Statistical methods and analyses in autoimmune diseases

University of Oxford

Brasenose College

DPhil in Clinical Medicine

Jamie Inshaw

04 10 2019

Contents

1 Introduction and background **1**

1.1 Background 1

1.1.1 Genetic Introduction 2

1.1.2 Types of genetic variation 2

1.1.3 The central dogma of molecular biology 4

1.1.4 Linkage disequilibrium and recombination 5

1.1.5 Hardy-Weinberg equilibrium 7

1.2 Genotyping methods 7

1.2.1 Sanger sequencing 7

1.2.2 Next generation sequencing 8

1.2.3 ImmunoChip 9

1.3 Statistical methods for genome-wide association studies 10

1.3.1 Genome-wide association studies 10

1.3.2 Bayesian statistics 12

1.4 Genetics of type 1 diabetes 15

1.4.1 Human leukocyte antigen 15

1.4.2 Regions of association outside the human leukocyte antigen 16

1.5 Thesis structure 17

2 Genetic analysis of age-at-diagnosis of type 1 diabetes **18**

2.1 Introduction and literature search 18

2.2 Methods 21

2.2.1 Genotype data and collections analysed . . . 21
2.2.2 Meta-analysis . . . 23
2.2.3 Residual-based method . . . 34
2.2.4 Imputed GWAS analysis . . . 35
2.2.5 Fine mapping . . . 36
2.2.6 Disease-association meta-analysis at various age-at-diagnosis groups 45
2.3 Results . . . 49
2.3.1 Principal component analyses . . . 49
2.3.2 Discovery analysis: meta-analysis . . . 52
2.3.3 Discovery analysis: residual-based method . . . 59
2.3.4 Imputed GWAS analysis . . . 64
2.3.5 Fine mapping . . . 65
2.3.6 Haplotype analyses . . . 72
2.3.7 Disease-association meta-analysis at various age-at-diagnosis groups 76
2.4 Discussion . . . 81

3 Type 1 diabetes variants acting primarily in early-diagnosed individuals 85
3.1 Introduction . . . 85
3.2 Methods . . . 88
3.2.1 Individuals included in analysis . . . 88
3.2.2 Loci studied . . . 89
3.2.3 Logistic regression . . . 93
3.2.4 Multinomial logistic regression . . . 95
3.2.5 Likelihood ratio test for heterogeneity in effect size between AAD groups . . . 96
3.2.6 HLA imputation . . . 97
3.2.7 Sensitivity analyses . . . 98
3.2.8 Fine mapping . . . 99
3.2.9 Colocalisation . . . 100
3.2.10 Heritability analyses by age-at-diagnosis group . . . 102

3.2.11 Prediction . . . 105
3.3 Results . . . 107
3.3.1 HLA results . . . 107
3.3.2 Non-HLA results . . . 109
3.3.3 Sensitivity analyses . . . 111
3.3.4 Differential-association between males and females . . . 116
3.3.5 Fine mapping and colocalisation . . . 119
3.3.6 Heritability analyses . . . 126
3.3.7 Prediction . . . 127
3.4 Discussion . . . 130

4 ImmunoChip analysis of type 1 diabetes with over 60,000 individuals 135
4.1 Introduction . . . 135
4.2 Methods . . . 137
4.2.1 Genotyping . . . 137
4.2.2 Relatedness and ancestry calculation . . . 137
4.2.3 Defining ImmunoChip regions . . . 138
4.2.4 Imputation and post-imputation quality control . . . 138
4.2.5 Independent case-control analyses . . . 139
4.2.6 Transmission disequilibrium test: family-based analysis . . . 140
4.2.7 Phase I and II Meta-analysis . . . 141
4.2.8 Exploration of genomic inflation factor statistic . . . 142
4.2.9 Significance thresholds . . . 142
4.2.10 Fine mapping, colocalisation and haplotype analyses . . . 143
4.2.11 ATAC-seq data processing . . . 146
4.2.12 Enrichment of prioritised T1D variants with ATAC-seq peaks . . . 150
4.2.13 Differential openness between activated and non-activated cells overlapping potentially causal T1D variants . . . 152
4.3 Results . . . 157
4.3.1 Study population . . . 157

4.3.2 Variant Quality control 161
4.3.3 ImmunoChip-wide meta-analysis 161
4.3.4 Fine mapping - European only using GUESSFM 176
4.3.5 T1D variant enrichment ATAC-seq enrichment in ATAC-seq peaks 193
4.3.6 Stimulated vs unstimulated differential openness in ATAC-seq peaks overlapping T1D prioritised variants 199
4.3.7 Colocalisation analyses 201
4.4 Discussion 203

5 Conclusions **208**

Abstract

Statistical methods and analyses in autoimmune diseases
Jamie Inshaw, DPhil in Clinical Medicine, Trinity Term, 2019
Brasenose College

Type 1 diabetes (T1D) is a common, heritable autoimmune disease, with approximately 60 chromosome regions identified that show association with disease. Cases diagnosed at < 10 years most rapidly lose the ability to endogenously produce insulin and suffer the greatest complications of the disease, including decreased life expectancy.

I therefore analysed genotype data from $15,645$ cases and tested the association of each variant at a genome-wide level with age-at-diagnosis, in an attempt to reveal genes and pathways implicated in early-diagnosed T1D. Variants within the human leucocyte antigen (HLA) region, as expected, showed association with age-at-diagnosis, but I also found a novel T1D region on chromosome 6q22.33, which associated with both age-at-diagnosis and also disease risk if examining only cases diagnosed at < 5 years.

Based on prior pancreas histology results, which showed cases diagnosed at < 7 years have a distinct immune cell infiltration signature to those diagnosed at ≥ 13 years, I then predefined specific age-at-diagnosis groups, < 7, 7-13 and ≥ 13, and restricted my analysis to examine only known T1D regions, thereby reducing the multiple testing adjustment required in genome-wide studies. Six alleles within the HLA region were more strongly associated in individuals diagnosed at < 7 years compared to ≥ 13 years, as well as six regions outside the HLA. The identified candidate causal genes indicate effects not only in the immune system but also in the islet beta cells and thymus.

Finally, I performed an ImmunoChip-wide meta-analysis of $59,974$ individuals, approximately twice the sample size of the previous largest T1D genetic analysis, and identified 74 chromosome regions associated with T1D, including 22 novel regions. I fine mapped each region and confirmed that prioritised variants were enriched in open chromatin regions in lymphoid cells.

Ongoing and future research aims to go from maps, to mechanisms, to medicines.

Acknowledgements

I would like to thank my Supervisor, Professor John Todd, for guidance and support throughout the entirety of my DPhil. He showed great faith in me to allow me to study in his group with no prior genetics experience and I have found it thoroughly fulfilling and enjoyable. Doctor Leonardo Bottolo provided support to me during my first year in Cambridge, as well as Doctor Chris Wallace. In my final year, Doctor Daniel Crouch has been extremely dedicated and helpful as my secondary supervisor, engaging in discussion about specific analyses and statistics generally, as well as reading my thesis and making suggestions. Other members of the lab have also been very helpful and supportive over the course of the last four years, namely Professor Linda Wicker, Doctor Tony Cutler, David Flores and Daniel Rainbow.

Chapter four in this Document was performed in collaboration with the University of Virginia. I worked closely with Catherine Robertson and learnt a lot from her, as well as receiving support and suggestions from Professor Steve Rich, Doctor Suna Onengut-Gumuscu and Doctor Wei-Min Chen.

Finally, I would like to thank my family and fiancée for their support throughout this process. Without their combined kindness, generosity and understanding, there is no way I could have achieved this.

Chapter 1

Introduction and background

1.1 Background

Type 1 diabetes (T1D) is a common multifactorial autoimmune disease with a prevalence of approximately 0.4%, although this varies greatly depending on the country, with incidences ranging from one case per $100,000$ per year in China [205] to ≈ 50 cases per $100,000$ per year in Finland and Sardinia [77, 34]. The disease is characterised by the destruction of pancreatic beta cells by the immune system, leading to a lack of insulin production from the pancreas and therefore increased blood sugar levels. If left untreated, the lack of insulin production will rapidly lead to death. However, in 1922, insulin from an Ox pancreas was administered to a child with T1D and a decrease in blood sugar levels was observed. In 1978, the first synthetic human insulin was developed and variations of this are still used today as the most effective way to treat T1D. However, insulin injections are costly, time consuming and prone to human error with regards to dose. More effective treatments are being sought after that could prevent T1D diagnosis or slow the disease course with a less demanding and invasive treatment regime.

T1D is a heritable disease, the probability of the offspring of a childhood-diagnosed T1D father being diagnosed with T1D is approximately 0.14 [76], compared to 0.004 in the general population. Uncovering which DNA variants are more prevalent amongst T1D individuals compared to controls is an active area of research. Highlighting variants more common in T1D individuals can point to clues as to which genes are impacting T1D risk

and how they are acting. A more complete understanding of the genetic determinants of T1D could lead to identification of biological pathways that could be targeted with therapeutic agents and potential discovery of new treatments to prevent or slow T1D progression.

In this thesis I will further explore the genetic determinants of T1D by examining the main genetic drivers of early-diagnosed T1D, which has the most aggressive disease course, and also by vastly increasing the number of individuals examined in genetic analyses of T1D.

1.1.1 Genetic Introduction

Each cell in the human body (except red blood cells) contains a nucleus that contains DNA, comprising 22 pairs of chromosomes, with an additional two X chromosomes for females and an X and Y chromosome for males. With exceptions that arise from somatic mutations, each cell in the body has an identical DNA sequence of nitrogenous bases (Adenine, Guanine, Cytosine and Thymine).

DNA is structured in a double helix, where two strands of complimentary DNA are bound to a sugar phosphate backbone, Adenine always couples with Thymine, whilst Guanine always couples with Cytosine. The vast majority of this genetic code ($> 99.9\%$) is identical between humans, but there are approximately 88 million potential differences in the genetic code between humans [45].

1.1.2 Types of genetic variation

Single nucleotide polymorphisms

Single nucleotide polymorphisms (SNPs) are the most common type of genetic variant, with approximately 80 million across the genome [45]. The definition of a SNP is when there are multiple potential nucleotides at a base. For example on one chromosome, an individual may have an Adenine on one strand (and thus a Thymine on the other strand) and on the second chromosome they could have a Guanine on the same strand at the same position (accompanied by a Cytosine on the other). This individual would be heterozygous at the SNP, since they have one copy of the Adenine/Thymine base and

one copy of the Guanine/Cytosine base. Homozygous individuals at this SNP would have two copies of either the Adenine/Thymine base or the Guanine/Cytosine base. Since it is known that Adenine is always accompanied with Thymine, one of the bases is named, so the three potential genotypes in this example would be Adenine Adenine (AA), Adenine Guanine (AG), or Guanine Guanine (GG).

Across a population, variants will be present at particular frequencies; new mutations will be seen less often than historical mutations, whilst variants that were selected due to evolutionary advantage might be particularly common. The minor allele frequency (MAF) is used to determine how often a variant appears:

$$MAF = \frac{((2 \times N_{aa}) + N_{aA})}{2 \times (N_{aa} + N_{aA} + N_{AA})}, \tag{1.1}$$

where N_{aa}, N_{aA} and N_{AA} are the number of individuals with genotype aa, aA and AA, respectively; and MAFs can be compared across different ancestral populations.

Insertions and deletions

Another type of variation between human DNA can be in the form of an insertion or a deletion. This can be, for example, if one individual has the following code on one chromosome: AGCTCAC and the other chromosome has ACAC, there is a sequence, GCT, which is inserted in the first chromosome that is not present in the second. Homozygous individuals in these instances would either have AGCTCAC on both chromosomes or ACAC on both chromosomes. There are estimated to be $> 800,000$ insertions or deletions of < 50 base pairs in length in the human genome [38].

Inversions and duplications

Often referred to as structural variants, a third type of genetic variation is an inversion, where two parts of the genome have switched places with one another. For example, if a DNA sequence on one chromosome was AGTAGGCG and the other were GGCGAGTA, the two 4-base sequences, AGTA and GGCG, have switched in their position in the genome. They are typically many base pairs long, unlike this toy example, and there are

estimated to be between 136 [35] and 156 [38] of these in the human genome, although the application of long-read DNA sequencing technologies is beginning to reveal variants that were not detected by short read sequencing [126]. Another type of structural variant is a duplication, where a, usually long, DNA sequence is repeated.

1.1.3 The central dogma of molecular biology

Particular stretches of DNA can form sequences that give rise to genes. There are approximately $20,000$ genes in the human genome [67]. When required, a gene is expressed in a cell by producing a complementary strand of RNA. RNA is similar to DNA but the nitrogenous base Uracil replaces Thymine and it is most often single-stranded as opposed to the double-stranded DNA. Transcription of RNA is initiated when RNA polymerase, as well as other molecules called transcription factors, bind to the start of the gene, known as the promoter or 5' untranslated region (UTR) and the complementary RNA is generated until the end of the gene, the 3' UTR. The complementary strand of RNA is called precursor messenger RNA. The precursor messenger RNA is transformed into messenger RNA (mRNA) within the nucleus by removal of the introns of the genes, keeping only the exons, which is called splicing.

The mRNA molecule is translated into a protein after entering a structure called the ribosome. However, the mRNA can serve a number of different purposes by itself, for example acting as an enhancer, which can initiate transcription of other genes (enhancer RNA), or mediating translation of other mRNAs into proteins (microRNA).

Variants that disrupt a protein coding sequence are called missense or non-synonymous variants, and these variants have been shown to be enriched for common disease associations, for example $5-10\%$ of common metabolic phenotypes are located in coding regions of the genome (synonymous or non-synonymous), despite only approximately 1.2% of the genome encoding for protein [16, 46]. If a missense variant is identified that is disease-associated, one can assume, in most cases, but not all, that it is involved in perturbing disease risk, since the structure of the downstream translated protein is altered, and this could lead to an altered biological process that increases or decreases disease risk.

However, many disease-associated variants do not change the coding sequence of the translated protein. Therefore, it is thought that many disease-associated variants can act in other ways to alter disease risk, for example they could alter the function or efficiency of long non-coding RNAs (lncRNAs), of which there are tens of thousands in the human genome [92], they could affect splicing, giving rise to increased number of proteins with different numbers of exons included in the translation, or they could be located in transcription factor binding sites, affecting the ability of transcription factors to bind to promoters and thus affecting the expression levels of a particular mRNA.

1.1.4 Linkage disequilibrium and recombination

During human meiosis, the germ-line cells from the mother and father each replicate their DNA inside the nucleus. However, an exact replicate of each chromosome is not passed down from parent to offspring. Instead, recombination occurs, where the DNA from the two sets of DNA replicates crossover and bind to one another, forming chromosomes that are different from the parental chromosome. One of these two chromosomes from each of the 23 pairs in humans is then passed down to form the gamete. It has been shown that recombination is not uniform across the genome, and instead there are approximately $25,000$ recombination hotspots, where crossover events are more likely than in other regions [134]. The physically closer two variants are to one another, the less likely it is that there will be a recombination event between the variants and therefore the more likely it is that the two variants will appear together in the offspring chromosome. As a result of this, nearby variants are often inherited together more than expected based on their frequencies in the population, which is termed linkage disequilibrium (LD).
Examining variants at the population level gives the opportunity to calculate LD between variants. Two metrics are used to calculate LD, the first is D', defined by first defining D:

$$D = p_{ab} - p_a p_b, \tag{1.2}$$

where p_{ab} is the proportion of times the ab haplotype appears in the population, where

a haplotype is the combination of variants along a single chromosome, and p_a and p_b are the allele frequencies of a and b in the population, respectively. The haplotype can be inferred for each combination of alleles except for the combination where both variants are heterozygous. As from genotype data alone the haplotypes are ambiguous, to estimate the haplotypes, an Expectation-Maximisation (EM) algorithm is typically used, which is explained in more detail in chapter 2. The D' value is a normalised form of D which is bounded $(-1, 1)$ and has the following definition:

$$D' = \begin{cases} \frac{D}{D_{max}} \ if \ D \geq 0 \\ \frac{D}{D_{min}} \ if \ D < 0, \end{cases} \tag{1.3}$$

where

$$\begin{cases} D_{max} = min(p_a(1-p_b), (1-p_a)p_b) \\ D_{min} = max(-p_a p_b, -(1-p_a)(1-p_b)). \end{cases} \tag{1.4}$$

Under this definition of LD, if one variant was very rare and the other very common, and it happened that the rare variant was only ever present when the common variant is present in the population under study, the D' value would be 1, but in reality this is likely to be a chance finding due to the rareness of the first variant. To overcome this limitation, another measure of LD is often used, r^2:

$$r^2 = \frac{D^2}{p_a p_b (1-p_a)(1-p_b)}, \tag{1.5}$$

which also takes into consideration the MAF of both variants. LD between variants can make identification of the causal variant for a disease difficult to identify, since if one variant in the block in high LD is associated with disease, all the variants in the block will show some disease association.

