A.E., Kwong, A., Vrieze, S.I., Chew, E.Y., Levy, S., McGue, M., Schlessinger, D., Stambolian, D., Loh, P.R., Iacono, W.G., Swaroop, A., Scott, L.J., Cucca, F., Kronenberg, F., Boehnke, M., Abecasis, G.R. and Fuchsberger, C. (2016) 'Next-generation genotype imputation service and methods', *Nat Genet*, 48(10), pp. 1284-1287.

de Lange, K.M., Moutsianas, L., Lee, J.C., Lamb, C.A., Luo, Y., Kennedy, N.A., Jostins, L., Rice, D.L., Gutierrez-Achury, J., Ji, S.G., Heap, G., Nimmo, E.R., Edwards, C., Henderson, P., Mowat, C., Sanderson, J., Satsangi, J., Simmons, A., Wilson, D.C., Tremelling, M., Hart, A., Mathew, C.G., Newman, W.G., Parkes, M., Lees, C.W., Uhlig, H., Hawkey, C., Prescott, N.J., Ahmad, T., Mansfield, J.C., Anderson, C.A. and Barrett, J.C. (2017) 'Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease', *Nat Genet*, 49(2), pp. 256-261.

de Los Campos, G., Vazquez, A.I., Fernando, R., Klimentidis, Y.C. and Sorensen, D. (2013) 'Prediction of complex human traits using the genomic best linear unbiased predictor', *PLoS Genet*, 9(7), p. e1003608.

Dhana, K., Braun, K.V.E., Nano, J., Voortman, T., Demerath, E.W., Guan, W., Fornage, M., van Meurs, J.B.J., Uitterlinden, A.G., Hofman, A., Franco, O.H. and Dehghan, A. (2018) 'An Epigenome-Wide Association Study (EWAS) of Obesity-Related Traits', *Am J Epidemiol*.

Dudbridge, F. (2013) 'Power and predictive accuracy of polygenic risk scores', *PLoS Genet*, 9(3), p. e1003348.
Duncan, L., Shen, H., Gelaye, B., Meijsen, J., Ressler, K., Feldman, M., Peterson, R. and Domingue, B. (2019) 'Analysis of polygenic risk score usage and performance in diverse human populations', *Nat Commun*, 10(1), p. 3328.
Fleischer, T., Tekpli, X., Mathelier, A., Wang, S., Nebdal, D., Dhakal, H.P., Sahlberg, K.K., Schlichting, E., Sauer, T., Geisler, J., Hofvind, S., Bathen, T.F., Engebraaten, O., Garred, Ø., Geitvik, G.A., Langerød, A., Kåresen, R., Mælandsmo, G.M., Russnes, H.G., Sørlie, T., Lingjærde, O.C., Skjerven, H.K., Park, D., Fritzman, B., Børresen-Dale, A.-L., Borgen, E., Naume, B., Eskeland, R., Frigessi, A., Tost, J., Hurtado, A., Kristensen, V.N. and Oslo Breast Cancer Research, C. (2017) 'DNA methylation at enhancers identifies distinct breast cancer lineages', *Nature Communications*, 8(1), p. 1379.
Fransen, K., Visschedijk, M.C., van Sommeren, S., Fu, J.Y., Franke, L., Festen, E.A.M., Stokkers, P.C.F., van Bodegraven, A.A., Crusius, J.B.A., Hommes, D.W., Zanen, P., de Jong, D.J., Wijmenga, C., van Diemen, C.C. and Weersma, R.K. (2010) 'Analysis of SNPs with an effect on gene expression identifies UBE2L3 and BCL3 as potential new risk genes for Crohn's disease', *Human Molecular Genetics*, 19(17), pp. 3482-3488.
Freytag, V., Vukojevic, V., Wagner-Thelen, H., Milnik, A., Vogler, C., Leber, M., Weinhold, L., Bohmer, A.C., Riedel-Heller, S., Maier, W., de Quervain, D.J., Ramirez, A. and Papassotiropoulos, A. (2018) 'Genetic estimators of DNA methylation provide insights into the molecular basis of polygenic traits', *Transl Psychiatry*, 8(1), p. 31.
Fryett, J.J., Inshaw, J., Morris, A.P. and Cordell, H.J. (2018) 'Comparison of methods for transcriptome imputation through application to two common complex diseases', *Eur J Hum Genet*, 26(11), pp. 1658-1667.
Fryett, J.J., Morris, A.P. and Cordell, H.J. (2020) 'Investigation of prediction accuracy and the impact of sample size, ancestry, and tissue in transcriptome-wide association studies', *Genet Epidemiol*, 44(5), pp. 425-441.
Gallagher, M.D. and Chen-Plotkin, A.S. (2018) 'The Post-GWAS Era: From Association to Function', *Am J Hum Genet*, 102(5), pp. 717-730.
Gamazon, E.R., Wheeler, H.E., Shah, K.P., Mozaffari, S.V., Aquino-Michaels, K., Carroll, R.J., Eyler, A.E., Denny, J.C., Nicolae, D.L., Cox, N.J. and Im, H.K. (2015) 'A gene-based association method for mapping traits using reference transcriptome data', *Nat Genet*, 47(9), pp. 1091-8.

Gaunt, T.R., Shihab, H.A., Hemani, G., Min, J.L., Woodward, G., Lyttleton, O., Zheng, J., Duggirala, A., McArdle, W.L., Ho, K., Ring, S.M., Evans, D.M., Davey Smith, G. and Relton, C.L. (2016) 'Systematic identification of genetic influences on methylation across the human life course', *Genome Biol*, 17, p. 61.
Giambartolomei, C., Vukcevic, D., Schadt, E.E., Franke, L., Hingorani, A.D., Wallace, C. and Plagnol, V. (2014) 'Bayesian test for colocalisation between pairs of genetic association studies using summary statistics', *PLoS Genet*, 10(5), p. e1004383.
Giambartolomei, C., Zhenli Liu, J., Zhang, W., Hauberg, M., Shi, H., Boocock, J., Pickrell, J., Jaffe, A.E., Pasaniuc, B. and Roussos, P. (2018) 'A Bayesian framework for multiple trait colocalization from summary association statistics', *Bioinformatics*, 34(15), pp. 2538-2545.
Grundberg, E., Meduri, E., Sandling, J.K., Hedman, A.K., Keildson, S., Buil, A., Busche, S., Yuan, W., Nisbet, J., Sekowska, M., Wilk, A., Barrett, A., Small, K.S., Ge, B., Caron, M., Shin, S.Y., Lathrop, M., Dermitzakis, E.T., McCarthy, M.I., Spector, T.D., Bell, J.T. and Deloukas, P. (2013) 'Global analysis of DNA methylation variation in adipose tissue from twins reveals links to disease-associated variants in distal regulatory elements', *Am J Hum Genet*, 93(5), pp. 876-90.
GTEx Consortium (2015) 'Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans', *Science*, 348(6235), pp. 648-60.
Guo, Y., Wei, Z., Keating, B.J. and Hakonarson, H. (2016) 'Machine learning derived risk prediction of anorexia nervosa', *BMC Med Genomics*, 9, p. 4.
Gusev, A., Ko, A., Shi, H., Bhatia, G., Chung, W., Penninx, B.W., Jansen, R., de Geus, E.J., Boomsma, D.I., Wright, F.A., Sullivan, P.F., Nikkola, E., Alvarez, M., Civelek, M., Lusis, A.J., Lehtimaki, T., Raitoharju, E., Kahonen, M., Seppala, I., Raitakari, O.T., Kuusisto, J., Laakso, M., Price, A.L., Pajukanta, P. and Pasaniuc, B. (2016) 'Integrative approaches for large-scale transcriptome-wide association studies', *Nat Genet*, 48(3), pp. 245-52.
Gusev, A., Lawrenson, K., Lin, X., Lyra, P.C., Jr., Kar, S., Vavra, K.C., Segato, F., Fonseca, M.A.S., Lee, J.M., Pejovic, T., Liu, G., Karlan, B.Y., Freedman, M.L., Noushmehr, H., Monteiro, A.N., Pharoah, P.D.P., Pasaniuc, B. and Gayther, S.A. (2019) 'A transcriptome-wide association study of high-grade serous epithelial ovarian cancer identifies new susceptibility genes and splice variants', *Nat Genet*, 51(5), pp. 815-823.

Gutierrez-Arcelus, M., Lappalainen, T., Montgomery, S.B., Buil, A., Ongen, H., Yurovsky, A., Bryois, J., Giger, T., Romano, L., Planchon, A., Falconnet, E., Bielser, D., Gagnebin, M., Padioleau, I., Borel, C., Letourneau, A., Makrythanasis, P., Guipponi, M., Gehrig, C., Antonarakis, S.E. and Dermitzakis, E.T. (2013) 'Passive and active DNA methylation and the interplay with genetic variation in gene regulation', *Elife*, 2, p. e00523.

Hannon, E., Knox, O., Sugden, K., Burrage, J., Wong, C.C.Y., Belsky, D.W., Corcoran, D.L., Arseneault, L., Moffitt, T.E., Caspi, A. and Mill, J. (2018) 'Characterizing genetic and environmental influences on variable DNA methylation using monozygotic and dizygotic twins', *PLoS Genet*, 14(8), p. e1007544.

Hannon, E., Weedon, M., Bray, N., O'Donovan, M. and Mill, J. (2017) 'Pleiotropic Effects of Trait-Associated Genetic Variation on DNA Methylation: Utility for Refining GWAS Loci', *Am J Hum Genet*, 100(6), pp. 954-959.

Harley, J.B., Alarcón-Riquelme, M.E., Criswell, L.A., Jacob, C.O., Kimberly, R.P., Moser, K.L., Tsao, B.P., Vyse, T.J., Langefeld, C.D., Nath, S.K., Guthridge, J.M., Cobb, B.L., Mirel, D.B., Marion, M.C., Williams, A.H., Divers, J., Wang, W., Frank, S.G., Namjou, B., Gabriel, S.B., Lee, A.T., Gregersen, P.K., Behrens, T.W., Taylor, K.E., Fernando, M., Zidovetzki, R., Gaffney, P.M., Edberg, J.C., Rioux, J.D., Ojwang, J.O., James, J.A., Merrill, J.T., Gilkeson, G.S., Seldin, M.F., Yin, H., Baechler, E.C., Li, Q.Z., Wakeland, E.K., Bruner, G.R., Kaufman, K.M. and Kelly, J.A. (2008) 'Genome-wide association scan in women with systemic lupus erythematosus identifies susceptibility variants in ITGAM, PXK, KIAA1542 and other loci', *Nat Genet*, 40(2), pp. 204-10.

He, X., Fuller, C.K., Song, Y., Meng, Q., Zhang, B., Yang, X. and Li, H. (2013) 'Sherlock: detecting gene-disease associations by matching patterns of expression QTL and GWAS', *Am J Hum Genet*, 92(5), pp. 667-80.

Hewitt, A.W., Januar, V., Sexton-Oates, A., Joo, J.E., Franchina, M., Wang, J.J., Liang, H., Craig, J.E. and Saffery, R. (2017) 'DNA methylation landscape of ocular tissue relative to matched peripheral blood', *Sci Rep*, 7, p. 46330.

Hirschfield, G.M., Dyson, J.K., Alexander, G.J.M., Chapman, M.H., Collier, J., Hübscher, S., Patanwala, I., Pereira, S.P., Thain, C., Thorburn, D., Tiniakos, D., Walmsley, M., Webster, G. and Jones, D.E.J. (2018) 'The British Society of Gastroenterology/UK-PBC primary biliary cholangitis treatment and management guidelines', *Gut*, 67(9), pp. 1568-1594.

Hoerl, A.E. and Kennard, R.W. (2000) 'Ridge Regression: Biased Estimation for Nonorthogonal Problems', *Technometrics*, 42(1), pp. 80-86.
Hormozdiari, F., van de Bunt, M., Segre, A.V., Li, X., Joo, J.W.J., Bilow, M., Sul, J.H., Sankararaman, S., Pasaniuc, B. and Eskin, E. (2016) 'Colocalization of GWAS and eQTL Signals Detects Target Genes', *Am J Hum Genet*, 99(6), pp. 1245-1260.
Howey, R., Shin, S.Y., Relton, C., Davey Smith, G. and Cordell, H.J. (2020) 'Bayesian network analysis incorporating genetic anchors complements conventional Mendelian randomization approaches for exploratory analysis of causal relationships in complex data', *PLoS Genet*, 16(3), p. e1008198.
Hueber, W., Patel, D.D., Dryja, T., Wright, A.M., Koroleva, I., Bruin, G., Antoni, C., Draelos, Z., Gold, M.H., Durez, P., Tak, P.P., Gomez-Reino, J.J., Foster, C.S., Kim, R.Y., Samson, C.M., Falk, N.S., Chu, D.S., Callanan, D., Nguyen, Q.D., Rose, K., Haider, A. and Di Padova, F. (2010) 'Effects of AIN457, a fully human antibody to interleukin-17A, on psoriasis, rheumatoid arthritis, and uveitis', *Sci Transl Med*, 2(52), p. 52ra72.
Iain, M.J. and Titterington, D.M. (2009) 'Statistical challenges of high-dimensional data', *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 367(1906), pp. 4237-4253.
Jansen, R., Batista, S., Brooks, A.I., Tischfield, J.A., Willemsen, G., van Grootheest, G., Hottenga, J.J., Milaneschi, Y., Mbarek, H., Madar, V., Peyrot, W., Vink, J.M., Verweij, C.L., de Geus, E.J., Smit, J.H., Wright, F.A., Sullivan, P.F., Boomsma, D.I. and Penninx, B.W. (2014) 'Sex differences in the human peripheral blood transcriptome', *BMC Genomics*, 15, p. 33.
Johansson, Å., Enroth, S., Palmblad, M., Deelder, A.M., Bergquist, J. and Gyllensten, U. (2013) 'Identification of genetic variants influencing the human plasma proteome', *Proc Natl Acad Sci U S A*, 110(12), pp. 4673-8.
Jorgenson, E., Melles, R.B., Hoffmann, T.J., Jia, X., Sakoda, L.C., Kvale, M.N., Banda, Y., Schaefer, C., Risch, N. and Shen, L. (2016) 'Common coding variants in the HLA-DQB1 region confer susceptibility to age-related macular degeneration', *Eur J Hum Genet*, 24(7), pp. 1049-55.
Kirsten, H., Al-Hasani, H., Holdt, L., Gross, A., Beutner, F., Krohn, K., Horn, K., Ahnert, P., Burkhardt, R., Reiche, K., Hackermuller, J., Loffler, M., Teupser, D., Thiery, J. and Scholz, M. (2015) 'Dissecting the genetics of the human transcriptome identifies novel trait-related trans-eQTLs and corroborates the regulatory relevance of non-protein coding locidagger', *Hum Mol Genet*, 24(16), pp. 4746-63.