1.1.5 Hardy-Weinberg equilibrium

The Hardy-Weinberg Equilibrium (HWE) principal states that, in the absence of evolutionary factors and with the assumption of random mating, genotype allele frequencies will remain stable across generations. If a variant is consistent with HWE, then the frequency, f of the genotypes AA, aa and aA in the population will be approximately equal to:

$$\begin{cases} f_{AA} = p_A^2 \\ f_{aa} = p_a^2 \\ f_{aA} = 2 \times p_a \times p_A, \end{cases} \tag{1.6}$$

where p_A and p_a are the allele frequency of A and a, respectively. If a variant has genotype frequencies that are inconsistent with HWE then either the variant could have been subject to very strong natural selection, which is unlikely, or there may have been inbreeding (more homozygotes) or outbreeding (more heterozygotes), or alternatively it can be an indicator of poor genotyping of that variant. Therefore HWE checks are performed on each variant prior to analysis, with those strongly inconsistent with HWE often excluded.

1.2 Genotyping methods

1.2.1 Sanger sequencing

A breakthrough in sequencing DNA came in 1977 when Frederick Sanger proposed a method to sequence fragments of DNA [165]. This method was taken up widely and has become known as Sanger sequencing. A brief outline of Sanger sequencing is given below. To sequence a fragment of DNA of interest using Sanger sequencing, the sequence upstream of the sequence of interest must be known. A 'library' is then prepared that contains many copies of the upstream sequence and the sequence of interest for an individual. A primer (a short piece of DNA of known sequence) can then be designed which

will anneal to the known upstream sequence containing complementary bases. Once the primer has bound to the upstream sequence, DNA polymerase begins to add bases to the primer which are complementary to the DNA being sequenced. By itself, this is not informative, since there is no way of knowing which base has been added to the primer. However, terminating bases are included in the solution, which are fluorescently labelled, with a different colour for each base. When a termination base is added, which happens randomly, no further bases are added to the DNA strand as they are modified from naturally occuring bases to have an additional 'deoxy' group, which blocks the polymerase from adding further bases. Since this happens many times, the termination base will eventually appear in each position in the sequence of interest many times and thus there will be fluorescent labels at each base in the sequence.

To identify the order in which these bases appear, each DNA fragment is placed in polyacrylamide gel and an electric current is passed through the gel. Since DNA fragments are negatively charged, the fragments will move towards the positive pole in the gel, with longer fragments encountering more resistance and therefore travelling less far. Therefore, after the electric current has been passed through the gel, the DNA fragments will be ordered by fragment size and fluorescently labelled according to the complementary base of the fragment of interest. Shining ultraviolet light on the gel allows these colours to be observed and therefore the DNA sequence to be inferred.

1.2.2 Next generation sequencing

The limitation of Sanger sequencing is that, although it is accurate, to sequence and assemble the entire genome is very timely and costly. To sequence the entire genome in a more efficient way, next generation sequencing (NGS) was developed [131]. I will outline Illumina NGS, but a variety of other approaches exist.

The first part of the procedure is to cut the genome into smaller fragments of DNA using restriction enzymes. Using a low concentration of restriction enzyme will ensure that the DNA is not cut at every possible point and therefore there will likely be overlapping sequences along the genome, which allows assembly of the DNA sequences into the ap-

propriate order once sequencing has occurred. For each fragment of DNA, an adaptor sequence of DNA is attached to one end of the sequence and a primer is attached to the other end. The adaptor sequence allows the DNA fragments to attach to a glass plate, on which all the DNA fragments are attached, and the primer is attached so that the DNA can be replicated via a polymerase chain reaction (PCR). PCR is required to ensure the downstream fluorescent signal is sufficiently strong to determine the genotype at each base.

NGS then differs from Sanger sequencing by using only terminator bases in the DNA polymerase, so that a single base will attach to the chain and omit a fluorescent light when ultraviolet light is shone on it, the colour of which will be determined by which base is attached. A picture is taken of the fluorescent light omitted from each cluster of DNA fragments on the glass plate, before the terminator base is chemically amended to become a standard base. The DNA polymerase is then able to attach another terminator base to the chain, complimentary to the fragment being sequenced. This process is repeated until the end of the fragment, and the sequence of the fragment can be inferred by examining each picture of fluorescent lights.

1.2.3 ImmunoChip

Most genotyping platforms uniformly genotype variants across the genome. Whole genome sequencing (WGS) would be the gold standard to test every variant in the human genome for association with disease, but WGS is still expensive to do, so the more affordable genome-wide platforms that do not sequence every variant but have reasonable coverage across the genome are often used. Genotype imputation can be carried out to infer the probability of each possible genotype at variants not present on the genotyping platform, which is explained in more detail in chapter 2, but there is less confidence drawing conclusions about associations with variants that have been imputed compared to directly genotyped variants.

The ImmunoChip is a genotyping platform that is designed to densely genotype regions of the genome that are of particular interest for autoimmune diseases [50]. It was designed

after the initial Wellcome Trust Case Control Consortium (WTCCC) genome-wide association study (GWAS) that examined the genetic architecture of seven common diseases [48]. It contains 196,524 variants, 718 of which are small insertions or deletions, the rest are SNPs. It was much cheaper to genotype DNA samples using the ImmunoChip than other GWAS arrays, as well as being well-suited to identify with more confidence the causal variant in the typed regions, owing to the density of variants typed in these regions, thus there was a number of publications for various immune-related disorders that genotyped the DNA of cases and controls using the ImmunoChip [118, 193, 143, 18, 144]. It is important to note that due to the design of this chip, there is bias in any analyses that attempt to make genome-wide inferences about where the disease-associated variants lie, since only previously identified immune-related regions are analysed.

1.3 Statistical methods for genome-wide association studies

1.3.1 Genome-wide association studies

One of the first GWAS analysis was performed in 2007, where approximately 2,000 cases from each of seven common diseases were compared to 3,000 controls without any of the seven diseases [48]. The analysis methods used in this paper are largely similar to how GWAS are performed today. They assessed the association of each variant on the genotyping platform with disease, one variant at a time.

Multiple testing burden

A key factor that must be accounted for in any GWAS is the multiple testing burden. If each variant is tested for association with disease, by chance, many of the 469,557 variants assessed in [48] will show nominal association with disease ($p < 0.05$). Reporting every nominal association would lead to high false-positive and irreproducibility rates amongst the reported variants, and therefore a more stringent threshold of $p < 5 \times 10^{-8}$ was adopted for GWAS loci to be declared associated with disease (details in chapter 2).

Statistical power

Sample size, MAF and variant effect size are all factors that affect the statistical power of a test. Statistical power is defined as (1−type II error), where type II error is the probability of failing to reject a null hypothesis which is actually false. In a single GWAS, in which each variant is assessed on approximately the same number of individuals, the statistical power when assessing the association of a rare variant compared to a common variant is less if the effect size of the two variants are identical. Intuitively this is logical, since if there are insufficient cases and controls with a particular allele, it is difficult to distinguish an effect of the genotype on disease from chance. Similarly, if two variants in the same analysis have an identical MAF, but one variant has a large effect size on disease risk and the other has a subtle effect, the statistical power will be higher to reject the null hypothesis for the variant with a large effect size.

Even at the time, it was recognised that the 2007 GWAS did not have sufficient statistical power to detect true effects at many of the tested loci [48], as the effect sizes were assumed to very rarely be above an odds ratio of two. The design of $14,000$ cases and $3,000$ shared controls was determined by the fact that the consortium wanted to be inclusive, examining as many common conditions as possible in the budget available. One of the diseases examined in this analysis was Rheumatoid arthritis (RA), for which 4 regions were found to be associated with disease. As sample sizes have increased to over $100,000$ individuals included in the largest RA genetic analysis, 101 regions have been convincingly associated with disease. This shows the importance of having sufficient number of individuals to detect and fine map the underlying genetic architecture of the disease [143].

Fine mapping

After the initial variant discovery GWAS, it is typical that in each chromosome region associated with disease, there will be more than one variant showing association. This is usually because the causal variant is in LD with other variants in the region, causing the other variants to appear to be disease-associated when they are not truly causal for

disease risk. Many methods, termed 'fine-mapping analyses', have been developed that attempt to prioritise the more likely causal variants in each region, either through inducing sparsity in models, which prioritises more strongly those variants with a stronger disease association [132, 124] and/or through incorporation of other data types to prioritise those variants located in functional regions in the genome [104]. For example, those variants that are in enhancers in a relevant cell type, or through elucidation of the number of causal signals in the region by exploring many different models rather than assuming the variant with the strongest association in the GWAS analysis, the index variant, is causal [202, 8]. These methods are explored in more detail in chapter 2.
Fine mapping is performed for two reasons. The first reason is that if a casual variant is part of a large LD block, this means researchers attempting to identify the mechanism of this association on disease risk have many possible variants to consider in their follow-up experiment. If the number of potentially causal variants can be reduced, this will be beneficial with regards to time and cost saving. The second reason is that, in combination with other data types, it can help to point to a cell type that the disease variant is acting in. If each of the likely disease-associated variants are in quiescent chromatin in one cell type, but many are in functional chromatin regions in another disease relevant cell type, this might point researchers towards the second cell type to conduct their functional work in. Reducing the number of potential causally associated variants gives a finer resolution to this analysis.

1.3.2 Bayesian statistics

Bayesian statistics is a branch of statistics where inference is not only drawn from the data but also from a prior probability (i.e. degree of belief), which is used in conjunction with the data to calculate posterior probability distributions. Bayesian methodologies are often useful since they calculate probability distributions for unknown parameters of interest, which can be distributed in an infinite number of ways and they are also directly comparable across studies. Frequentist methodologies usually estimate a parameter as a point estimate and show a 95% confidence interval around the point estimate, leaving

less scope for the entire distribution of this parameter to be explored. This exploration of the entire distribution of a parameter is more informative and especially useful in fine mapping analyses when different models are being compared, since the models with less probability are not ruled out entirely as they might be if frequentists methods were used. Bayes theorem was first derived in 1763 [17] and states the following:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}. \tag{1.7}$$

In Equation 1.7, $P(A|B)$ is referred to as the posterior probability of A given B, which is a product of the likelihood (the probability of B conditional on A), $P(B|A)$, and the prior for A, $P(A)$, divided by a normalising factor, $P(B)$, which is the probability of B under all possible values of A, which is equivalent to $\int P(B|A)P(A)dA$. Equation 1.7 could be restated in terms of a data matrix, $\mathbf{X}$ and a parameter of interest, θ:

$$P(\theta|\mathbf{X}) = \frac{P(\mathbf{X}|\theta)P(\theta)}{\int P(\mathbf{X}|\theta)P(\theta)d\theta}. \tag{1.8}$$

If there is a prior belief about the distribution of the unknown parameter, θ, this is incorporated into the posterior distribution. The integral in Equation 1.8 can often be extremely difficult or impossible to calculate in practise. Therefore, an alternative to exact calculation of this integral is to sample from the posterior distribution using a Markov Chain Monte-Carlo (MCMC) scheme [73], which will eventually converge to samples from a distribution that approximates the exact posterior distribution. A feature of this mechanism is that it samples from each parameter in the model and therefore the marginal posterior distribution of interest is averaged across all the other possible values of all the other parameters in the model. This marginal posterior distribution is not conditional on the other parameters in the model and thus is simple to interpret.

Markov-Chain Monte Carlo

MCMC methods work by attempting to sample from the unknown target distribution, the posterior probability of the parameter of interest. They work by initially sampling

from a proposal distribution, which must have enough variance to cover the majority of the posterior distribution, since if, for example, the proposal distribution is bounded $(0, 1)$ yet the unknown target distribution is bounded $(10, 20)$, sampling from this proposal will never identify any points that have high posterior probability in the target distribution. The 'Monte-Carlo' aspect of MCMC methods is the nature of the random selection of a value from the proposal distribution, sampled as θ_t, for the t'th iteration of the algorithm. The next random draw is assumed to depend on the data and the previous Monte-Carlo draw only. If the probability of the next draw from the proposal distribution, $P(\theta_{t+1}|\mathbf{y})$, is larger than the current probability, $P(\theta_t|\mathbf{y})$, the algorithm accepts the new value, θ_{t+1}, and repeats the process. However, if the probability of the new parameter value given the data, $P(\theta_{t+1}|\mathbf{y})$, is less than the current probability, $P(\theta_t|\mathbf{y})$, an acceptance probability is generated, $a = \frac{P(\theta_{t+1}|\mathbf{y})}{P(\theta_t|\mathbf{y})}$. The new value, θ_{t+1}, is accepted if the acceptance probability is less than a randomly generated number bounded $(0, 1)$. This means that even if the new value, θ_{t+1}, is less probable given the data than the current value, there is a chance the algorithm will still explore this space, and thus the algorithm can escape from local maxima in a high dimensional space and explore other parts of the posterior.

MCMC algorithms usually converge to a distribution that approximates the posterior once all the accepted draws for θ are collated. There are many varieties of MCMC, including the Metropolis algorithm, which has the condition that the proposal distribution must be symmetric, Metropolis-Hastings algorithm, which is similar to the Metropolis but without the restriction on the proposal distribution, and the Gibbs sampler, which is a special case of the Metropolis-Hastings algorithm that samples through each parameter separately and accepted/rejects the value of the new parameter, θ_{t+1}, conditional on the previous iteration values from the other parameters (which could be, for example β_t, α_t).

Bayes Factors

In frequentists statistics, the p value is a measure of strength of association, where the p value is the probability of observing the data (or more extreme) if the null hypothesis were true. In Bayesian statistics, the strength of association is tested using a Bayes Factor.

The Bayes Factor is defined as the ratio of the marginal likelihood under an alternative model to the marginal likelihood under the null model, therefore the higher the Bayes Factor, the greater the strength of evidence against the null hypothesis.

1.4 Genetics of type 1 diabetes

1.4.1 Human leukocyte antigen

The first genetic association with T1D was with the Human Leukocyte Antigen (HLA) region in the 1970s [179, 183, 173], a long time before the advent of NGS, specifically with the class II DRB1 and DQB1 genes. HLA class II molecules are classically expressed in professional antigen presenting cells (APC), and bind to peptides that are consequently presented to $CD4^+$ T cells. If a $CD4^+$ T cell has a particular receptor that will bind to the peptide-HLA complex, it can set off a signalling cascade which leads to the essential help for B lymphocytes to make antibodies against the peptide or protein in infection. This is a central process in the adaptive immune system, allowing the organism to recognise foreign pathogens, bacteria, fungi, parasites and viruses. Autoimmunity occurs when the peptide carried to the surface by class II molecules is an unharmful 'self' peptide, which, if responded to by $CD4^+$ T cells, can result in creation of autoantibodies and lead to islet beta-cell destruction and T1D diagnosis. In T1D, there are four well-established autoantibodies that are often observed at around the time of diagnosis: preproinsulin (PPI), insulinoma-associated antigen 2 (IA2), glutamic acid decarboxylase (GAD) and Zinc transporter 8 (ZnT8) [71, 105, 206]. There are a number of disease susceptible or protective class II haplotypes. The combination of haplotypes with the largest effect on T1D risk is the combination of DR3-DQ2 on one chromosome and DR4-DQ8 on the other chromosome, which is a diplotype often referred to as 'DR3/4' [141].