Kochi, Y., Yamada, R., Suzuki, A., Harley, J.B., Shirasawa, S., Sawada, T., Bae, S.C., Tokuhiro, S., Chang, X., Sekine, A., Takahashi, A., Tsunoda, T., Ohnishi, Y., Kaufman, K.M., Kang, C.P., Kang, C., Otsubo, S., Yumura, W., Mimori, A., Koike, T., Nakamura, Y., Sasazuki, T. and Yamamoto, K. (2005) 'A functional variant in FCRL3, encoding Fc receptor-like 3, is associated with rheumatoid arthritis and several autoimmunities', *Nat Genet*, 37(5), pp. 478-85.
Labbé, C., Goyette, P., Lefebvre, C., Stevens, C., Green, T., Tello-Ruiz, M.K., Cao, Z., Landry, A.L., Stempak, J., Annese, V., Latiano, A., Brant, S.R., Duerr, R.H., Taylor, K.D., Cho, J.H., Steinhart, A.H., Daly, M.J., Silverberg, M.S., Xavier, R.J. and Rioux, J.D. (2008) 'MAST3: a novel IBD risk factor that modulates TLR4 signaling', *Genes Immun*, 9(7), pp. 602-12.
Lappalainen, T., Sammeth, M., Friedlander, M.R., t Hoen, P.A., Monlong, J., Rivas, M.A., Gonzalez-Porta, M., Kurbatova, N., Griebel, T., Ferreira, P.G., Barann, M., Wieland, T., Greger, L., van Iterson, M., Almlof, J., Ribeca, P., Pulyakhina, I., Esser, D., Giger, T., Tikhonov, A., Sultan, M., Bertier, G., MacArthur, D.G., Lek, M., Lizano, E., Buermans, H.P., Padioleau, I., Schwarzmayr, T., Karlberg, O., Ongen, H., Kilpinen, H., Beltran, S., Gut, M., Kahlem, K., Amstislavskiy, V., Stegle, O., Pirinen, M., Montgomery, S.B., Donnelly, P., McCarthy, M.I., Flicek, P., Strom, T.M., Lehrach, H., Schreiber, S., Sudbrak, R., Carracedo, A., Antonarakis, S.E., Hasler, R., Syvanen, A.C., van Ommen, G.J., Brazma, A., Meitinger, T., Rosenstiel, P., Guigo, R., Gut, I.G., Estivill, X. and Dermitzakis, E.T. (2013) 'Transcriptome and genome sequencing uncovers functional variation in humans', *Nature*, 501(7468), pp. 506-11.
Lewis, M.J., Vyse, S., Shields, A.M., Boeltz, S., Gordon, P.A., Spector, T.D., Lehner, P.J., Walczak, H. and Vyse, T.J. (2015) 'UBE2L3 polymorphism amplifies NF-κB activation and promotes plasma cell development, linking linear ubiquitination to multiple autoimmune diseases', *American journal of human genetics*, 96(2), pp. 221-234.
Lin, D., Chen, J., Perrone-Bizzozero, N., Bustillo, J.R., Du, Y., Calhoun, V.D. and Liu, J. (2018) 'Characterization of cross-tissue genetic-epigenetic effects and their patterns in schizophrenia', *Genome Med*, 10(1), p. 13.
Liu, J.Z., Sommeren, S., Huang, H., Ng, S.C., Alberts, R. and Takahashi, A. (2015a) 'Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations', *Nat Genet.*, 47.
Liu, X., Finucane, H.K., Gusev, A., Bhatia, G., Gazal, S., O'Connor, L., Bulik-Sullivan, B., Wright, F.A., Sullivan, P.F., Neale, B.M. and Price, A.L. (2017) 'Functional

Architectures of Local and Distal Regulation of Gene Expression in Multiple Human Tissues', *Am J Hum Genet*, 100(4), pp. 605-616.
Liu, X., Li, Y.I. and Pritchard, J.K. (2019) 'Trans Effects on Gene Expression Can Drive Omnigenic Inheritance', *Cell*, 177(4), pp. 1022-1034.e6.
Liu, Y., Buil, A., Collins, B.C., Gillet, L.C., Blum, L.C., Cheng, L.Y., Vitek, O., Mouritsen, J., Lachance, G., Spector, T.D., Dermitzakis, E.T. and Aebersold, R. (2015b) 'Quantitative variability of 342 plasma proteins in a human twin population', *Mol Syst Biol*, 11(1), p. 786.
Lundberg, M., Eriksson, A., Tran, B., Assarsson, E. and Fredriksson, S. (2011) 'Homogeneous antibody-based proximity extension assays provide sensitive and specific detection of low-abundant proteins in human blood', *Nucleic Acids Res*, 39(15), p. e102.
Luo, C., Hajkova, P. and Ecker, J.R. (2018) 'Dynamic DNA methylation: In the right place at the right time', *Science*, 361(6409), pp. 1336-1340.
Mancuso, N., Freund, M.K., Johnson, R., Shi, H., Kichaev, G., Gusev, A. and Pasaniuc, B. (2019) 'Probabilistic fine-mapping of transcriptome-wide association studies', *Nat Genet*, 51(4), pp. 675-682.
Mancuso, N., Shi, H., Goddard, P., Kichaev, G., Gusev, A. and Pasaniuc, B. (2017) 'Integrating Gene Expression with Summary Association Statistics to Identify Genes Associated with 30 Complex Traits', *Am J Hum Genet*, 100(3), pp. 473-487.
Manor, O. and Segal, E. (2013) 'Robust prediction of expression differences among human individuals using only genotype information', *PLoS Genet*, 9(3), p. e1003396.
Martin, A.R., Gignoux, C.R., Walters, R.K., Wojcik, G.L., Neale, B.M., Gravel, S., Daly, M.J., Bustamante, C.D. and Kenny, E.E. (2017) 'Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations', *Am J Hum Genet*, 100(4), pp. 635-649.
Martin, A.R., Kanai, M., Kamatani, Y., Okada, Y., Neale, B.M. and Daly, M.J. (2019) 'Clinical use of current polygenic risk scores may exacerbate health disparities', *Nat Genet*, 51(4), pp. 584-591.
Maurano, M.T., Humbert, R., Rynes, E., Thurman, R.E., Haugen, E., Wang, H., Reynolds, A.P., Sandstrom, R., Qu, H., Brody, J., Shafer, A., Neri, F., Lee, K., Kutyavin, T., Stehling-Sun, S., Johnson, A.K., Canfield, T.K., Giste, E., Diegel, M., Bates, D., Hansen, R.S., Neph, S., Sabo, P.J., Heimfeld, S., Raubitschek, A., Ziegler, S., Cotsapas, C., Sotoodehnia, N., Glass, I., Sunyaev, S.R., Kaul, R. and

Stamatoyannopoulos, J.A. (2012) 'Systematic localization of common disease-associated variation in regulatory DNA', *Science*, 337(6099), pp. 1190-5.
McCarthy, S., Das, S., Kretzschmar, W., Delaneau, O., Wood, A.R., Teumer, A., Kang, H.M., Fuchsberger, C., Danecek, P., Sharp, K., Luo, Y., Sidore, C., Kwong, A., Timpson, N., Koskinen, S., Vrieze, S., Scott, L.J., Zhang, H., Mahajan, A., Veldink, J., Peters, U., Pato, C., van Duijn, C.M., Gillies, C.E., Gandin, I., Mezzavilla, M., Gilly, A., Cocca, M., Traglia, M., Angius, A., Barrett, J.C., Boomsma, D., Branham, K., Breen, G., Brummett, C.M., Busonero, F., Campbell, H., Chan, A., Chen, S., Chew, E., Collins, F.S., Corbin, L.J., Smith, G.D., Dedoussis, G., Dorr, M., Farmaki, A.E., Ferrucci, L., Forer, L., Fraser, R.M., Gabriel, S., Levy, S., Groop, L., Harrison, T., Hattersley, A., Holmen, O.L., Hveem, K., Kretzler, M., Lee, J.C., McGue, M., Meitinger, T., Melzer, D., Min, J.L., Mohlke, K.L., Vincent, J.B., Nauck, M., Nickerson, D., Palotie, A., Pato, M., Pirastu, N., McInnis, M., Richards, J.B., Sala, C., Salomaa, V., Schlessinger, D., Schoenherr, S., Slagboom, P.E., Small, K., Spector, T., Stambolian, D., Tuke, M., Tuomilehto, J., Van den Berg, L.H., Van Rheenen, W., Volker, U., Wijmenga, C., Toniolo, D., Zeggini, E., Gasparini, P., Sampson, M.G., Wilson, J.F., Frayling, T., de Bakker, P.I., Swertz, M.A., McCarroll, S., Kooperberg, C., Dekker, A., Altshuler, D., Willer, C., Iacono, W., Ripatti, S., et al. (2016) 'A reference panel of 64,976 haplotypes for genotype imputation', *Nat Genet*, 48(10), pp. 1279-83.
McCartney, D.L., Walker, R.M., Morris, S.W., McIntosh, A.M., Porteous, D.J. and Evans, K.L. (2016) 'Identification of polymorphic and off-target probe binding sites on the Illumina Infinium MethylationEPIC BeadChip', *Genom Data*, 9, pp. 22-4.
Mele, M., Ferreira, P.G., Reverter, F., DeLuca, D.S., Monlong, J., Sammeth, M., Young, T.R., Goldmann, J.M., Pervouchine, D.D., Sullivan, T.J., Johnson, R., Segre, A.V., Djebali, S., Niarchou, A., Wright, F.A., Lappalainen, T., Calvo, M., Getz, G., Dermitzakis, E.T., Ardlie, K.G. and Guigo, R. (2015) 'Human genomics. The human transcriptome across tissues and individuals', *Science*, 348(6235), pp. 660-5.
Mikhaylova, A.V. and Thornton, T.A. (2019) 'Accuracy of Gene Expression Prediction From Genotype Data With PrediXcan Varies Across and Within Continental Populations', *Front Genet*, 10, p. 261.
Mogil, L.S., Andaleon, A., Badalamenti, A., Dickinson, S.P., Guo, X., Rotter, J.I., Johnson, W.C., Im, H.K., Liu, Y. and Wheeler, H.E. (2018) 'Genetic architecture of gene expression traits across diverse populations', *PLoS Genet*, 14(8), p. e1007586.

Momozawa, Y., Dmitrieva, J., Théâtre, E., Deffontaine, V., Rahmouni, S., Charloteaux, B., Crins, F., Docampo, E., Elansary, M., Gori, A.S., Lecut, C., Mariman, R., Mni, M., Oury, C., Altukhov, I., Alexeev, D., Aulchenko, Y., Amininejad, L., Bouma, G., Hoentjen, F., Löwenberg, M., Oldenburg, B., Pierik, M.J., Vander Meulen-de Jong, A.E., Janneke van der Woude, C., Visschedijk, M.C., Lathrop, M., Hugot, J.P., Weersma, R.K., De Vos, M., Franchimont, D., Vermeire, S., Kubo, M., Louis, E. and Georges, M. (2018) 'IBD risk loci are enriched in multigenic regulatory modules encompassing putative causative genes', *Nat Commun*, 9(1), p. 2427.

Nagpal, S., Meng, X., Epstein, M.P., Tsoi, L.C., Patrick, M., Gibson, G., De Jager, P.L., Bennett, D.A., Wingo, A.P., Wingo, T.S. and Yang, J. (2019) 'TIGAR: An Improved Bayesian Tool for Transcriptomic Data Imputation Enhances Gene Mapping of Complex Traits', *Am J Hum Genet*.

Nicolae, D.L., Gamazon, E., Zhang, W., Duan, S., Dolan, M.E. and Cox, N.J. (2010) 'Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS', *PLoS Genet*, 6(4), p. e1000888.

Pasaniuc, B., Zaitlen, N., Shi, H., Bhatia, G., Gusev, A., Pickrell, J., Hirschhorn, J., Strachan, D.P., Patterson, N. and Price, A.L. (2014) 'Fast and accurate imputation of summary statistics enhances evidence of functional enrichment', *Bioinformatics*, 30(20), pp. 2906-14.

Pividori, M., Rajagopal, P.S., Barbeira, A., Liang, Y., Melia, O., Bastarache, L., Park, Y., Wen, X. and Im, H.K. (2019) 'PhenomeXcan: Mapping the genome to the phenome through the transcriptome', *bioRxiv*, p. 833210.

Quon, G., Lippert, C., Heckerman, D. and Listgarten, J. (2013) 'Patterns of methylation heritability in a genome-wide analysis of four brain regions', *Nucleic Acids Res*, 41(4), pp. 2095-104.

Rawlik, K., Rowlatt, A. and Tenesa, A. (2016) 'Imputation of DNA Methylation Levels in the Brain Implicates a Risk Factor for Parkinson's Disease', *Genetics*, 204(2), pp. 771-781.

Relton, C.L. and Davey Smith, G. (2012) 'Two-step epigenetic Mendelian randomization: a strategy for establishing the causal role of epigenetic processes in pathways to disease', *Int J Epidemiol*, 41(1), pp. 161-76.

Relton, C.L., Gaunt, T., McArdle, W., Ho, K., Duggirala, A., Shihab, H., Woodward, G., Lyttleton, O., Evans, D.M., Reik, W., Paul, Y.L., Ficz, G., Ozanne, S.E., Wipat, A., Flanagan, K., Lister, A., Heijmans, B.T., Ring, S.M. and Davey Smith, G. (2015) 'Data

Resource Profile: Accessible Resource for Integrated Epigenomic Studies (ARIES)', *Int J Epidemiol*, 44(4), pp. 1181-90.
Richardson, T.G., Shihab, H.A., Hemani, G., Zheng, J., Hannon, E., Mill, J., Carnero-Montoro, E., Bell, J.T., Lyttleton, O., McArdle, W.L., Ring, S.M., Rodriguez, S., Campbell, C., Smith, G.D., Relton, C.L., Timpson, N.J. and Gaunt, T.R. (2016) 'Collapsed methylation quantitative trait loci analysis for low frequency and rare variants', *Hum Mol Genet*, 25(19), pp. 4339-4349.
Rowlatt, A., Hernandez-Suarez, G., Sanabria-Salas, M.C., Serrano-Lopez, M., Rawlik, K., Hernandez-Illan, E., Alenda, C., Castillejo, A., Soto, J.L., Haley, C.S. and Tenesa, A. (2016) 'The heritability and patterns of DNA methylation in normal human colorectum', *Hum Mol Genet*, 25(12), pp. 2600-2611.
Ruffieux, H., Carayol, J., Popescu, R., Harper, M.-E., Dent, R., Saris, W.H.M., Astrup, A., Hager, J., Davison, A.C. and Valsesia, A. (2020) 'A fully joint Bayesian quantitative trait locus mapping of human protein abundance in plasma', *bioRxiv*, p. 524405.
Sarkar, R.K., Rao, A.R., Meher, P.K., Nepolean, T. and Mohapatra, T. (2015) 'Evaluation of random forest regression for prediction of breeding value from genomewide SNPs', *J Genet*, 94(2), pp. 187-92.
Scott, R.A., Scott, L.J., Mägi, R., Marullo, L., Gaulton, K.J., Kaakinen, M., Pervjakova, N., Pers, T.H., Johnson, A.D., Eicher, J.D., Jackson, A.U., Ferreira, T., Lee, Y., Ma, C., Steinthorsdottir, V., Thorleifsson, G., Qi, L., Van Zuydam, N.R., Mahajan, A., Chen, H., Almgren, P., Voight, B.F., Grallert, H., Müller-Nurasyid, M., Ried, J.S., Rayner, N.W., Robertson, N., Karssen, L.C., van Leeuwen, E.M., Willems, S.M., Fuchsberger, C., Kwan, P., Teslovich, T.M., Chanda, P., Li, M., Lu, Y., Dina, C., Thuillier, D., Yengo, L., Jiang, L., Sparso, T., Kestler, H.A., Chheda, H., Eisele, L., Gustafsson, S., Frånberg, M., Strawbridge, R.J., Benediktsson, R., Hreidarsson, A.B., Kong, A., Sigurðsson, G., Kerrison, N.D., Luan, J., Liang, L., Meitinger, T., Roden, M., Thorand, B., Esko, T., Mihailov, E., Fox, C., Liu, C.T., Rybin, D., Isomaa, B., Lyssenko, V., Tuomi, T., Couper, D.J., Pankow, J.S., Grarup, N., Have, C.T., Jørgensen, M.E., Jørgensen, T., Linneberg, A., Cornelis, M.C., van Dam, R.M., Hunter, D.J., Kraft, P., Sun, Q., Edkins, S., Owen, K.R., Perry, J.R.B., Wood, A.R., Zeggini, E., Tajes-Fernandes, J., Abecasis, G.R., Bonnycastle, L.L., Chines, P.S., Stringham, H.M., Koistinen, H.A., Kinnunen, L., Sennblad, B., Mühleisen, T.W., Nöthen, M.M., Pechlivanis, S., Baldassarre, D., Gertow, K., Humphries, S.E., Tremoli, E., Klopp, N., Meyer, J., Steinbach, G., et al. (2017) 'An Expanded Genome-