HLA class I molecules, by contrast, are expressed in almost every cell in the human body. They bind to peptides present in the cytoplasm and transport them to the cell surface, upon which class I molecules are bound to and the cell is destroyed by cytotoxic $CD8^+$ T cells. In most instances this results in protection from unwanted or harmful peptides,

but also could result in autoimmunity if the presented peptide is a 'self' peptide. There are strong associations with T1D in class I region also, with classical alleles within the HLA-A and HLA-B genes showing association with disease after accounting for the effects of the class II haplotypes [135, 83]. For example the HLA-A*02:01 allele is associated with increased disease risk and also has been shown to bind to part of the PPI molecule [177].

1.4.2 Regions of association outside the human leukocyte antigen

Prior to the development of GWAS chips, four additional chromosome regions and the assumed candidate causal genes had been associated with T1D, insulin (INS) [194], Protein Tyrosine Phosphatase Non-Receptor Type 22 (PTPN22) [24, 178], Cytotoxic T-Lymphocyte Associated Protein 4 (CTLA4) [140] and Interleukin-2 Receptor alpha (IL2RA) [197]. These regions were found using the candidate gene approach, in which particular genes are chosen because of their known function in the immune system (or in the case of INS, diabetes). The association near *INS* tags a variable number tandem repeat that alters the expression of INS mRNA in the thymus, where the T1D risk allele is associated with reduced mRNA levels and therefore potentially a subsequent lack of negative selection of insulin self-reactive T cells in the thymus [194]. At *PTPN22*, which is associated with many autoimmune diseases, the increased risk is likely as a result of changing T cell receptor signalling thresholds [25, 6], potentially allowing self-reactive T cells to avoid negative selection in the thymus. To date, the *CTLA4* region association is not fully resolved with regards to how it impacts T1D risk, but it may be to do with a decreased suppressive capacity of T regulatory cells with the risk variant(s) [115]. *IL2RA* forms part of the interleukin-2 receptor complex. It is expressed on the surface of most T cells and particularly highly expressed on regulatory T cells and therefore this gene could increase disease risk via a number of mechanisms in a variety of cell types.

Once GWAS were developed, the number of T1D-associated genetic regions increased from these four to approximately 60 [48, 15, 30, 190, 144]. The associations near *PTPN22*

and *INS* remain the variants outside the HLA with the largest effect sizes, but other regions were identified with large effect sizes, for example near the Interferon Induced With Helicase C Domain 1 (IFIH1) and Tyrosine Kinase 2 (TYK2) genes. The majority of the genetic risk of T1D can be explained by the HLA alleles, *INS* and *PTPN22* polymorphisms, but with increasing sample sizes, increasing numbers of variants with smaller effect sizes on disease are likely to be identified. This could lead to novel therapeutic targets or pathways to consider in the understanding of T1D aetiology.

1.5 Thesis structure

I will structure the remainder of my thesis into three chapters, each relating to a research project, followed by an overall conclusion. Each project will contain an introduction and literature search, an explanation of the statistical methodology implemented and details of the analysis, a results section and a conclusion/discussion.

In chapter 2 I performed an analysis uncovering the genetic determinants of age-at-diagnosis of T1D, something that has never been examined in a GWAS fashion before.

In chapter 3, in a related analysis, I examined the known T1D association regions and identified those that are particularly associated with early-diagnosed T1D.

In chapter 4 I performed the largest genetic analysis of T1D to date, using $> 60,000$ individuals.

In chapter 5 I discuss my findings and place them in context of the current literature and state any further work I think is of greatest urgency in the objective of understanding T1D aetiology and pointing towards therapeutic targets that could be beneficial in years to come.

Chapter 2

Genetic analysis of age-at-diagnosis of type 1 diabetes

2.1 Introduction and literature search

The genetics of T1D have been examined in a number of studies, with over 55 regions altering susceptibility to the disease [48, 190, 15, 30, 144]. The region with the largest effect size is the HLA region, whilst the regions containing the Insulin (INS), Protein Tyrosine Phosphatase Non-Receptor Type 22 (PTPN22) and Interleukin-2 Receptor Alpha (IL-2RA) genes also have strong association with disease risk. Combining the HLA and non-HLA risk in a genetic risk score, leads to accurate prediction of who is at most risk of developing T1D [172].

However, an individual who is diagnosed with T1D early in life will, on average, have a faster disease progression than someone diagnosed later. This is measured by the patients ability to naturally produce insulin. The level of a marker called C-peptide, which joins the two chains of the proinsulin molecule, and is present in circulation as a byproduct of proinsulin converting to insulin, indicates the individual's endogenous insulin production. It has been shown that individuals diagnosed at a young age have lower circulating C-peptide levels and thus less endogenous insulin production at diagnosis [52, 121, 14, 54, 175]. On top of this, the slope of C-peptide decline following diagnosis is steeper in

individuals diagnosed at a young age compared to older [121, 14, 54], suggesting a faster destruction of pancreatic beta cells with early diagnosis. Slowing the decline of C-peptide and therefore the reliance on insulin injections is considered a key longstanding aim in the treatment of T1D.

In addition to the differences in endogenous insulin production between those diagnosed at a younger or older age, clinical differences have been observed with a decrease in life expectancy of around 15 years for individuals diagnosed at < 10 years compared to matched controls and an increased prevalence of cardiovascular disease [95, 157], although some smaller studies have reported lower prevalence of diabetic retinopathy with early diagnosis [184].

Given these biological and clinical differences between T1D individuals diagnosed at different ages, it might be expected that there are alleles that predispose to younger T1D age-at-diagnosis (AAD). Identifying the genes and pathways could point towards therapeutic targets to slow the decline of C-peptide, with a view to decreasing the probability of complications in T1D individuals, as some endogenous C-peptide production is associated with protection from complications [106].

It is already known that some T1D-associated alleles are also associated with younger AAD [195, 135, 83], specifically, the HLA class II haplotypes DR3-DQ2 and DR4-DQ8, their diplotype: DR3-DQ2/DR4-DQ8, as well as the HLA class I classical alleles A*24:02 and B*39:06. Outside the HLA, there has been some evidence of AAD effects near the Interleukin 2 (IL-2) and Interleukin 21 (IL-21) genes, specifically at SNP rs2069763 (C>A), which is now known to tag two functional haplotypes, containing SNPs rs6819058 (A>G) and rs67797421 ($C > T$) [31] and also near the renalase, FAD-dependent amine oxidase (RNLS) gene, specifically SNP rs10509540 (T>C) [84]. In addition, an analysis of C-peptide decline, adjusting for T1D AAD suggested both the HLA region and a region on chromosome 1, specifically variant rs559047 (T>A), near the Zona Pellucida Glycoprotein 4 (ZP4) gene, were associated with C-peptide decline [163].

However, previous studies examining the genetics of AAD of T1D have focussed on specific known T1D-risk variants. It has never been studied before in an ImmunoChip or

genome-wide fashion. Given the number of T1D cases with both ImmunoChip and AAD data available in the present dataset, it offered the perfect opportunity to study this question in depth for the first time.

2.2 Methods

2.2.1 Genotype data and collections analysed

ImmunoChip data had been collected for individuals who enrolled into various collections. I excluded all individuals without AAD data available, as well as all individuals without T1D, since this analysis focussed exclusively on T1D individuals and the genetic determinants of AAD of T1D. I also excluded all individuals with missing sex data, since sex was included as a covariate in all models as males had an older average AAD than females in this dataset (females median AAD = 8.32, IQR= $(5.00, 12.00)$, males median= 9.00, IQR= $(5.00, 13.00)$). This left a total of 15,645 individuals included in the analysis (Table 2.1).

Individual and Genotype Quality Control

Genotyping was performed in three batches, the Genetic Resource Investigating Diabetes (GRID) collection [190] was genotyped in one batch, the Type 1 Diabetes Genetics Consortium (T1DGC) [160] collection was genotyped in a second batch and the third batch contained individuals from the Northern Ireland GRID (NI GRID) collection, Finnish individuals from the Tyypin 1 Diabetekseen Sairastuneita Perheenjasenineen (IDDMGEN) and Tyypin 1 Diabeteksen Genetiikka (T1DGEN) collections [152] and UK-based individuals from the Warren collection [13]. Quality control was performed by genotype batch.

The samples were removed from each batch separately using the following criteria, following guidance from [5]:

i) Discrepancies in sex between named sex and expected homozygosity on the X chromosome for a female. The expected homozygosity rate on the X chromosome of males is 1, since they have just one copy of the X chromosome and around 0.8 for females, which is equal to 1 minus the typical expected rate of heterozygosity across the genome, roughly 0.2. If the homozygosity rate is too high for females or too low for males, this could imply either an incorrect sex labelling for that individual or a possible sample mix

up. Therefore samples labelled as female with greater than 0.93 homozygosity on the X chromosome and samples labelled as male with less than 0.93 homozygosity on the X chromosome were removed (Figure 2.1).

ii) Samples with $> 10\%$ missingness, since lots of missing values is an indication of poor and therefore unreliable genotyping for that sample (Figure 2.1).

iii) The expected heterozygosity rate across the genome is approximately 0.2, and large deviations from this expected rate could indicate poor genotyping for that particular sample or indicating inbreeding or admixture for that sample. Thus I removed all individuals with ImmunoChip wide heterozygosity rate of less than 0.18 or greater than 0.24 (Figure 2.1).

Figure 2.1: Per-sample quality control by genotype batch. Top left=GRID, Top right=T1DGC, Bottom=NI GRID, IDDMGEN, T1DGEN and Warren. Dots outside red dashed lines were removed from the analysis.

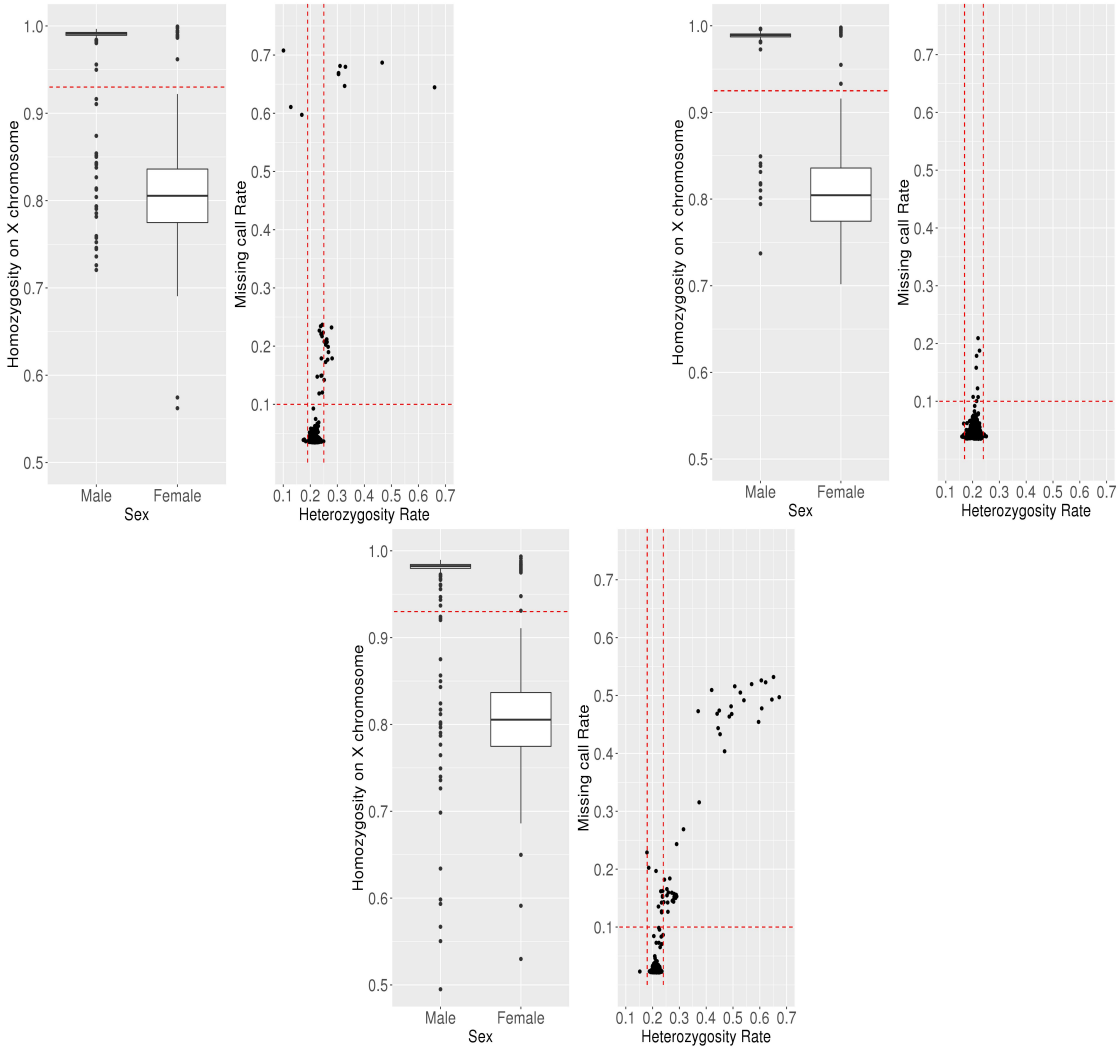


Table 2.1: Baseline characteristics of individuals included in AAD ImmunoChip analysis. Variants are number passing quality control. ASP=affected sibling pairs.

Cohort	Country	Type	Included	AAD: median (IQR)	Variants
GRID	UK	Independent	6728	8 (4, 11)	137401
IDDMGEN	Finland	Independent	1073	9.1 (5.5, 12.1)	137636
NI GRID	NI	Independent	509	7 (4, 10)	137636
T1DGEN	Finland	Independent	249	16 (10.4, 24.5)	137636
T1DGC	USA	ASP	2530	9 (4, 13)	138548
T1DGC	Europe	ASP	2485	11 (6, 17)	138548
T1DGC	Asia-Pacific	ASP	890	9.5 (5, 14)	138548
Warren	UK	ASP	839	10 (5.3, 15)	137636
T1DGC	UK	ASP	342	8 (4, 11)	138548
All	All	All	15645	9 (5, 12)	131357

After performing sample QC, I removed specific genetic variants which were genotyped with less certainty or lower quality. I first removed variants with $> 5\%$ missing values across all individuals in a batch and secondly variants with a low MAF (< 0.01), since spurious associations are easy to detect if very rare variants are included in the analysis as the approximations used in the models examining association of these variants are not designed for low cell counts. I did not remove variants that were out of HWE, since the population included in this analysis are T1D cases and some disease-associated variants will be truly inconsistent with HWE, rather than inconsistent due to genotyping error. However, the signal plots were inspected for variants associated with AAD to ensure the genotyping looked of good quality for the associated regions.

After performing SNP QC by batch, the intersect of SNPs passing QC was taken forward for association testing.

2.2.2 Meta-analysis

To estimate the effect of each variant on T1D AAD, I assessed the association of each variant with T1D AAD in each collection separately, before combining results from each collection in a meta-analysis. In collections of unrelated individuals, linear regression models were fitted, and in collections of related individuals, linear mixed models were fitted to account for non-independence of observations, one of the underlying assumptions of the standard linear model. Since the dataset was a combination of individuals

from multiple ancestries, in each analysis I accounted for the population structure in the dataset by adjusting for Principal Components (PCs) generated from genotype data. I first present the rationale and methods for calculating the PCs, then explain how I fitted standard linear models for collections of unrelated individuals and linear mixed models for collections of related individuals.

Principal Components Analysis

Principal components analysis (PCA) is a method used to reduce the dimensionality of a matrix by generating a number of orthogonal vectors and corresponding scalers that explain a high proportion of the overall variance with a much smaller set of variables. To perform PCA on, for example, an $N \times M$ matrix of genotypes, $\mathbf{X}$ containing N individuals and M SNPs, first the variance-covariance matrix of $\mathbf{X}$ must be calculated, using a matrix, $\mathbf{Z}$, where each column in $\mathbf{Z}$, $\mathbf{z_j}$, is the genotype value, $\mathbf{x_j}$ minus the mean of the genotypes for that SNP, $\bar{x}_j$. By the definition of covariance ($\frac{\sum_{i=1}^{N}(x_i-\bar{x})(y_i-\bar{y})}{N}$), it is evident that $\frac{1}{M}\mathbf{ZZ^T}$ is the variance-covariance matrix of $\mathbf{X}$. Performing an eigendecomposition of $\frac{1}{M}\mathbf{ZZ^T}$ identifies a set of orthogonal eigenvectors, and corresponding eigenvalues that combine so $\frac{1}{M}\mathbf{ZZ^T}$ can be expressed in the following way:

$$\frac{1}{M}\mathbf{ZZ^T} = \mathbf{Q\Lambda Q^{-1}}, \tag{2.1}$$

where $\mathbf{Q}$ is the matrix of eigenvectors of $\frac{1}{M}\mathbf{ZZ^T}$ and $\mathbf{\Lambda}$ is a diagonal matrix with the eigenvalues of $\frac{1}{M}\mathbf{ZZ^T}$ on the diagonal and 0 elsewhere.

The largest PCs (as defined by the eigenvectors with the corresponding largest eigenvalues) in a genetic covariance matrix will extract information about population structure in the dataset. Population structure in the dataset comes about since alleles may differ in frequency between ancestry divergent populations. Adjusting for these PCs in association analyses by including them as covariates should account for the population structure without overfitting the data by using all genetic variants as predictors, since adjusting for a small number of PCs explains much of the variance in the genetic matrix.

However, prior to performing PCA on the genetic matrix, it is important to first remove

variants in high LD with each other. This is because some haplotypes will be very large, and if you inherit one allele on the haplotype, you will almost certainly inherit the entirety of the haplotype, which can comprise many variants. If two individuals shared alleles at, say, five large haplotypes, they could cluster together in the space of the eigenvalues corresponding to the largest few eigenvectors, as a result of sharing many alleles at these haplotypes. In reality, the clustering should consider just five variants which tag the haplotypes so that these five genetic signals do not overwhlemingly drive the clustering on the largest PCs. Therefore I pruned the genetic matrix, $\mathbf{X}$, so that variants with $r^2 > 0.2$ with each other were excluded (and also excluded the HLA region entirely due to the complex LD structure in the region) before performing PCA on it. The PLINK software was used to carry out this entire procedure [150].

In the collections of unrelated individuals, this standard PCA approach was used to generate (eigen)vectors to adjust for in the analysis examining genetic determinants of AAD of T1D.

Principal Components Analysis in related samples (PC-AiR)

Principal component analysis in related samples (PC-AiR) is a method for generating PCs from a genetic matrix when the individuals in the analysis could be related to each other [44]. Under these circumstances, standard PCA can be distorted as a result of having related individuals in the analysis, which will naturally cluster together. PC-AiR first identifies a subset of unrelated individuals in the dataset, $\mathbf{U}$, whilst still including individuals with diverse ancestry in the population. It does this using a two-step procedure, firstly, the relatedness between each pair of individuals, i and j in the LD pruned matrix, $\mathbf{X}$, is estimated so a genetic relationship matrix (GRM) is generated [125]:

$$\hat{\phi}_{ij} = \frac{N_{heti,hetj} - 2N_{homiA,homja}}{2N^i_{het}} + \frac{1}{2} - \frac{1}{4}\frac{N^i_{het} + N^j_{het}}{N^i_{het}}, \tag{2.2}$$

with $N_{heti,hetj}$ the number of times individual i and j are both heterozygous, $N_{homiA,homja}$ the number of times individual i and j are homozygous with the opposite allele to each other, and N^i_{het} (or N^j_{het}) are the number of times individuals i (or j) are heterozygous

across the M variants. Secondly, in order to estimate the ancestral divergence between individuals i and j, the following formula is used [44]:

$$\hat{\kappa}_{ij} = \frac{1}{2}\left(1 - \frac{\sum_{s=1}^{M}(g_{is} - g_{js})^2}{N_{het}^{i} + N_{het}^{j}}\right), \tag{2.3}$$

where M is the total number of pruned variants in $\mathbf{X}$, s indexes the variant under consideration and g is the genotype. A negative value of $\hat{\kappa}_{ij}$ indicates that two individuals are from ancestrally divergent populations.

To calculate the unrelated set of individuals, the GRM is used to infer related individuals, where I chose a parameter threshold of 0.022 so that anyone related to $\hat{\phi}_{ij} > 0.022$ was considered related, but then of the unrelated individuals, those individuals with a higher overall ancestral diversity are prioritised for inclusion in $\mathbf{U}$. This prioritisation is performed for each individual, i, by summing the number of people they are ancestrally divergent from. Those individuals with a $\hat{\kappa}_{ij}$ value of < -0.022 (i.e. more likely ancestrally divergent) with more individuals are prioritised for inclusion in $\mathbf{U}$ over those individuals with a $\hat{\kappa}_{ij}$ value of < -0.022 with fewer individuals.

Standard PCA is carried out using $\mathbf{U}$, where the eigenvectors $\mathbf{Q}$ and eigenvalues $\mathbf{\Lambda}$ are calculated from the covariance matrix of $\mathbf{U}$, which I will call $\frac{1}{M}\mathbf{Z}\mathbf{Z}^{\mathbf{T}}$, to be consistent with Equation 2.1.

The eigenvectors are then 'projected' onto the related individuals in the following way: a weighting matrix (also sometimes referred to a 'SNP loadings'), $\mathbf{W}$, is generated by multiplying the mean-centred genetic matrix, $\mathbf{Z}$, by the eigenvectors, $\mathbf{Q}$, so $\mathbf{W} = \mathbf{Z}^{\mathbf{T}}\mathbf{Q}$. Rearrangement of Equation 2.1 gives:

$$\mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^{-\mathbf{1}} = \frac{1}{M}\mathbf{Z}\mathbf{Z}^{\mathbf{T}}, \tag{2.4}$$

$$\mathbf{Q} = \frac{1}{M}\mathbf{Z}\mathbf{Z}^{\mathbf{T}}\mathbf{Q}\mathbf{\Lambda}^{-1} = \frac{1}{M}\mathbf{Z}\mathbf{W}\mathbf{\Lambda}^{-1}. \tag{2.5}$$

Now taking the related individuals, $\mathbf{R}$, so that the individuals from $\mathbf{R}$ and $\mathbf{U}$ combine and together contain genotype data from all individuals in $\mathbf{X}$, the PCs for these individuals,

$\mathbf{Q_R}$, can be obtained using the mean-centred genotype data of $\mathbf{R}$, $\mathbf{Z_R}$ and the weighting matrix (or SNP loadings) and eigenvalues from the unrelated individuals:

$$\mathbf{Q_R} = \frac{1}{M}\mathbf{Z_R}\mathbf{W}\mathbf{\Lambda}^{-1} \tag{2.6}$$

The combined set of PCs for $\mathbf{X}$ is the matrix $\begin{bmatrix} \mathbf{Q} \\ \mathbf{Q_R} \end{bmatrix}$. In collections with related individuals, this method was employed to generate PCs. The five largest PCs were included as covariates in the linear regression models within each collection in an attempt to account for the population structure present in this dataset.

Standard linear models for collections of unrelated individuals

The gaussian likelihood function for a simple linear regression framework can be expressed in the following way for N obervations:

$$L(\boldsymbol{\beta}, \mathbf{X}, \sigma|\mathbf{y}) = (2\pi\sigma^2)^{-\frac{N}{2}} e^{-\frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^T(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})} \tag{2.7}$$

where $\mathbf{y}$ is the outcome, $\mathbf{X}$ is the matrix of predictors, including an $N \times 1$ vector $\mathbf{1}$ and $\boldsymbol{\beta}$ is a vector of fixed effects parameter estimates, the first element of which corresponds to an estimate of the intercept. This likelihood function assumes that errors (differences between predicted and true values) are normally distributed with constant variance. The likelihood is maximised by differentiating the log-likelihood:

$$log(L(\boldsymbol{\beta}, \mathbf{X}, \sigma|\mathbf{y})) = -\frac{N}{2}log(2\pi) - \frac{N}{2}log(\sigma^2) - \frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}), \tag{2.8}$$

with respect to each model parameter and setting the equations to 0. In this analysis, each observation is assumed independent. This leads to the following maximum likelihood estimates for $\boldsymbol{\beta}$:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X^T X})^{-1}(\mathbf{X^T y}). \tag{2.9}$$

The estimated variance around these estimates is:

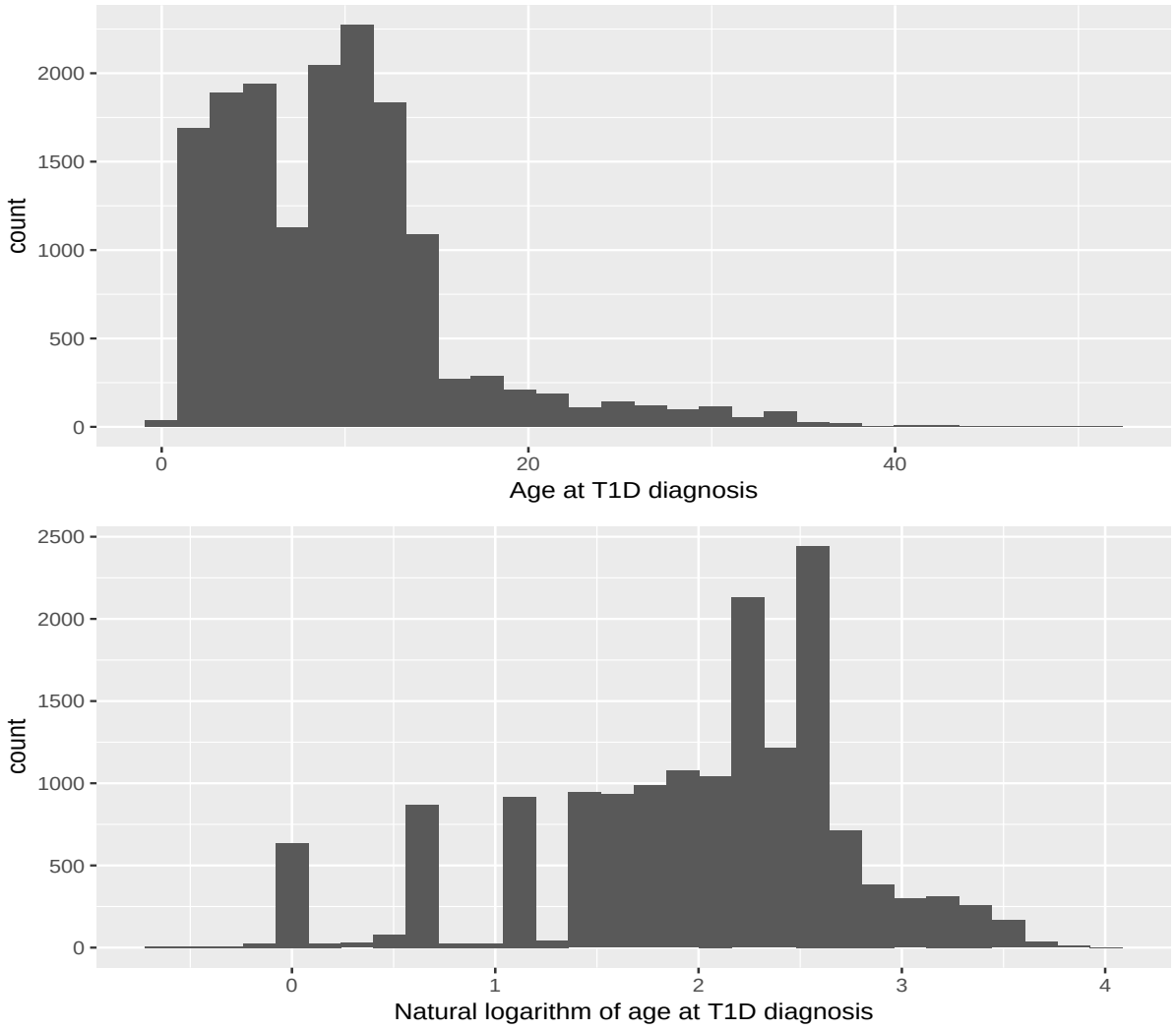
$$Var(\hat{\boldsymbol{\beta}}) = \hat{\sigma}^2(\mathbf{X^T X})^{-1}, \tag{2.10}$$

where $\hat{\sigma}^2$ is the estimated residual variance, calculated as $\frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{(N-p)}$, where $\hat{y}_i = \mathbf{x_i}\hat{\boldsymbol{\beta}}$, N is the number of observations and p is the number of covariates included in the model, including the intercept term (i.e. the number of elements of $\hat{\boldsymbol{\beta}}$).

In collections comprising unrelated individuals, simple linear regression was used, including the five largest PCs derived from genotype data as described above to account for any population structure so as to guard against reporting spurious associations. To be clear, for each SNP on autosomal chromosomes that passed QC on the ImmonoChip, a simple linear model was fitted with a p of 8, including an intercept, 5 PCs, sex and the genetic variant under consideration included as covariates.

The outcome variable in all instances was the natural logarithm of AAD ($log_e AAD$), meaning less emphasis would be placed on extreme values of AAD. Examining the distribution of AAD across all individuals in all collections, it can be seen that the small number of individuals included with older AAD could have large impacts on the results if the untransformed AAD was used (Figure 2.2). In addition, the residuals from fitting a linear model including sex as a covariate and including only the GRID collection (used as it is the largest ancestry-homogeneous and unrelated collection) are more consistent with the Normal distribution when the outcome is $log_e AAD$ as opposed to AAD, as assessed by the Shaprio-Wilks test (W statistic= 0.964 untransformed versus 0.894, log-transformed) [170].

Figure 2.2: AAD distribution across all individuals included in the analysis. Top panel: raw AAD values. Bottom panel: log_e transformed AAD values.



Linear mixed models for collections of related individuals

In many statistical problems, there is not independence between observations. For example, if individuals are siblings and one of them has T1D, the probability of the other being diagnosed with T1D is higher than an individual from the general population. Another possible way that observations could be related to each other is through individuals all being from one ancestral background. In some of our collections, individuals were affected sibling pairs (ASPs) and also from different ancestries, which needed to be accounted for. One way in which to do this is via a linear mixed model, fitted in the following way:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{b} + \boldsymbol{\epsilon}, \tag{2.11}$$

where $\mathbf{b}$ is a random effect, distributed as $\mathbf{b} \sim N(0, h^2\mathbf{R})$ and $\boldsymbol{\epsilon} \sim N(0, \sigma_e^2)$. The issue of non-independence between individuals is handled by the inclusion of $\mathbf{b}$ in the model, where the variance-covariance matrix of this distribution comprises the heritability, h^2 and the fixed relationship matrix, $\mathbf{R}$, which is estimated from genotype data as in Equation 2.2. The log-likelihood function becomes:

$$log\ L(\boldsymbol{\beta}, \mathbf{X}, \sigma, \mathbf{b}|\mathbf{y}) = -\frac{N}{2}log(2\pi) - \frac{1}{2}log|\boldsymbol{\Omega}| - \frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T\boldsymbol{\Omega}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}), \qquad (2.12)$$

with $\boldsymbol{\Omega} = \sigma^2(h^2\mathbf{R} + (1 - h^2)\mathbf{I})$ being the variance-covariance matrix of the random effect parameters, σ^2 the total trait variance and $|\boldsymbol{\Omega}|$ being the determinant of $\boldsymbol{\Omega}$.