Wide Association Study of Type 2 Diabetes in Europeans', *Diabetes*, 66(11), pp. 2888-2902.
Serre, D., Montpetit, A., Paré, G., Engert, J.C., Yusuf, S., Keavney, B., Hudson, T.J. and Anand, S. (2008) 'Correction of population stratification in large multi-ethnic association studies', *PLoS One*, 3(1), p. e1382.
Shi, J., Marconett, C.N., Duan, J., Hyland, P.L., Li, P., Wang, Z., Wheeler, W., Zhou, B., Campan, M., Lee, D.S., Huang, J., Zhou, W., Triche, T., Amundadottir, L., Warner, A., Hutchinson, A., Chen, P.H., Chung, B.S., Pesatori, A.C., Consonni, D., Bertazzi, P.A., Bergen, A.W., Freedman, M., Siegmund, K.D., Berman, B.P., Borok, Z., Chatterjee, N., Tucker, M.A., Caporaso, N.E., Chanock, S.J., Laird-Offringa, I.A. and Landi, M.T. (2014) 'Characterizing the genetic basis of methylome diversity in histologically normal human lung tissue', *Nat Commun*, 5, p. 3365.
Shin, S.-Y., Fauman, E.B., Petersen, A.-K., Krumsiek, J., Santos, R., Huang, J., Arnold, M., Erte, I., Forgetta, V., Yang, T.-P., Walter, K., Menni, C., Chen, L., Vasquez, L., Valdes, A.M., Hyde, C.L., Wang, V., Ziemek, D., Roberts, P., Xi, L., Grundberg, E., Waldenberger, M., Richards, J.B., Mohney, R.P., Milburn, M.V., John, S.L., Trimmer, J., Theis, F.J., Overington, J.P., Suhre, K., Brosnan, M.J., Gieger, C., Kastenmüller, G., Spector, T.D., Soranzo, N. and The Multiple Tissue Human Expression Resource, C. (2014) 'An atlas of genetic influences on human blood metabolites', *Nature Genetics*, 46(6), pp. 543-550.
Story Jovanova, O., Nedeljkovic, I., Spieler, D., Walker, R.M., Liu, C., Luciano, M., Bressler, J., Brody, J., Drake, A.J., Evans, K.L., Gondalia, R., Kunze, S., Kuhnel, B., Lahti, J., Lemaitre, R.N., Marioni, R.E., Swenson, B., Himali, J.J., Wu, H., Li, Y., McRae, A.F., Russ, T.C., Stewart, J., Wang, Z., Zhang, G., Ladwig, K.H., Uitterlinden, A.G., Guo, X., Peters, A., Raikkonen, K., Starr, J.M., Waldenberger, M., Wray, N.R., Whitsel, E.A., Sotoodehnia, N., Seshadri, S., Porteous, D.J., van Meurs, J., Mosley, T.H., McIntosh, A.M., Mendelson, M.M., Levy, D., Hou, L., Eriksson, J.G., Fornage, M., Deary, I.J., Baccarelli, A., Tiemeier, H. and Amin, N. (2018) 'DNA Methylation Signatures of Depressive Symptoms in Middle-aged and Elderly Persons: Meta-analysis of Multiethnic Epigenome-wide Studies', *JAMA Psychiatry*, 75(9), pp. 949-959.
Sun, B.B., Maranville, J.C., Peters, J.E., Stacey, D., Staley, J.R., Blackshaw, J., Burgess, S., Jiang, T., Paige, E., Surendran, P., Oliver-Williams, C., Kamat, M.A., Prins, B.P., Wilcox, S.K., Zimmerman, E.S., Chi, A., Bansal, N., Spain, S.L., Wood, A.M., Morrell, N.W., Bradley, J.R., Janjic, N., Roberts, D.J., Ouwehand, W.H., Todd,

J.A., Soranzo, N., Suhre, K., Paul, D.S., Fox, C.S., Plenge, R.M., Danesh, J., Runz, H. and Butterworth, A.S. (2018) 'Genomic atlas of the human plasma proteome', *Nature*, 558(7708), pp. 73-79.
Taylor, D.L., Jackson, A.U., Narisu, N., Hemani, G., Erdos, M.R., Chines, P.S., Swift, A., Idol, J., Didion, J.P., Welch, R.P., Kinnunen, L., Saramies, J., Lakka, T.A., Laakso, M., Tuomilehto, J., Parker, S.C.J., Koistinen, H.A., Davey Smith, G., Boehnke, M., Scott, L.J., Birney, E. and Collins, F.S. (2019) 'Integrative analysis of gene expression, DNA methylation, physiological traits, and genetic variation in human skeletal muscle', *Proc Natl Acad Sci U S A*, 116(22), pp. 10883-10888.
Tibshirani, R. (1996) 'Regression Shrinkage and Selection via the Lasso', *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1), pp. 267-288.
Urbut, S.M., Wang, G., Carbonetto, P. and Stephens, M. (2019) 'Flexible statistical methods for estimating and testing effects in genomic studies with multiple conditions', *Nature Genetics*, 51(1), pp. 187-195.
van der Pouw Kraan, T.C., Zwiers, A., Mulder, C.J., Kraal, G. and Bouma, G. (2009) 'Acute experimental colitis and human chronic inflammatory diseases share expression of inflammation-related genes with conserved Ets2 binding sites', *Inflamm Bowel Dis*, 15(2), pp. 224-35.
van Dongen, J., Nivard, M.G., Willemsen, G., Hottenga, J.J., Helmer, Q., Dolan, C.V., Ehli, E.A., Davies, G.E., van Iterson, M., Breeze, C.E., Beck, S., Suchiman, H.E., Jansen, R., van Meurs, J.B., Heijmans, B.T., Slagboom, P.E. and Boomsma, D.I. (2016) 'Genetic and environmental influences interact with age and sex in shaping the human methylome', *Nat Commun*, 7, p. 11115.
van Rheenen, W., Shatunov, A., Dekker, A.M., McLaughlin, R.L., Diekstra, F.P., Pulit, S.L., van der Spek, R.A., Võsa, U., de Jong, S., Robinson, M.R., Yang, J., Fogh, I., van Doormaal, P.T., Tazelaar, G.H., Koppers, M., Blokhuis, A.M., Sproviero, W., Jones, A.R., Kenna, K.P., van Eijk, K.R., Harschnitz, O., Schellevis, R.D., Brands, W.J., Medic, J., Menelaou, A., Vajda, A., Ticozzi, N., Lin, K., Rogelj, B., Vrabec, K., Ravnik-Glavač, M., Koritnik, B., Zidar, J., Leonardis, L., Grošelj, L.D., Millecamps, S., Salachas, F., Meininger, V., de Carvalho, M., Pinto, S., Mora, J.S., Rojas-García, R., Polak, M., Chandran, S., Colville, S., Swingler, R., Morrison, K.E., Shaw, P.J., Hardy, J., Orrell, R.W., Pittman, A., Sidle, K., Fratta, P., Malaspina, A., Topp, S., Petri, S., Abdulla, S., Drepper, C., Sendtner, M., Meyer, T., Ophoff, R.A., Staats, K.A., Wiedau-Pazos, M., Lomen-Hoerth, C., Van Deerlin, V.M., Trojanowski, J.Q., Elman, L., McCluskey, L., Basak, A.N., Tunca, C., Hamzeiy, H., Parman, Y., Meitinger, T.,