To estimate the unknown parameters by maximising the log likelihood with respect to the unknowns, an iterative procedure is required, where the heritability parameter, h^2, is estimated and then the log likelihood is maximised with respect to the other parameters with the heritability held constant. The heritability parameter is then re-estimated by maximising the log likelihood with the values of the other parameters fixed at the levels from the previous iteration. This procedure continues until convergence, which can be a lengthy and computationally intensive process.

It would therefore be impractical to fit over 100,000 of these models (one for each variant on the ImmunoChip), each with sex and the top five genetic PCs and the genetic variant under consideration included as fixed effects. As a result, many methods have been developed to approximate the association of each variant by first fitting a model including all covariates except the genetic variant to obtain estimates of h^2 and the fixed effects, then using these estimates to develop association statistics for each genetic variant one at a time [98, 217, 99, 212]. EMMAX [99], an extension to EMMA [98] uses a score test to obtain the association statistics for each variant with the phenotype but has no means to efficiently produce effect size estimates and respective standard errors for each variant. GCTA [212] has the focus of accurately estimating the heritability parameter, h^2 and the association of each variant with phenotype is explored less fully. I therefore decided to use genABEL, GRAMMAR-gamma method, [185] to examine the effect of each variant

with T1D AAD in collections of related individuals.

The method is a three-step procedure. The first step estimates the unknown parameters from the maximised log-likelihood function (Equation 2.12) which includes any covariates adjusted for as fixed effects and also estimates of the heritability and residual variance, σ_e^2. The residuals from this model, $\widetilde{\mathbf{y}}*$ are then taken in step two:

$$\widetilde{\mathbf{y}}* = \hat{\sigma_e}^2 \hat{\mathbf{\Omega}}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}). \tag{2.13}$$

Since this model accounts for relatedness between individuals, the residuals are no longer correlated, and thus standard linear models can be fitted using the outcome variable as $\widetilde{\mathbf{y}}*$ and the effect of each SNP, i, on the phenotype can be tested using the following model:

$$\widetilde{\mathbf{y}}* = \alpha + \mathbf{g}_i\beta_i + \boldsymbol{\epsilon}. \tag{2.14}$$

A score test for the association of the SNP on the phenotype has been developed, where $T_i^2 = \frac{\beta_i^2}{var(\beta_i)}$. Here,

$$\beta_i = \frac{\widetilde{\mathbf{g_i}}^T \widetilde{\mathbf{y}}*}{\widetilde{\mathbf{g_i}}^T \widetilde{\mathbf{g_i}}}, \tag{2.15}$$

with $\widetilde{\mathbf{g_i}} = \mathbf{g_i} - E(\mathbf{g_i})$, with $\mathbf{g_i}$ the genotype vector at variant i, and

$$var(\beta_i) = \frac{\widetilde{\mathbf{y}}*^T \widetilde{\mathbf{y}}*}{N(\widetilde{\mathbf{g_i}}^T \widetilde{\mathbf{g_i}})}. \tag{2.16}$$

So the score test for the GRAMMAR test [4, 10], which concludes steps one and two, is:

$$T_i^2 = \frac{(\widetilde{\mathbf{g_i}}^T \widetilde{\mathbf{y}}*)^2}{(\widetilde{\mathbf{g_i}}^T \widetilde{\mathbf{g_i}})^2} \Big/ \frac{\widetilde{\mathbf{y}}*^T \widetilde{\mathbf{y}}*}{N(\widetilde{\mathbf{g_i}}^T \widetilde{\mathbf{g_i}})} = \frac{(\widetilde{\mathbf{g_i}}^T \widetilde{\mathbf{y}}*)^2}{(\widetilde{\mathbf{g_i}}^T \widetilde{\mathbf{g_i}})^2} \times \frac{N(\widetilde{\mathbf{g_i}}^T \widetilde{\mathbf{g_i}})}{\widetilde{\mathbf{y}}*^T \widetilde{\mathbf{y}}*} = \frac{N(\widetilde{\mathbf{g_i}}^T \widetilde{\mathbf{y}}*)^2}{(\widetilde{\mathbf{g_i}}^T \widetilde{\mathbf{g_i}})(\widetilde{\mathbf{y}}*^T \widetilde{\mathbf{y}}*)}. \tag{2.17}$$

However, it was demonstrated that the association statistics generated from these formula were conservative, and so step three of the analysis corrects the score statistic in order to maximise power to detect associated SNPs.

In [185], this correction factor is derived and the association effect size and variance are:

$$\beta_{new} = \gamma\beta \tag{2.18}$$

and

$$var(\beta_{new}) = \gamma var(\beta), \tag{2.19}$$

where γ is the correction factor:

$$\gamma = \frac{1}{\hat{\sigma}^2\hat{h}^2}(1 - \frac{1-\hat{h}^2}{N-1}trace(\hat{\sigma}^2\hat{\mathbf{\Omega}}^{-1})). \tag{2.20}$$

It is these estimates of the effect sizes and standard errors that were generated using collections of related individuals and combined in the meta-analysis with independent case-control collections, as presented below.

Inverse-variance weighted meta-analysis

Once association testing of each variant with T1D AAD was performed, using linear regression in collections of unrelated individuals and the GRAMMAR-gamma linear mixed models in collections of related individuals, each variant had an estimate and a standard error around that estimate for it's effect on $log_e AAD$. The standard error of the estimate is a function of the number of individuals in the analysis, and thus a larger study is expected to have, on average, smaller standard errors around the effect estimates. It is appropriate that larger studies with more precise estimates are given more weight in the overall meta-analysis compared to smaller studies with less precise estimates. An effective way to do this is to weight the effects by the inverse of the squared standard error (i.e. variance) and average the effect across collections in this weighted fashion, as performed in, for example [208]. If there are C collections in the meta-analysis, the weighted average of the effect size across all collections, β, can be obtained using the following formula:

$$\beta = \frac{\Sigma_{i=1}^{C}\beta_i w_i}{\Sigma_{i=1}^{C} w_i}, \tag{2.21}$$

where $w_i = 1/se_i^2$, se_i=standard error around the β_i estimate from collection i. The standard error around this overall estimate, se, is:

$$se = \sqrt{\frac{1}{\Sigma_{i=1}^{C} w_i}}. \tag{2.22}$$

Therefore, results from each collection can be combined and overall significance of a variant can be tested using a Wald test, dividing the overall β estimate by it's standard error to obtain an overall z-score which should be normally distributed under the null hypothesis. Comparisons of the estimated overall z-score to the normal distribution gives a 2-sided probability of observing the z-score if the null hypothesis were true.

False discovery rate

Owing to the fact that in GWAS analyses many variants are tested for their association with the phenotype, the genome-wide significance threshold of 5×10^{-8} has been widely adopted to declare a locus associated with disease to adjust for the multiple comparisons. This was initially based on a Bonferroni correction at 5% significance, whereby the significance threshold is divided by the number of tests carried out, forming the Bonferroni significance threshold. Any tests associated with a p value less than the Bonferroni significance threshold can therefore be declared associated and overcome the multiple testing burden in a stringent way. However, there are fewer variants on the ImmunoChip than most GWAS platforms, and it would make sense to take the total number of variants examined in the analysis and divide 0.05 by this number to obtain an ImmunoChip Bonferroni significance threshold. In the meta-analysis, there were 130999 variants with association results available, so I also examined any variants that were associated to $p < 0.05/130999 (= 3.82 \times 10^{-7})$.

In addition, since the Bonferroni correction is conservative, I also examined regions that were associated to a False Discovery Rate (FDR) of < 0.05. The FDR was first introduced in [19] by Benjamini and Hochberg, and is often referred to as the Benjamini-Hochberg FDR (BH-FDR). The premise is that if you perform many hypothesis tests, some will be truly associated and others will be false-positive associations due to random chance. If

you perform M tests and identify R tests with a p value < 0.05, and therefore reject the null hypothesis, R will consist of T true associations and F false-positives, so $T + F = R$. In reality, T and F are unknown, but the expected proportion of false positives can be controlled by setting $E(\frac{F}{R}) = t$, where t is the researcher-defined threshold at which they are comfortable that this proportion of associated tests are false positives. Commonly used thresholds are 0.01, 0.05 or 0.1, indicating the FDR-associated results will contain on average 1%, 5% or 10% false positives. In practise, this is done by ordering the association p values, $i = 1, ...M$, from smallest to largest, then taking all those tests with $p_i < \frac{i \times t}{M}$ as FDR associated to $< t$. I chose a threshold, t, of 5% as I felt this was suitably conservative.

Another complicating factor is that the BH-FDR procedure assumes each test is independent, which is not the case in GWAS analyses, since there is LD between variants and therefore test statistics will be correlated. An extension to BH-FDR was proposed by Benjamini and Yekutieli [20] (BY-FDR) that accounts for the fact that there is correlation between test statistics. It does this by dividing the threshold, t, by the sum of number of variants already tested for in the procedure, so for test i, the threshold the p value needs to be smaller than to be declared associated under BY-FDR is $\frac{i}{M} \times \frac{t}{\Sigma_i i}$. For example, if 100 variants are tested, the first variant would be FDR associated at the 5% level if the association p value was $< \frac{1}{100} \times \frac{0.05}{1} = 0.0005$, the second would be declared associated if the p value was $< \frac{2}{100} \times \frac{0.05}{1+2} = 0.0003$, and so on.

Finally, including the HLA region, many variants of which were strongly AAD-associated, in the FDR calculation, lowers the threshold for which to declare statistics FDR-associated. Given variants within the HLA are undoubtedly associated with AAD, I removed the entirety of the region before performing a BY-FDR calculation on the remaining variants. Any variants associated to BY-FDR< 0.05 were declared AAD-associated.

2.2.3 Residual-based method

An alternative model was also used as a sensitivity analysis to the primary meta-analysis described above. The approach was simpler and complementary to the meta-analysis.

Instead of stratifying the dataset into collections, I combined all individuals and fitted a linear mixed model with three random intercept parameters, one for the family identifier, one for the country the individuals were enrolled in and one for the collection the individuals were enrolled into. So the following model was fitted, with $\mathbf{X}$ being a $N \times 2$ matrix with a $N \times 1$ column vector of 1's for the intercept and a sex identifier, $\boldsymbol{\beta}$ a 2×1 vector of fixed effect sizes for the intercept and sex effect, $\mathbf{Z}$, an $N \times 3$ matrix containing data on the family ID, country and collection the individuals are enrolled into and $\mathbf{U}$ a 3×1 vector of random effects, where $u_i \sim N(0, \sigma_i^2)$ for each random effect i=1,2,3:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{U} + \boldsymbol{\epsilon}. \tag{2.23}$$

The idea was that by fitting this model, the residuals would be approximately independent and population structure accounted for by including the three random effects. I used the residuals as the outcome to assess the association of each variant on the ImmunoChip with these residuals.

Another reason for this alternative approach was that I wanted to attempt fine mapping at any region that showed association with AAD. One way to do this would be to use these approximately independent residuals as the outcome variable in fine mapping, maximising power since all individuals would be included with an attempt to adjust for any population structure and non-independence in the data.

2.2.4 Imputed GWAS analysis

To examine whether any regions not on the ImmunoChip showed association with AAD, I used GWAS data from two cohorts to perform a UK-ancestry analysis. A subset of 1768 cases from the GRID collection had been genotyped using the Affymetrix GeneChip Mapping 500K genotype platform and 3833 had been genotyped using the Illumina 550K Infinium microarray platform [15]. Variants not present on the genotype platforms were imputed and the association of these variants with $log_e AAD$ was examined, adjusting for the three largest genetic PCs as covariates in the model. The results from the individuals genotyped in both batches were combined in an inverse-variance weighted fixed-effects

meta-analysis.

Genotype imputation

Since these platforms do not genotype all the same variants (though there is much overlap), it was appropriate to impute the missing variants along the genome stratified by genotype platform. IMPUTE2 [82] [81] was used to impute the missing genotypes genome-wide, using the 1000 Genomes Project reference panel (Phase 3) as reference haploypes.

Genotype imputation can be performed using a Hidden Markov Model (HMM), where haplotypes from individuals in the analysis are examined for consistency with haplotypes in a reference panel. The reference panel of haplotypes can be considered hidden states and a section of an individuals DNA can be matched to the haplotype from a given state if the other variants on the DNA strand are consistent with those in the hidden state haplotype. The assumed state for the individual can then transition from one hidden state to another at recombination hotspots, where the individuals genotypes might better match to a different reference haplotype. The imputation was performed in 1Mb chunks across the genome.

Dr. Nicholas Cooper carried out the imputation [49].

2.2.5 Fine mapping

Fine mapping is a branch of methodology that attempts to uncover the variants that are most strongly associated with the phenotype from a set of variants, usually in LD, that show association with the phenotype and therefore are more likely to be causal. Occasionally, it is not the most associated variant in the region that is causal, either due to random statistical fluctuation, or due to a variant being in LD with two variants that are both associated with the phenotype [202, 8]. Even when the most associated variant is causal, fine mapping is useful as it can characterise how strongly to prioritise the variant over other SNPs in LD with the index variant.

A number of methods have been developed that assume a single causal variant in each

region [132, 124], but my focus will be on methods that can account for more than one causal variant in a region; this is when fine mapping can be of most use elucidating multiple signals, as if there is one signal, marginal summary statistics can provide a good estimate of the most likely causal variants. Amongst methods that can identify more than one causal variant are PAINTOR [104], CAVIAR [80], CAVIARBF [41] and FINEMAP [21]. These methods vary slightly in their implementation, for example PAINTOR can include empirical priors to increase probability of variants that sit in functional regions of the genome (in the appropriate cell type). However, they all work under the same premise, which is they take as input a set of marginal summary statistics and also an LD matrix from the ancestry population the individuals from the GWAS are from. The association z-scores ($\frac{\hat{\beta}}{\hat{\sigma}_\beta}$) are assumed to follow a multivariate normal distribution, with the covariance matrix being the LD matrix in the population, $\mathbf{\Sigma}$:

$$\mathbf{z} \sim N(\boldsymbol{z}, \mathbf{\Sigma}). \tag{2.24}$$

Then the posterior probability of causality is calculated for each possible combination of variants included in the model $\mathbf{c}*$, where $\mathbf{c}*$ is a $1 \times M$ vector of 0s and 1s indicating whether the variant is included in the model (1) or not (0). Each combination $\mathbf{c}*$, has a prior, $p(c*)$ and all possible models are denoted C. The posterior for each combination is thus calculated as:

$$p(\mathbf{c}*|\hat{\mathbf{z}}) = \frac{p(\hat{\mathbf{z}}|\mathbf{c}*)p(\mathbf{c}*)}{\Sigma_{c \in C} p(\hat{\mathbf{z}}|\mathbf{c})p(\mathbf{c})}, \tag{2.25}$$

where $\hat{\mathbf{z}}$ is the estimated z-score for each variant. The reason I chose to use GUESSFM [202] over the other available software was that raw genotype data was available as opposed to association summary statistics alone, and GUESSFM harnesses this raw data availability in a way that PAINTOR, CAVIARBF and FINEMAP do not. I will outline the GUESSFM algorithm below.

I applied the GUESSFM fine-mapping algorithm in an attempt to identify the most likely causal variants associated with AAD in those regions associated with AAD in

the variant-discovery analysis ($p < 5 \times 10^{-8}$). In the primary fine-mapping analysis, I made use of all available data by including all individuals in the analysis and using the residuals from the residual-based method described above as the outcome. However, as a sensitivity analysis, I repeated the analysis using only individuals from the GRID cohort. These individuals are from a more ancestry-homogeneous population and therefore the fine-mapping algorithm is less likely to identify spurious associations due to population structure alone. The outcome in the sensitivity analysis was $log_e AAD$ rather than the residuals as in the primary analysis, since $log_e AAD$ in the GRID cohort are independent between individuals and there is relatively little population structure in the dataset.

Overview of GUESS

The overall goal of GUESSFM is to identify the model(s), M, with the highest posterior probability of association with the phenotype, if there are T models to choose from $(t = 1, ...T)$, there is:

$$p(M_t|\mathbf{y}) = \frac{p(\mathbf{y}|M_t)p(M_t)}{\sum_{t=1}^{T} p(\mathbf{y}|M_t)p(M_t)}, \tag{2.26}$$

where:

$$p(\mathbf{y}|M_t) = \int p(\mathbf{y}|\theta_t, M_t)p(\theta_t|M_t)d\theta_t \tag{2.27}$$

is the marginal likelihood for the model M_t [42], integrated over the parameters in the model, θ_t to obtain the marginal likelihood.

GUESSFM works by implementing the GUESS algorithm [26], but first pruning the variants in the $N \times P$ genetic matrix, $\mathbf{X}$ so that there are no variants that perfectly predict each other in the dataset. It does this because in the downstream calculation of the marginal likelihood of the model space, one of the matrix calculations becomes unstable when the columns of the matrix perfectly predict each other. The LD value to prune the variants to is not fixed and can be altered based on, for example, the sample size in the analysis. I chose a value of $r^2 = 0.99$ to prune variants to, since this included a high proportion of variants in the GUESS procedure. Once the P variants have been

pruned, the matrix is an $N \times K$ genetic matrix, called $\mathbf{G}$.

The outcome, $\mathbf{y}$, is modelled as a linear combination of genetic predictors, $\boldsymbol{\beta}$, and an intercept term, α. To include every genetic variant in this model will lead to unstable and ungeneralisable results and the model will suffer from overfitting. Thus it is better to identify a smaller number of genetic variants that jointly explain the maximum variance in the outcome, $\mathbf{y}$. This is done with a Bayesian approach, with priors placed on the genetic variants which encourage a sparse model with few predictors. More formally:

$$\mathbf{y} = \alpha\mathbf{1} + \mathbf{G}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{2.28}$$

where $\boldsymbol{\epsilon}$ is the error term and $\boldsymbol{\epsilon} \sim N(0, \sigma^2)$. In order to choose which variables are selected to go into the model, the sparsity-inducing prior on $\boldsymbol{\beta}$ is made to depend on a $K \times 1$ latent binary vector, $\boldsymbol{\gamma}$, which is equal to 0 if the corresponding $\beta_k = 0$ and 1 if the corresponding $\beta_k \neq 0$. This latent binary vector now replaces M_t from Equation 2.26, since every combination of 0's and 1's corresponds to a different model. Equation 2.28 can be re-expressed in terms of the latent binary vector:

$$\mathbf{y} = \alpha\mathbf{1} + \mathbf{G}_\gamma\boldsymbol{\beta}_{\boldsymbol{\gamma}} + \boldsymbol{\epsilon}. \tag{2.29}$$

This model has as many predictors as $\boldsymbol{\gamma}$ has value 1, so as the latent binary vector changes, both the predictors in the model and the number of predictors in the model vary. A normal prior is placed on $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$:

$$p(\boldsymbol{\beta}_{\boldsymbol{\gamma}}|\boldsymbol{\gamma}, \sigma^2) = N(0, \sigma^2\Sigma_\gamma), \tag{2.30}$$

with an inverse-gamma prior placed on the error variance σ^2:

$$p(\sigma^2|\boldsymbol{\gamma}) = InvGa(a_\sigma, b_\sigma), \tag{2.31}$$

where a_σ and b_σ are hyperparameters larger than 0. Σ_γ is the parameter for the covariance structure between the variables included in the model, $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$, and is estimated from the data

itself, so $\Sigma_\gamma = g(\mathbf{G}_\gamma^\mathbf{T}\mathbf{G}_\gamma)^{-1}$, with scaler g.

Another prior to be determined is the prior for the latent binary vector, $p(\boldsymbol{\gamma})$, which is taken to be a beta-binomial prior, with probability parameter, ω, which is sampled from the Beta distribution with hyperparameters a_ω and b_ω, which are chosen to reflect the expected number of variants included in the model. In the primary analyses in this chapter, the expected number of causal variants in each region was taken to be 3, since there are many T1D-associated regions with more than one genetic association signal.

The final prior to be decided is that for the scaler which determines the scale of the variance and covariance of $\boldsymbol{\beta}_\boldsymbol{\gamma}$, g. Since most genetic effect sizes are relatively small, the variance of the $\boldsymbol{\beta}_\boldsymbol{\gamma}$ parameters is expected to be small and therefore an inverse-gamma prior on g with a shape parameter of 0.5 and a scale parameter of $\frac{N}{2}$ is appropriate since the majority of the time the variance will be less than one but on occasion the variance will be very large, covering all likely values of $\boldsymbol{\beta}_\boldsymbol{\gamma}$.

An MCMC scheme is used to explore the model space with this configuration of prior set up which has a number of moves that attempts to avoid becoming trapped in local maxima. The marginal distribution of interest in the MCMC scheme is that of $\boldsymbol{\gamma}$, but this depends on the scaler that determines the scale of the variance and covariance of $\boldsymbol{\beta}_\boldsymbol{\gamma}$, g, and the distribution of $\boldsymbol{\gamma}$ must therefore also be averaged over values of g. As a result of this, in the MCMC scheme, first a value for g is sampled, then the latent binary vector $\boldsymbol{\gamma}$ is updated, conditional on g, followed by an updating of g, and repeating the process examining $\boldsymbol{\gamma}$, and so on until convergence is reached.

The idea behind the evolutionary stochastic search that GUESS implements is that it is easy for standard MCMC methods to become 'stuck' in local maxima of the posterior density when examining a model selection problem with many possible predictors, and not explore the majority of the posterior density due to acceptance of the current $\boldsymbol{\gamma}$ value when in a local maxima the majority of the time. Therefore the method instead has more than one 'chain' being explored at any given time. These chains have a 'temperature' attached to them, whereby the higher the chain number, the higher the temperature attached to the chain. The temperature attached to the chain has the impact of flattening

the likelihood and therefore increasing the probability of changes to $\boldsymbol{\gamma}$ being accepted for that chain. The bottom chain, chain 1 of L has no temperature attached to it (known as the unheated chain) and the value of $\boldsymbol{\gamma}$ in this chain is the one that contributes towards the marginal posterior distribution recorded for $\boldsymbol{\gamma}$. However, by running L chains with different temperatures attached to them and then using a probabilistic mechanism to swap or exchange $\boldsymbol{\gamma_l}$ between chains, it ensures that the unheated chain can inherit some or all of a heated chain value of $\boldsymbol{\gamma_l}$, and thus explore a higher proportion of the posterior density. In all analyses carried out in this thesis, I have run three chains per stochastic search ($L = 3$).

The algorithm works by, once a value for g has been sampled, performing a 'local' move with probability 0.5. Local moves are defined as those where the variables selected potentially change within the same chain. If the local move is accepted, it is applied to all chains in the search. The local move implemented is called Fast Scan Metropolis-Hastings (FSMH), which is used because if you have many predictors in a model space and a small number of those predictors are included in the current model, some search methods would attempt to swap each variant and spend the majority of time attempting to change elements of $\boldsymbol{\gamma}$ (on any given chain) from 0 to 1, since most values will be 0. To take every variable in the model space and propose a new value, which is accepted based on a probability would take a long time and usually result in a 0 being resampled to a 0. Therefore FSMH is used, which randomly samples a subset of variants to propose a change (either 0 to 1 or 1 to 0), then accepts the change if the posterior probability is higher from the new proposal, and accepts the change with probability equal to the ratio of new posterior probability to old posterior probability if the posterior probability of the new proposal is lower than the posterior probability in the old value. The sampled variants to propose changing depends on the current value of the selected variant (more likely to propose changing those currently with $\gamma_j = 1$) and the current model size (larger model size would lead to more variants proposed).

If the local move is rejected, which would happen with probability 0.5, a global move is attempted, where global is defined as any exchange of elements of $\boldsymbol{\gamma}$ across chains. There

are two types of global moves, the crossover operator or the exchange operator.

The crossover operator swaps elements of a chain with another. Two chains in the search are sampled, where those chains with higher posterior probability are more likely to be selected than chains with lower posterior probability at that point in the algorithm. There are two types of exchange operator that could then be performed between these two sampled chains, the first is a k-points crossover, which chooses a random break point within the variants and swaps the values of all the variants prior to the break point between the two chains. The second type of crossover is a block crossover, where a variant is selected and then the value of all variants in high LD with this variant are swapped between chains. The new chain is accepted based on a function that takes into account the temperature of each chain and the ratio of posterior probability after the exchange to the posterior probability before the exchange. If this value is greater than one, the move is accepted, if it is less than one, it is accepted with probability equal to the value of the function.

The other type of global move considered is the exchange operator, which is a complete swapping of $\boldsymbol{\gamma}$ values between two chains. In order to attempt to swap chains that could potentially have very different temperatures attached to them, a process called delayed rejection is used to select which chains to try and swap. Firstly, two chains are sampled randomly. If the switch of these two chains is rejected based on the same criteria as the crossover exchange, the first chain is again taken and a second proposal is put forward to exchange it with the chain immediately above it in the temperature ladder (e.g. chain 2 will be attempted to swap with chain 3), this gives two chances for the chain to be exchanged, firstly with a chain that is potentially distal from the chain under consideration with regards to temperature, the second with a chain that is similar in temperature to the current chain.

Overview of GUESSFM

GUESSFM is an extension to GUESS, where the models with the highest marginal posterior probability are saved and all variants that were filtered from the genetic matrix $\mathbf{X}$

that are in LD with a variant which appears in a model or models with high posterior probability are reintroduced. The set of models with high marginal posterior probability are then expanded to include all possible combinations of the current most probable models (which include only tag variants) and also models where the tag variants are replaced by variants in high LD with the tag variant in the models, since the variants in high LD were dropped arbitrarily prior to the stochastic search, leaving a total of M models.

Frequentist linear regression models are then fitted for each of the M models present in this expanded set of high posterior probability models from the stochastic search and their equivalents with variants in LD with the tags swapped in. Approximate Bayes Factors (ABFs) for each model are calculated by first obtaining the Bayesian Information Criterion (BIC) for each model under consideration as well as the null model with no genetic variants included in the model, then $ABF = e^{\frac{-BIC}{2}}$.

Finally, from these ABFs, the posterior probability of each model can be obtained using the following formula:

$$p(model_i) = \frac{ABF_i \times \pi_i}{\Pi_{i=1}^{M} ABF_i \times \pi_i}, \tag{2.32}$$

where π_i is the prior for model i. In reality, this is performed taking the log of all the above terms, making the calculation a summation rather than a product. The variant marginal posterior probability is then simply then sum of the posterior probability of each model that the variant appears in.

GUESSFM then selects the variant with the highest marginal posterior probability and orders the remaining variants by the LD r^2 to this lead variant. In an attempt to include all variants that could potentially be casually associated with disease, it then removes variants that have a joint marginal probability of inclusion of > 0.02 with the top variant (as these may constitute a second independent phenotype-associated signal) and steps away from the lead variant, keeping every variant and including it in the group until either the r^2 falls below 0.5, or the variant marginal posterior probability of inclusion falls below 0.001, for two variants in a row. This procedure is repeated until no variant is left with a posterior probability of > 0.01.

The top 10,000 most visited models were saved from the GUESS stochastic search for downstream analysis in the GUESSFM procedure.

Haplotype analyses

Once the most likely causal variants are identified by the fine-mapping algorithm, it is helpful to visualise the effect of each haplotype on the phenotype. Although I refer to these analyses as haplotype analyses, I do not consider non-additive effects as might be considered in classical haplotype analyses. I am instead attempting to show how combinations of alleles affect the phenotype, in what direction they are affecting the phenotype and showing the LD pattern between independent association signals. Only additive effects are considered by GUESSFM, which is used to define which variants to consider in the haplotype analyses, therefore the assumption of additive effects are the only ones considered in this complementary analysis. The difficulty in doing this is that individuals have genotype data available and not the haplotype data, so for example, if someone is heterozygous at SNP A, and heterozygous at SNP B, it is possible that their haplotype could be {01, 10} or {00, 11}, where $\{WX, YZ\}$ represents the haplotypes on the first (WX) and second (YZ) of the pair of chromosomes for positions A and B.
It is possible to estimate the haplotype frequency in the population by using an Expectation-Maximisation (EM) algorithm. This is necessary, since complete data (haplotype phase) is being inferred from a set of partially complete data (genotype data). The algorithm begins by choosing an initial guess for the frequencies of each haplotype, $\boldsymbol{\theta}^0$, then maximises the likelihood $L(\boldsymbol{\theta}^1|\mathbf{y}, \boldsymbol{\theta}^0)$, where $\mathbf{y}$ is the set of observed genotypes. The previous iteration $\boldsymbol{\theta}^{t-1}$ can be thought of as a prior to estimate the posterior, $\boldsymbol{\theta}^t$. Each individual will have a posterior distribution for each possible haplotype, which can be sampled from. With samples from individuals posterior probability of each haplotype, I fitted a linear mixed model with $log_e AAD$ as the outcome and included a fixed effect for each haplotype and sex, as well as the three random effects used in the residual-based method. I averaged over the effect size of ten estimates for each haplotype to make an estimate of the overall effect size of each haplotype on AAD risk. In haplotype estimation, I included index

variants from groups prioritised by the GUESSFM analysis. Each haplotype effect was estimated relative to the most common haplotype in the population.

I repeated the analysis but this time including only individuals from the GRID cohort in order to study a relatively ancestry-homogeneous population. I included variants prioritised from the primary GUESSFM analysis as generated using all individuals, to investigate whether or not the same haplotypes appeared to be impacting AAD in this more ancestry-homogeneous population as they appeared to be in the combined analysis of all individuals. When estimating the effect size of each haplotype, instead of using a linear mixed model, I fitted a linear model and included the five largest genetic PCs as covariates in the model as well as the haplotypes.

2.2.6 Disease-association meta-analysis at various age-at-diagnosis groups

In genetic regions that were found to be AAD-associated ($p < 5 \times 10^{-8}$), I wanted to explore how the T1D disease risk varied by AAD. I therefore stratified the dataset into individuals diagnosed at < 5 years, $\geq 5- < 10$ years, $\geq 10- < 15$ years, $\geq 15- < 20$ years and ≥ 20 years and examined disease association between the genetic variants of interest and T1D diagnosed at each AAD group. A collection of $19,022$ common controls from a number of collections (Table 2.2) were used in this analysis. Effects from each of the four collections were combined in an inverse-variance weighted fixed-effects meta-analysis, as described in section 2.2.2. The NI GRID collection was excluded due to small sample sizes in each AAD group. A brief explanation of how the independent and related collections were analysed is presented below.

Table 2.2: Baseline characteristics of individuals included in T1D risk analysis examining AAD-associated variants on T1D risk by AAD group.

Cohort	N controls	N cases	Control type
T1DGC	6282	6275	Parents/siblings of ASPs
IDDMGEN	2389	1082	Parents/siblings of cases
Warren	735	844	Parents/siblings of ASPs
GRID/1958BC/UKBS	9616	6734	Independent cases and controls
All	19022	15184	Various

Logistic regression for case-control collections

Logistic regression models are explained in more depth in chapter 3, since they are widely used in that analysis. However, they were also used in this analysis for collections of independent cases and controls (Table 2.2) to examine the association of each variant with T1D risk at each AAD group. The logistic regression models included the five largest PCs as calculated from genotype data within each collection (using standard PCA in collections of unrelated cases and controls and PC-AiR in collections of related individuals) in an attempt to account for any population structure present within each collection.

Generalised linear mixed model association test (GMMAT) for related collections

Logistic regression assumes independence of observations in the analysis. In collections where individuals are related, an alternative approach is required which accounts for the relatedness between individuals. A number of ways have been suggested to do this by assuming the outcome is linear and fitting a linear mixed model accounting for relatedness [98, 217, 99, 185, 212]. These methods assume a constant error variance across all means, for example if the outcome was height and the mean height was different between two categories, the variance around the mean would be assumed to be the same in linear regression. However, for binary traits, the variance is a function of the mean, $var(p) = p(1-p)$, thus the constant variance assumption is violated if the mean differs between ancestrally divergent groups. This is particularly important when there is population structure in the data, with potentially different prevalence of disease between ancestry group. Since there are individuals from multiple ancestries in the present dataset, it was inappropriate to make this normality assumption in the outcome.