Lichtner, P., Radivojkov-Blagojevic, M., Andres, C.R., Maurel, C., Bensimon, G., Landwehrmeyer, B., Brice, A., Payan, C.A., Saker-Delye, S., Dürr, A., Wood, N.W., Tittmann, L., Lieb, W., Franke, A., Rietschel, M., Cichon, S., Nöthen, M.M., Amouyel, P., Tzourio, C., Dartigues, J.F., Uitterlinden, A.G., Rivadeneira, F., Estrada, K., Hofman, A., Curtis, C., Blauw, H.M., van der Kooi, A.J., et al. (2016) 'Genome-wide association analyses identify new risk variants and the genetic architecture of amyotrophic lateral sclerosis', *Nat Genet*, 48(9), pp. 1043-8.
Verbanck, M., Chen, C.Y., Neale, B. and Do, R. (2018) 'Detection of widespread horizontal pleiotropy in causal relationships inferred from Mendelian randomization between complex traits and diseases', *Nat Genet*, 50(5), pp. 693-698.
Vervier, K. and Michaelson, J.J. (2016) 'SLINGER: large-scale learning for predicting gene expression', *Sci Rep*, 6, p. 39360.
Visscher, P.M., Wray, N.R., Zhang, Q., Sklar, P., McCarthy, M.I., Brown, M.A. and Yang, J. (2017) '10 Years of GWAS Discovery: Biology, Function, and Translation', *Am J Hum Genet*, 101(1), pp. 5-22.
Volkov, P., Olsson, A.H., Gillberg, L., Jorgensen, S.W., Brons, C., Eriksson, K.F., Groop, L., Jansson, P.A., Nilsson, E., Ronn, T., Vaag, A. and Ling, C. (2016) 'A Genome-Wide mQTL Analysis in Human Adipose Tissue Identifies Genetic Variants Associated with DNA Methylation, Gene Expression and Metabolic Traits', *PLoS One*, 11(6), p. e0157776.
Watanabe, K., Taskesen, E., van Bochoven, A. and Posthuma, D. (2017) 'Functional mapping and annotation of genetic associations with FUMA', *Nature Communications*, 8(1), p. 1826.
Wei, Z., Wang, W., Bradfield, J., Li, J., Cardinale, C., Frackelton, E., Kim, C., Mentch, F., Van Steen, K., Visscher, P.M., Baldassano, R.N. and Hakonarson, H. (2013) 'Large sample size, wide variant spectrum, and advanced machine-learning technique boost risk prediction for inflammatory bowel disease', *Am J Hum Genet*, 92(6), pp. 1008-12.
Wellcome Trust Case Control Consortium (2007) 'Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls', *Nature*, 447(7145), pp. 661-78.
Westra, H.J., Peters, M.J., Esko, T., Yaghootkar, H., Schurmann, C., Kettunen, J., Christiansen, M.W., Fairfax, B.P., Schramm, K., Powell, J.E., Zhernakova, A., Zhernakova, D.V., Veldink, J.H., Van den Berg, L.H., Karjalainen, J., Withoff, S., Uitterlinden, A.G., Hofman, A., Rivadeneira, F., Hoen, P.A.C., Reinmaa, E., Fischer,

K., Nelis, M., Milani, L., Melzer, D., Ferrucci, L., Singleton, A.B., Hernandez, D.G., Nalls, M.A., Homuth, G., Nauck, M., Radke, D., Volker, U., Perola, M., Salomaa, V., Brody, J., Suchy-Dicey, A., Gharib, S.A., Enquobahrie, D.A., Lumley, T., Montgomery, G.W., Makino, S., Prokisch, H., Herder, C., Roden, M., Grallert, H., Meitinger, T., Strauch, K., Li, Y., Jansen, R.C., Visscher, P.M., Knight, J.C., Psaty, B.M., Ripatti, S., Teumer, A., Frayling, T.M., Metspalu, A., van Meurs, J.B.J. and Franke, L. (2013) 'Systematic identification of trans eQTLs as putative drivers of known disease associations', *Nat Genet*, 45(10), pp. 1238-1243.

Wheeler, H.E., Shah, K.P., Brenner, J., Garcia, T., Aquino-Michaels, K., Cox, N.J., Nicolae, D.L. and Im, H.K. (2016) 'Survey of the Heritability and Sparse Architecture of Gene Expression Traits across Human Tissues', *PLoS Genet*, 12(11), p. e1006423.

Wray, N.R., Ripke, S., Mattheisen, M., Trzaskowski, M., Byrne, E.M., Abdellaoui, A., Adams, M.J., Agerbo, E., Air, T.M., Andlauer, T.M.F., Bacanu, S.A., Bækvad-Hansen, M., Beekman, A.F.T., Bigdeli, T.B., Binder, E.B., Blackwood, D.R.H., Bryois, J., Buttenschøn, H.N., Bybjerg-Grauholm, J., Cai, N., Castelao, E., Christensen, J.H., Clarke, T.K., Coleman, J.I.R., Colodro-Conde, L., Couvy-Duchesne, B., Craddock, N., Crawford, G.E., Crowley, C.A., Dashti, H.S., Davies, G., Deary, I.J., Degenhardt, F., Derks, E.M., Direk, N., Dolan, C.V., Dunn, E.C., Eley, T.C., Eriksson, N., Escott-Price, V., Kiadeh, F.H.F., Finucane, H.K., Forstner, A.J., Frank, J., Gaspar, H.A., Gill, M., Giusti-Rodríguez, P., Goes, F.S., Gordon, S.D., Grove, J., Hall, L.S., Hannon, E., Hansen, C.S., Hansen, T.F., Herms, S., Hickie, I.B., Hoffmann, P., Homuth, G., Horn, C., Hottenga, J.J., Hougaard, D.M., Hu, M., Hyde, C.L., Ising, M., Jansen, R., Jin, F., Jorgenson, E., Knowles, J.A., Kohane, I.S., Kraft, J., Kretzschmar, W.W., Krogh, J., Kutalik, Z., Lane, J.M., Li, Y., Li, Y., Lind, P.A., Liu, X., Lu, L., MacIntyre, D.J., MacKinnon, D.F., Maier, R.M., Maier, W., Marchini, J., Mbarek, H., McGrath, P., McGuffin, P., Medland, S.E., Mehta, D., Middeldorp, C.M., Mihailov, E., Milaneschi, Y., Milani, L., Mill, J., Mondimore, F.M., Montgomery, G.W., Mostafavi, S., Mullins, N., Nauck, M., Ng, B., et al. (2018) 'Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression', *Nat Genet*, 50(5), pp. 668-681.

Wray, N.R., Yang, J., Hayes, B.J., Price, A.L., Goddard, M.E. and Visscher, P.M. (2013) 'Pitfalls of predicting complex traits from SNPs', *Nat Rev Genet*, 14(7), pp. 507-15.

Wu, C. and Pan, W. (2020) 'A powerful fine-mapping method for transcriptome-wide association studies', *Hum Genet*, 139(2), pp. 199-213.
Wu, L., Candille, S.I., Choi, Y., Xie, D., Jiang, L., Li-Pook-Than, J., Tang, H. and Snyder, M. (2013) 'Variation and genetic control of protein abundance in humans', *Nature*, 499(7456), pp. 79-82.
Xu, C.J., Soderhall, C., Bustamante, M., Baiz, N., Gruzieva, O., Gehring, U., Mason, D., Chatzi, L., Basterrechea, M., Llop, S., Torrent, M., Forastiere, F., Fantini, M.P., Carlsen, K.C.L., Haahtela, T., Morin, A., Kerkhof, M., Merid, S.K., van Rijkom, B., Jankipersadsing, S.A., Bonder, M.J., Ballereau, S., Vermeulen, C.J., Aguirre-Gamboa, R., de Jongste, J.C., Smit, H.A., Kumar, A., Pershagen, G., Guerra, S., Garcia-Aymerich, J., Greco, D., Reinius, L., McEachan, R.R.C., Azad, R., Hovland, V., Mowinckel, P., Alenius, H., Fyhrquist, N., Lemonnier, N., Pellet, J., Auffray, C., van der Vlies, P., van Diemen, C.C., Li, Y., Wijmenga, C., Netea, M.G., Moffatt, M.F., Cookson, W., Anto, J.M., Bousquet, J., Laatikainen, T., Laprise, C., Carlsen, K.H., Gori, D., Porta, D., Iniguez, C., Bilbao, J.R., Kogevinas, M., Wright, J., Brunekreef, B., Kere, J., Nawijn, M.C., Annesi-Maesano, I., Sunyer, J., Melen, E. and Koppelman, G.H. (2018) 'DNA methylation in childhood asthma: an epigenome-wide meta-analysis', *Lancet Respir Med*, 6(5), pp. 379-388.
Xu, Z., Wu, C. and Pan, W. (2017) 'Imaging-wide association study: Integrating imaging endophenotypes in GWAS', *Neuroimage*, 159, pp. 159-169.
Yang, C., Wan, X., Lin, X., Chen, M., Zhou, X. and Liu, J. (2019a) 'CoMM: a collaborative mixed model to dissecting genetic contributions to complex traits by leveraging regulatory information', *Bioinformatics*, 35(10), pp. 1644-1652.
Yang, C.Y., Ma, X., Tsuneyama, K., Huang, S., Takahashi, T., Chalasani, N.P., Bowlus, C.L., Yang, G.X., Leung, P.S., Ansari, A.A., Wu, L., Coppel, R.L. and Gershwin, M.E. (2014) 'IL-12/Th1 and IL-23/Th17 biliary microenvironment in primary biliary cirrhosis: implications for therapy', *Hepatology*, 59(5), pp. 1944-53.
Yang, J., Huang, T., Petralia, F., Long, Q., Zhang, B., Argmann, C., Zhao, Y., Mobbs, C.V., Schadt, E.E., Zhu, J. and Tu, Z. (2015) 'Synchronized age-related gene expression changes across multiple tissues in human and the link to complex diseases', *Sci Rep*, 5, p. 15145.
Yang, Y., Shi, X., Jiao, Y., Huang, J., Chen, M., Zhou, X., Sun, L., Lin, X., Yang, C. and Liu, J. (2019b) 'CoMM-S2: a collaborative mixed model using summary statistics in transcriptome-wide association studies', *Bioinformatics*.

Yao, C., Chen, G., Song, C., Keefe, J., Mendelson, M., Huan, T., Sun, B.B., Laser, A., Maranville, J.C., Wu, H., Ho, J.E., Courchesne, P., Lyass, A., Larson, M.G., Gieger, C., Graumann, J., Johnson, A.D., Danesh, J., Runz, H., Hwang, S.J., Liu, C., Butterworth, A.S., Suhre, K. and Levy, D. (2018) 'Genome-wide mapping of plasma protein QTLs identifies putatively causal genes and pathways for cardiovascular disease', *Nat Commun*, 9(1), p. 3268.

Zeng, P., Zhou, X. and Huang, S. (2017) 'Prediction of gene expression with cis-SNPs using mixed models and regularization methods', *BMC Genomics*, 18(1), p. 368.

Zheng, J., Richardson, T.G., Millard, L.A.C., Hemani, G., Elsworth, B.L., Raistrick, C.A., Vilhjalmsson, B., Neale, B.M., Haycock, P.C., Smith, G.D. and Gaunt, T.R. (2018) 'PhenoSpD: an integrated toolkit for phenotypic correlation estimation and multiple testing correction using GWAS summary statistics', *Gigascience*, 7(8).

Zhou, X., Carbonetto, P. and Stephens, M. (2013) 'Polygenic modeling with bayesian sparse linear mixed models', *PLoS Genet*, 9(2), p. e1003264.

Zhu, Z., Zhang, F., Hu, H., Bakshi, A., Robinson, M.R., Powell, J.E., Montgomery, G.W., Goddard, M.E., Wray, N.R., Visscher, P.M. and Yang, J. (2016) 'Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets', *Nat Genet*, 48(5), pp. 481-7.

Zou, H. and Hastie, T. (2005) 'Regularization and Variable Selection via the Elastic Net', *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 67(2), pp. 301-320.

Appendix A

CpG site	Gene	Phenotype	Step 1 effect size	Step 1 SE	Step 1 p value	Step 2 effect size	Step 2 SE	Step 2 p value
cg13896204	HLA-DQB1	Asthma	15.299	0.732	5.50E-97	-0.003	0.001	1.73E-06
cg09555323	HLA-DQB1	Asthma	20.916	1.001	5.50E-97	-0.003	0.001	1.73E-06
cg06598146	HLA-DQB1	Asthma	-19.065	1.203	1.49E-56	-0.003	0.000	1.98E-09
cg00944433	PRMT6	BMI	6.484	0.876	1.34E-13	-0.046	0.011	5.01E-05
cg09367891	PRMT6	BMI	-254.380	34.365	1.34E-13	-0.046	0.011	5.01E-05
cg15468180	PRMT6	BMI	-36.672	4.954	1.34E-13	-0.046	0.011	5.01E-05
cg00944433	PRMT6	BMI	7.897	1.067	1.34E-13	-0.046	0.011	5.01E-05
cg15468180	PRMT6	BMI	-38.065	5.265	4.85E-13	-0.046	0.011	5.01E-05
cg04414917	TUFM	BMI	-56.974	5.215	8.84E-28	0.033	0.007	3.27E-07
cg01499465	TUFM	BMI	-26.648	2.434	6.63E-28	0.033	0.007	3.27E-07
cg05684748	TUFM	BMI	-30.116	3.780	1.63E-15	0.062	0.008	4.82E-14
cg02115394	CDC16	BMR	-27.657	2.111	3.25E-39	-0.018	0.004	1.91E-05
cg24890222	CDC16	BMR	-31.264	2.370	9.70E-40	-0.018	0.004	1.91E-05
cg22958605	CDC16	BMR	-66.614	6.641	1.11E-23	-0.013	0.003	7.24E-05
cg02115394	CDC16	BMR	-102.657	8.297	3.65E-35	-0.018	0.004	1.91E-05
cg17744997	RNASET2	CD	-6.348	1.075	3.57E-09	-0.391	0.058	1.55E-11
cg16490823	RNASET2	CD	8.930	1.513	3.57E-09	-0.391	0.058	1.55E-11
cg00271210	RNASET2	CD	2.316	0.392	3.57E-09	-0.391	0.058	1.55E-11
cg16490823	RNASET2	CD	9.772	1.655	3.57E-09	-0.391	0.058	1.55E-11
cg12472693	RNASET2	CD	-22.946	3.255	1.79E-12	-0.513	0.083	5.13E-10
cg01612133	RNASET2	CD	5.238	0.887	3.57E-09	-0.391	0.058	1.55E-11
cg17991206	RNASET2	CD	-9.913	0.966	1.06E-24	-0.461	0.089	2.58E-07
cg00271210	RNASET2	CD	2.200	0.373	3.57E-09	-0.391	0.058	1.55E-11
cg19851602	SDCCAG8	DBP	-5.593	1.351	3.48E-05	0.061	0.010	5.22E-09
cg09070378	FCER1G	HAE	-10.759	2.506	1.76E-05	0.050	0.010	1.39E-06