The generalised linear mixed model association test (GMMAT) [39] fits a logistic mixed model. The method maximises a quasi-likelihood function with respect to the random intercept term first and then the fixed effect terms. The iterative process then updates the random intercept term to maximise the quasi-likelihood with the new fixed effect

parameters, and continues this procedure until convergence. The software was designed with the capability to test the association of each variant across the genome by first fitting the logistic mixed model without any genotype variables in order to obtain estimates for the fixed and random effect parameters; which are assumed constant from then onwards. Then a score test is used, which is a fast procedure involving one matrix multiplication, quickly generating association statistics for each variant.

However, this method only calculates association statistics against the null hypothesis of no association between variant and outcome and does not estimate effect sizes or standard errors around effect sizes for each variant. In this analysis, the related and independent cohorts needed to be combined in a meta-analysis. The preference was for this meta-analysis to be inverse-variance weighted and thus the effect size and standard error around the effect size for the variants of interest needed to be estimated. Therefore, instead of excluding the genetic variant from the initial logistic mixed model, I included the genetic variant and maximised the quasi-likelihood including this variant as well as the top five genetic PCs and a random effect incorporating relatedness between the individuals:

$$logit(\mathbf{p}) = \alpha + \mathbf{PC1}\beta_1 + \mathbf{PC2}\beta_2 + \mathbf{PC3}\beta_3 + \mathbf{PC4}\beta_4 + \mathbf{PC5}\beta_5 + \mathbf{g1}\beta_6 + \mathbf{b}, \quad (2.33)$$

where **g1** is the genotype for the variant under consideration and $\mathbf{b} \sim N(0, h^2\mathbf{R})$, with h^2 the heritability and $\mathbf{R}$ the GRM which is considered known and has been calculated from genotype data. Since this GRM needs to be positive semi-definite for the algorithm to be analytically tractable, a different calculation for individual relatedness was used in this analysis from in Equation 2.2, which provides robust estimates when population structure is present but the GRM generated is not necessarily positive-semi definite. Therefore, in this analysis, the following was used to estimate the GRM between individual i and j, with M variants in the pruned genetic matrix [9]:

$$\hat{\phi}_{ij} = \frac{1}{M}\sum_{m=1}^{M}\frac{(g_{m,i} - 2p_m)(g_{m,j} - 2p_m)}{2p_m(1-p_m)}. \quad (2.34)$$

Here p_m is the MAF at variant m.

To fit this model across all variants on the ImmunoChip would be very slow and computationally expensive, but since only a small number of variants appeared to be AAD-associated, this procedure could be performed for each variant and in each AAD-group in a reasonable time-frame.

To examine whether or not any association between SNP and T1D < 5 years was due to the threshold chosen to define 'early-diagnosed', as sensitivity analyses I varied the threshold and examined the association with both T1D diagnosed at < 4 years and < 6 years.

2.3 Results

2.3.1 Principal component analyses

The results from the PCA in independent cohorts and PC-AiR in collections of related individuals showed only one collection, the T1DGC had obvious population structure from studying the PC plots. Figure 2.3 shows the two largest PCs plotted against each other for the four collections of independent individuals and Figures 2.4 and 2.5 shows a scree plot and the two top PCs plotted against each other from the analysis using PC-AiR. It is clear there is population structure in the T1DGC collection (Figure 2.4) but very little in the Warren collection (Figure 2.5). The points are coloured by country of enrolment for the related individuals and it is clear that individuals enrolled from different regions cluster. Overall, adjusting for the five largest PCs seemed sufficient to account for the population structure in each collection.

Figure 2.3: Largest versus second largest genetic Principal Components for each collection of independent cases.

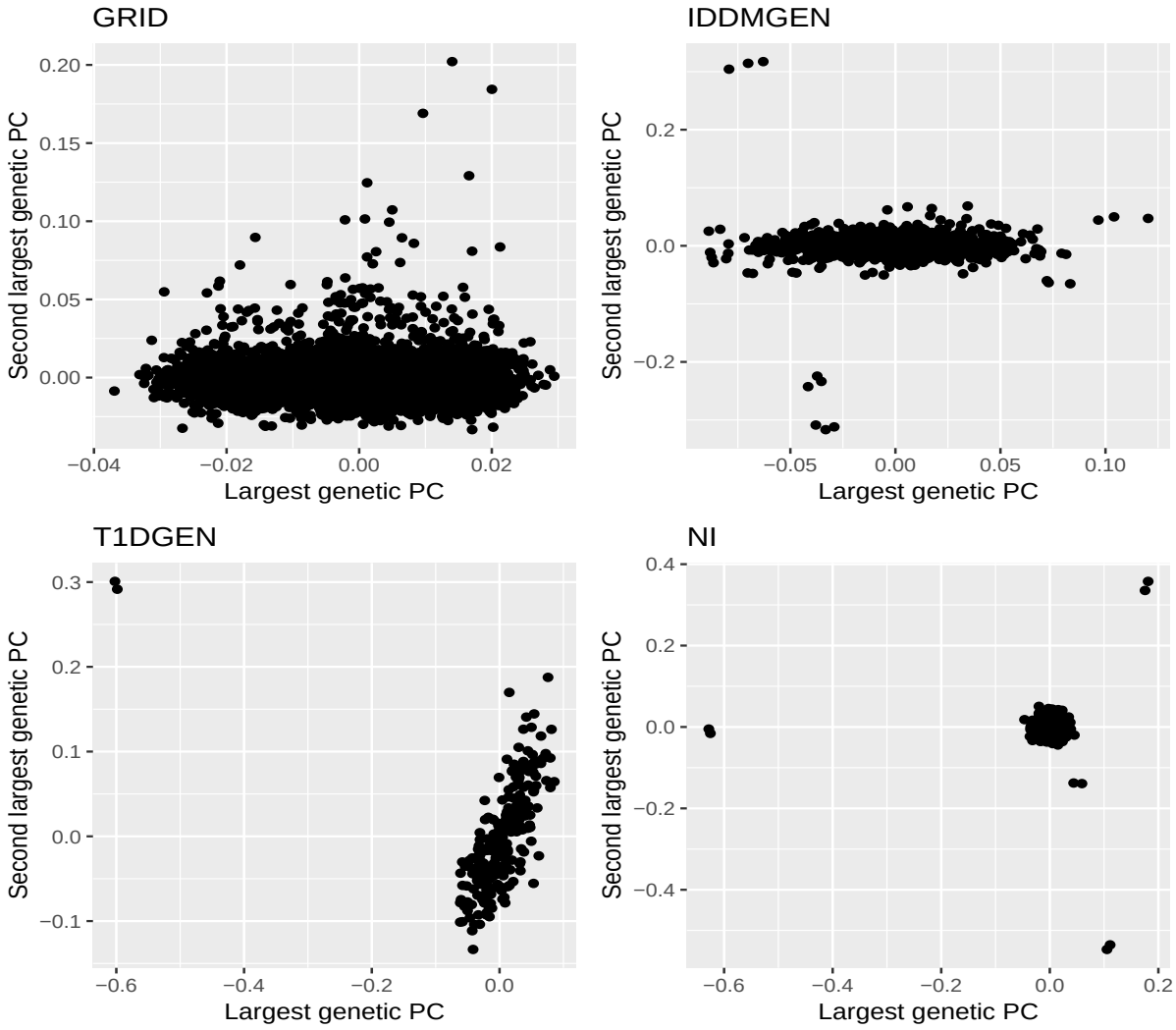


Figure 2.4: Top panel: scree plot for the T1DGC collection of related individuals showing proportion of variance explained by each eigenvector. Bottom panel: Largest versus second largest genetic PCs for the T1DGC collection, coloured by country of enrolment.

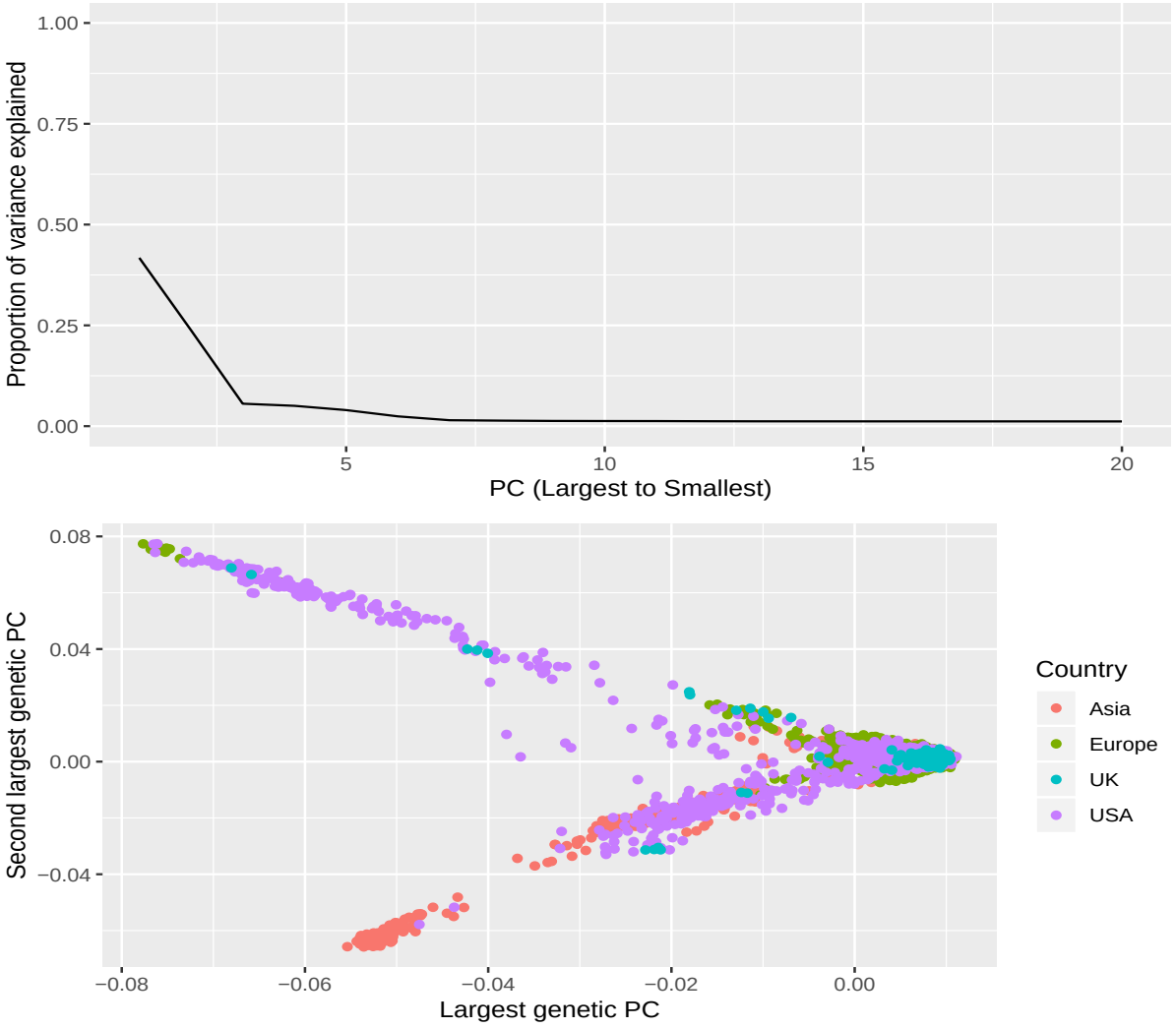
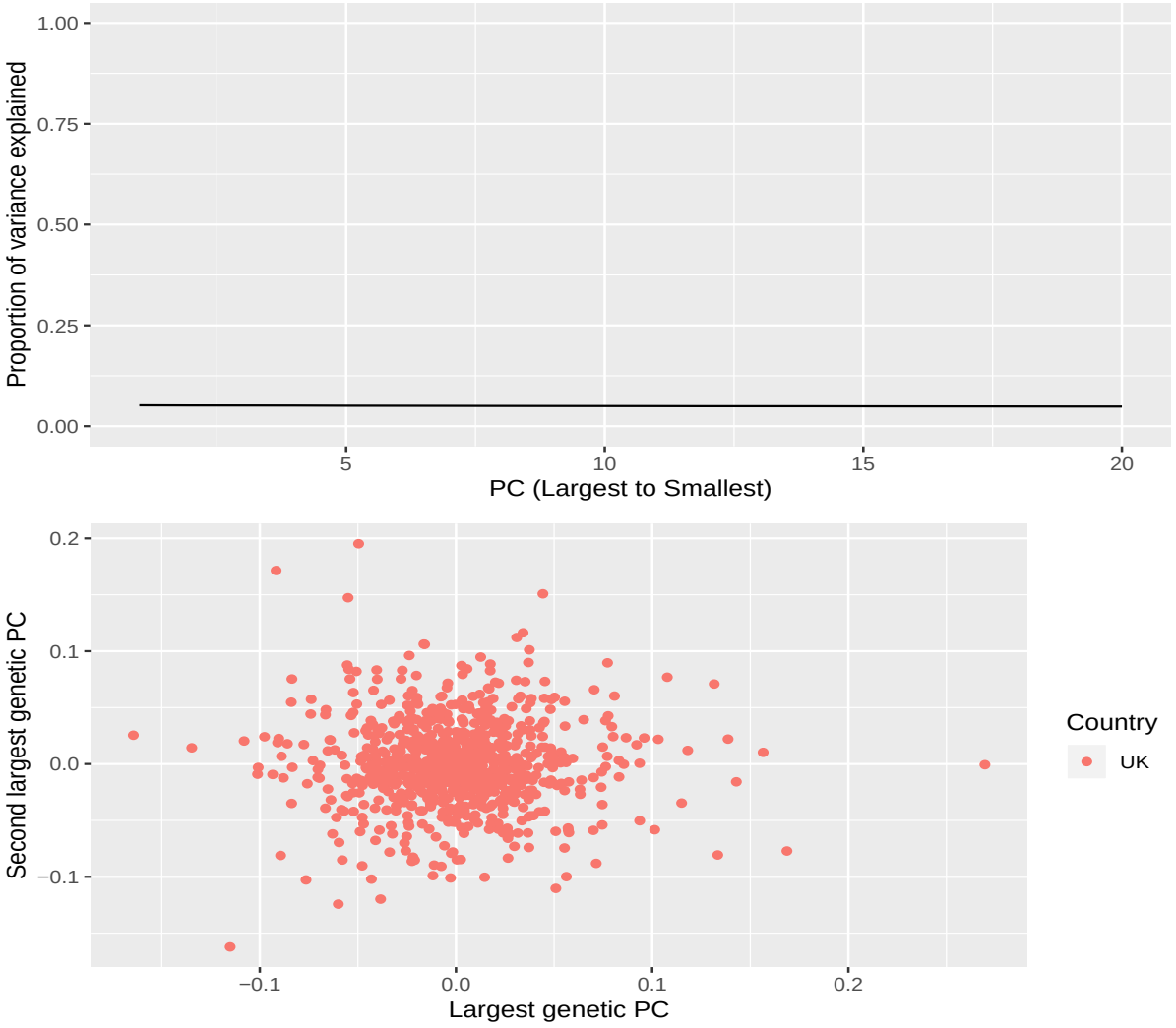


Figure 2.5: Top panel: scree plot for the Warren collection of related individuals showing proportion of variance explained by each eigenvector. Bottom panel: Largest versus second largest genetic PCs for the Warren collection, coloured by country of enrolment.