cg03198009	ZFP57	HBMD	-4.610	0.396	2.56E-31	0.022	0.005	1.07E-05
cg00588198	ZFP57	HBMD	-5.429	0.466	2.56E-31	0.022	0.005	1.07E-05
cg05863862	ZFP57	HBMD	31.843	2.736	2.56E-31	0.022	0.005	1.07E-05
cg08041448	ZFP57	HBMD	-3.123	0.268	2.56E-31	0.022	0.005	1.07E-05
cg25209112	IGHMBP2	HDL	-42.365	3.445	9.31E-35	-0.029	0.005	9.59E-09
cg19408572	DGKQ	HDL	6.505	0.442	5.91E-49	0.034	0.007	5.72E-06
cg00741675	DGKQ	HDL	5.283	0.359	5.91E-49	0.034	0.007	5.72E-06
cg18512352	C1QTNF4	Height	19.789	4.940	6.18E-05	0.038	0.005	1.11E-13
cg20135002	C1QTNF4	Height	25.516	6.129	3.14E-05	0.038	0.005	1.11E-13
cg20135002	C1QTNF4	Height	26.450	6.072	1.32E-05	0.038	0.005	1.11E-13
cg17688768	C1QTNF4	Height	7.061	1.763	6.18E-05	0.038	0.005	1.11E-13
cg27400689	C1QTNF4	Height	-20.248	4.783	2.30E-05	0.038	0.005	1.11E-13
cg27552578	C1QTNF4	Height	31.076	7.398	2.66E-05	0.038	0.005	1.11E-13
cg18512352	C1QTNF4	Height	19.469	4.469	1.32E-05	0.038	0.005	1.11E-13
cg21025488	CDC16	Height	-37.191	3.006	3.65E-35	-0.022	0.004	4.82E-07
cg02115394	CDC16	Height	-27.657	2.111	3.25E-39	-0.022	0.004	4.82E-07
cg04718414	CDC16	Height	-64.168	4.898	3.25E-39	-0.022	0.004	4.82E-07
cg22958605	CDC16	Height	-68.471	5.226	3.25E-39	-0.022	0.004	4.82E-07
cg05616608	CDC16	Height	-71.433	5.415	9.70E-40	-0.022	0.004	4.82E-07
cg00963291	CDC16	Height	50.844	5.088	1.64E-23	-0.023	0.005	3.35E-05
cg04718414	CDC16	Height	-132.128	13.172	1.11E-23	-0.025	0.004	2.46E-12
cg24890222	CDC16	Height	-18.335	1.400	3.25E-39	-0.022	0.004	4.82E-07
cg21025488	CDC16	Height	-22.737	1.701	9.29E-41	-0.022	0.004	4.82E-07
cg14412159	CDC16	Height	53.705	4.071	9.70E-40	-0.022	0.004	4.82E-07
cg24890222	CDC16	Height	-31.264	2.370	9.70E-40	-0.022	0.004	4.82E-07
cg22877807	CDC16	Height	-23.848	1.820	3.25E-39	-0.022	0.004	4.82E-07
cg08699608	CDC16	Height	25.395	1.938	3.25E-39	-0.022	0.004	4.82E-07
cg22958605	CDC16	Height	-66.614	6.641	1.11E-23	-0.025	0.004	2.46E-12
cg24704211	CDC16	Height	-44.821	4.468	1.11E-23	-0.025	0.004	2.46E-12
cg02115394	CDC16	Height	-102.657	8.297	3.65E-35	-0.022	0.004	4.82E-07

cg00339695	SLC5A11	HGS	19.665	2.048	7.85E-22	-0.014	0.003	6.70E-05
cg12267557	SLC5A11	HGS	86.491	7.642	1.08E-29	-0.014	0.003	6.70E-05
cg06028605	SLC5A11	HGS	10.018	0.885	1.08E-29	-0.014	0.003	6.70E-05
cg02320151	SLC5A11	HGS	61.636	5.446	1.08E-29	-0.014	0.003	6.70E-05
cg08112777	SLC5A11	HGS	28.220	2.939	7.85E-22	-0.014	0.003	6.70E-05
cg02591213	SLC5A11	HGS	30.295	2.677	1.08E-29	-0.014	0.003	6.70E-05
cg06505273	SLC5A11	HGS	39.528	3.493	1.08E-29	-0.014	0.003	6.70E-05
cg00339695	SLC5A11	HGS	18.815	1.662	1.08E-29	-0.014	0.003	6.70E-05
cg26199251	SLC5A11	HGS	22.748	2.369	7.85E-22	-0.014	0.003	6.70E-05
cg07099998	SLC5A11	HGS	39.628	3.502	1.08E-29	-0.014	0.003	6.70E-05
cg04756594	SLC5A11	HGS	23.250	2.054	1.08E-29	-0.014	0.003	6.70E-05
cg01109535	SLC5A11	HGS	32.895	2.907	1.08E-29	-0.014	0.003	6.70E-05
cg06028605	SLC5A11	HGS	10.961	0.969	1.08E-29	-0.014	0.003	6.70E-05
cg06505273	SLC5A11	HGS	76.536	8.239	1.55E-20	-0.014	0.003	6.70E-05
cg15355400	SLC5A11	HGS	72.743	6.428	1.08E-29	-0.014	0.003	6.70E-05
cg05134816	SLC5A11	HGS	58.668	5.184	1.08E-29	-0.014	0.003	6.70E-05
cg10929980	SLC5A11	HGS	40.752	3.804	8.75E-27	-0.014	0.003	6.70E-05
cg14744741	SLC5A11	SBP	48.722	4.467	1.08E-27	-0.025	0.005	1.01E-06
cg01109535	SLC5A11	SBP	44.046	4.039	1.08E-27	-0.025	0.005	1.01E-06
cg03993335	SLC5A11	SBP	53.223	4.880	1.08E-27	-0.025	0.005	1.01E-06
cg22233843	HLA-DQA1	T1D	-2.267	0.228	2.99E-23	-2.086	0.142	9.75E-49
cg06598146	HLA-DQA1	T1D	-8.672	0.873	2.99E-23	-2.086	0.142	9.75E-49
cg13896204	HLA-DQA1	T1D	6.198	0.624	2.99E-23	-2.086	0.142	9.75E-49
cg24631579	HLA-DQA1	T1D	8.320	0.838	2.99E-23	-2.086	0.142	9.75E-49
cg13353717	HLA-DQA1	T1D	-6.727	0.677	2.99E-23	-2.086	0.142	9.75E-49
cg12296550	HLA-DQA1	T1D	1.475	0.304	1.20E-06	-0.636	0.116	4.75E-08
cg12296550	HLA-DQB2	T1D	-4.338	0.583	1.03E-13	0.219	0.040	4.75E-08
cg13896204	HLA-DQB2	T1D	-16.591	1.086	1.16E-52	4.446	0.099	0
cg24631579	HLA-DQB2	T1D	-22.270	1.458	1.16E-52	4.446	0.099	0
cg23336481	HLA-DQB2	T1D	13.407	3.234	3.40E-05	0.920	0.048	3.81E-82

cg13353717	HLA-DQB2	T1D	18.006	1.179	1.16E-52	4.446	0.099	0
cg22233843	HLA-DQB2	T1D	6.068	0.397	1.16E-52	4.446	0.099	0
cg06598146	HLA-DQB2	T1D	23.212	1.520	1.16E-52	4.446	0.099	0
cg15997319	CDK2AP1	TC	-3.663	0.622	3.88E-09	0.034	0.007	1.82E-06
cg25717295	CDK2AP1	TC	24.416	3.481	2.33E-12	0.032	0.008	4.74E-05
cg13539460	PPP5C	TG	-18.446	2.153	1.06E-17	0.045	0.008	7.23E-08
cg22812614	HLA-DRB1	UC	1.649	0.273	1.48E-09	1.866	0.164	4.51E-30
cg13778567	HLA-DRB1	UC	-0.949	0.196	1.24E-06	1.146	0.113	4.03E-24
cg02293354	HLA-DQA2	UC	-81.027	4.624	9.53E-69	-0.156	0.034	5.02E-06
cg22812614	HLA-DQA2	UC	-8.446	0.365	1.88E-118	-0.156	0.034	5.02E-06
cg00944433	PRMT6	Weight	6.484	0.876	1.34E-13	-0.044	0.010	1.11E-05
cg09367891	PRMT6	Weight	-254.380	34.365	1.34E-13	-0.044	0.010	1.11E-05
cg15468180	PRMT6	Weight	-36.672	4.954	1.34E-13	-0.044	0.010	1.11E-05
cg00944433	PRMT6	Weight	7.897	1.067	1.34E-13	-0.044	0.010	1.11E-05
cg15468180	PRMT6	Weight	-38.065	5.265	4.85E-13	-0.044	0.010	1.11E-05
cg22958605	CDC16	Weight	-66.614	6.641	1.11E-23	-0.021	0.005	2.45E-06
cg01499465	TUFM	Weight	-26.648	2.434	6.63E-28	0.037	0.006	2.32E-10

Appendix A. Significant results from two-step Mendelian Randomisation analysis of 342 CpG-gene-trait trios