2.3.2 Discovery analysis: meta-analysis

Results from the inverse-variance weighted fixed-effects meta-analysis identified two regions associated to a genome-wide association threshold of $p < 5 \times 10^{-8}$, the HLA region and a region of chromosome 6q22.33, which contains the protein tyrosine phosphate receptor κ (PTPRK) and thymocyte-expressed molecule involved in selection (THEMIS) genes (Figure 2.6). No other regions were associated when lowering the significance threshold to the ImmunoChip Bonferroni significance threshold of 3.82×10^{-7}. There was some evidence of artificial inflation of association statistics, given the median squared z-score was 0.545, compared with 0.455 as expected under no inflation, as 0.455 is the expected

median chi-square statistic with 1 degree of freedom under the null hypothesis of no association. An inflation statistic, λ, can be calculated by dividing the median squared z-score by 0.455. If the λ statistic is calculated including the HLA region, λ was 1.20. However, the median squared z-score when excluding the HLA region (which was strongly AAD-associated and contains large LD blocks and can therefore have a large influence on λ), is 0.505 and $\lambda = 1.11$. I did not consider this worryingly high, especially given the ImmunoChip is designed to genotype regions associated with autoimmune diseases; it seems likely that, on average, therefore these regions have a higher probability to be AAD-associated than a random sample across the genome. However, to gauge which collections were driving the inflation, I examined inflation statistics within each collection. The results showed that the T1DGEN collection had the highest inflation factor, but this collection had just 249 individuals in it and thus would not be having a large influence on overall results. The next largest inflation factor was from the GRID collection and those collections with population structure and related individuals had a well controlled inflation factor (GRID $\lambda = 1.06$, NI GRID $\lambda = 0.95$, T1DGEN $\lambda = 1.08$, IDDMGEN $\lambda = 0.97$, Warren $\lambda = 1.04$, T1DGC $\lambda = 1.04$).

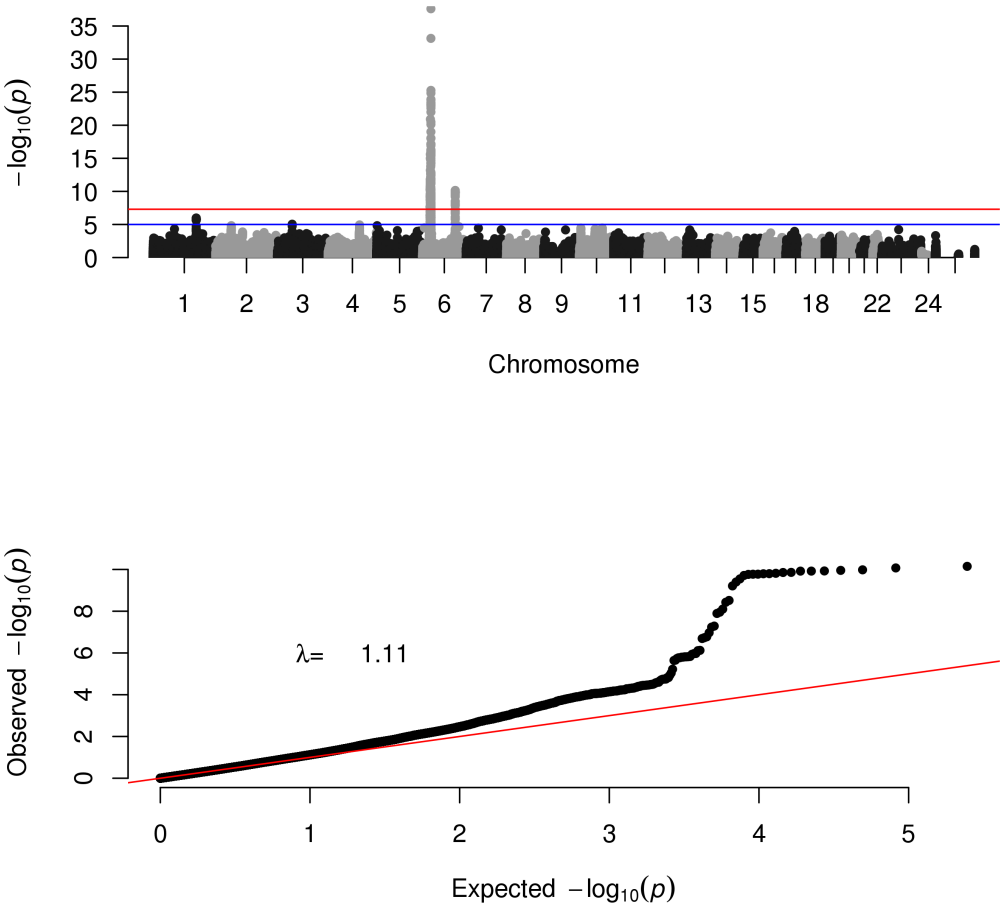
As an additional examination of inflation factor statistic, I repeated the GRID analysis but instead of adjusting for the five largest genetic PCs, I adjusted for geographical region in the UK, something that was collected as part of the GRID study. The inflation factor from this analysis was 1.077, higher than that when adjusting for PCs, providing support that the PC adjustment controls well for any population structure in this dataset. The inflation factor observed without any adjustment for population structure (neither geographical region nor PCs) is 1.081, indicating that the adjustment for either PCs or geographic region guarded slightly against artificially inflated association statistics, but that the population structure adjustment did not make too much difference in this relatively ancestry-homogeneous population of UK-based individuals.

The index variants at the two AAD-associated regions were rs9273363 (C>A) in the HLA region and rs72975913 (C>A) in the $6q22.33$ region. According to the results of the meta-analysis, the minor A allele at the HLA SNP rs9273363 was estimated to decrease

AAD by 12.0% (95% confidence interval 13.8%, 10.1%) and the minor A allele at the SNP in 6q22.33 was estimated to increase T1D AAD by 8.2% (95% confidence interval 5.7%, 10.6%). The associations remained genome-wide significant when adjusting for genomic control (performed by generating genomic control corrected p values by calculating $\frac{z^2}{1.11}$ and comparing this value to a chi-square distribution with one degree of freedom).

In addition to the two regions associated to $p < 5 \times 10^{-8}$, there was one other region that was associated to BY-FDR < 0.05 after removing the HLA region from the set of tests included in the FDR calculation. The index variant in this region on chromosome 1q24.3, rs10912265 (C>T), where the minor T allele was estimated to increase AAD by 5.74% (95% confidence interval 3.4%, 8.0%, $p = 1.06 \times 10^{-6}$). This variant is near the Tumor Necrosis Factor Ligand Superfamily Member 4 (TNFSF4) gene. This region has been associated with T1D before [66] but with a different index variant, rs78037977 (A>G). Upon examining this previously identified T1D-associated variant, I realised it was removed from the analysis due to failing QC in the Finnish genotype batch. To understand whether the AAD-associated signal was the same as the T1D risk signal, I performed a meta-analysis in the region including only the GRID and T1DGC collections. Indeed, in this analysis, rs78037977 was the index variant ($p = 2.69 \times 10^{-7}$ compared to rs10912265 $p = 7.28 \times 10^{-7}$), suggesting the signal identified here is the same as the previously identified disease-associated variant. I chose to focus on those regions that reached genome-wide significance, since these are most likely to be truly associated with AAD.

Figure 2.6: Results from ImmunoChip-wide meta-analysis. Top panel: Manhattan plot with each point showing the association of a genetic variant with T1D AAD and the x axis showing position along the genome. Blue line shows p value $< 1 \times 10^{-5}$ and red line shows p value $< 1 \times 10^{-8}$. Bottom panel: Quantile-quantile plot showing observed and expected $-log_{10}$ p values as well as the inflation factor, λ, for the analysis, calculated after excluding the HLA region.



Figures 2.7 and 2.9 show Forest plots with the effect size of the minor allele at each of the HLA and 6q22.33 region index variants for each collection. The effect sizes across collections is reasonably homogeneous, particularly in the two largest collections, GRID and T1DGC, which was reassuring that the effect was not driven by one collection. The MAF of the index variants by T1D AAD are plotted in Figures 2.8 and 2.10. The first thing to note is that the *A* allele at rs9273363 is not the minor allele amongst T1D cases, and secondly, the prevalence of this allele decreases as AAD increases. The opposite is true for rs72975913, where the prevalence of the allele increases with T1D AAD, but,

unlike the HLA variant, the effect appears to plateau at around AAD seven years.

Figure 2.7: Forest plot showing the log odds ratio (dots) and 95% confidence intervals (lines) for the addition of a minor A allele for the most associated variant with AAD of T1D in the HLA region, rs9273363.

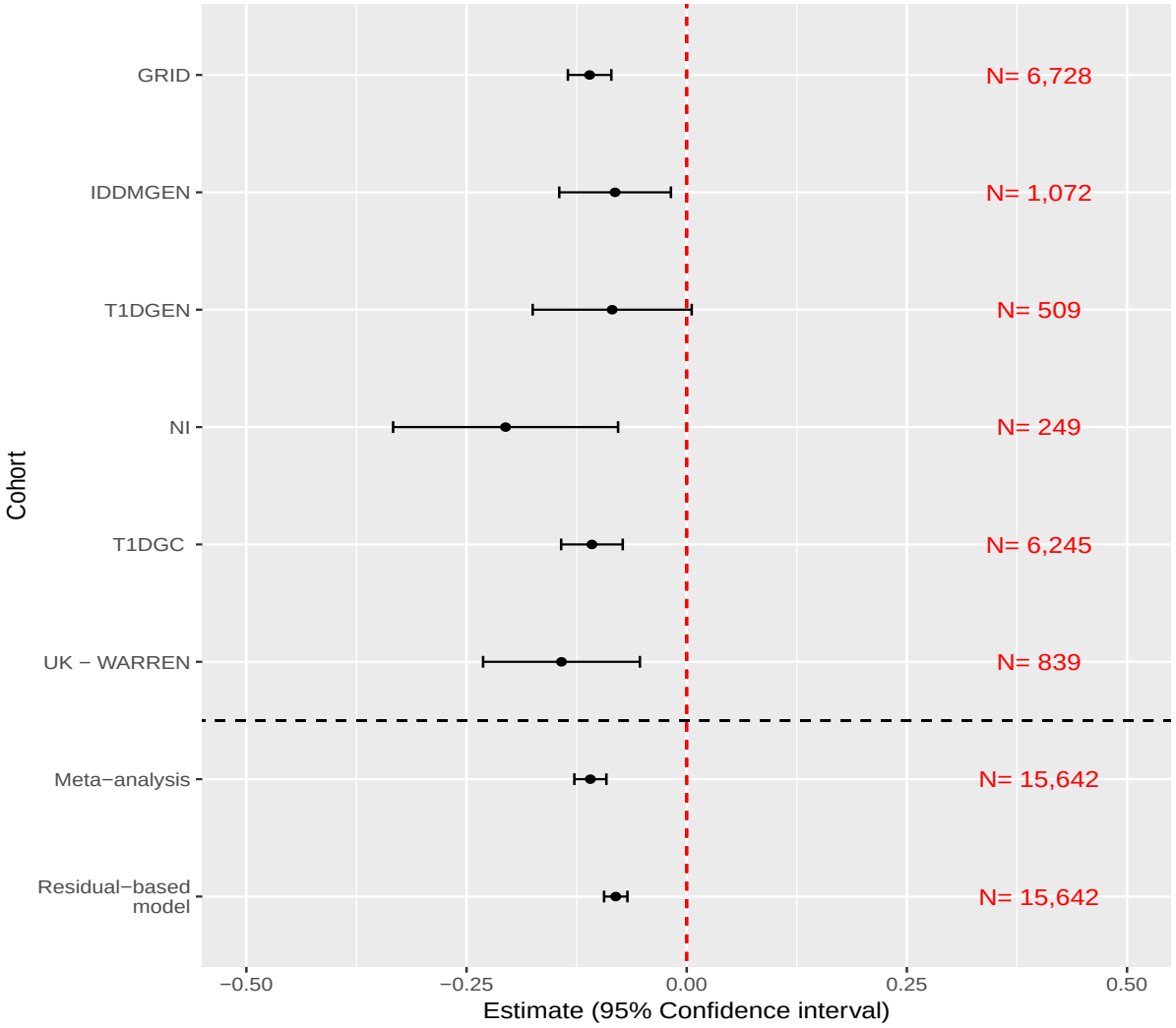


Figure 2.8: Minor allele frequency (95% confidence interval) of index variant in HLA region, rs9273363, by T1D AAD.

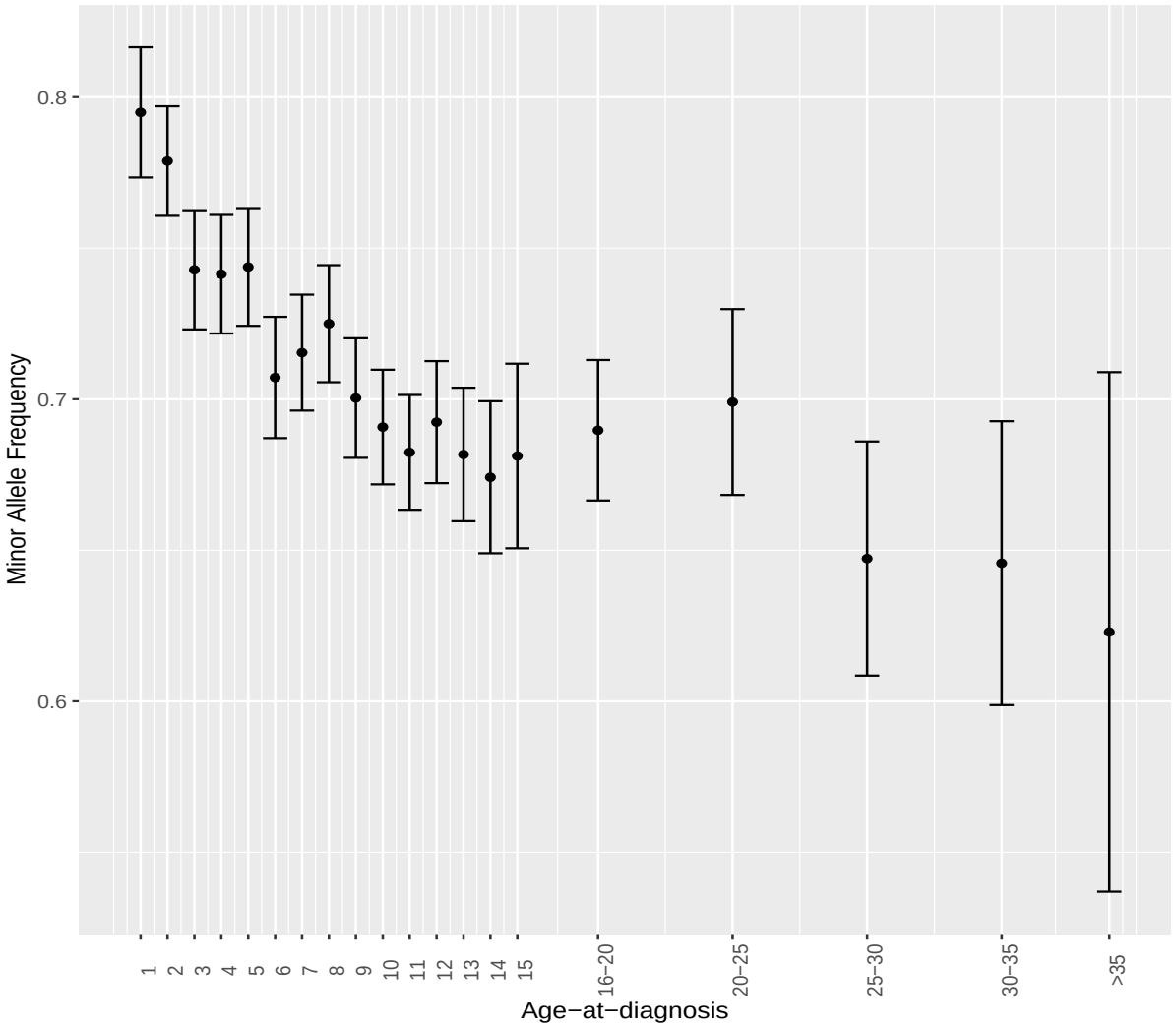


Figure 2.9: Forest plot showing the log odds ratio (dots) and 95% confidence intervals (lines) for the addition of a minor A allele for the most associated variant with AAD of T1D in the chromosome 6q22.33 region, rs72975913.

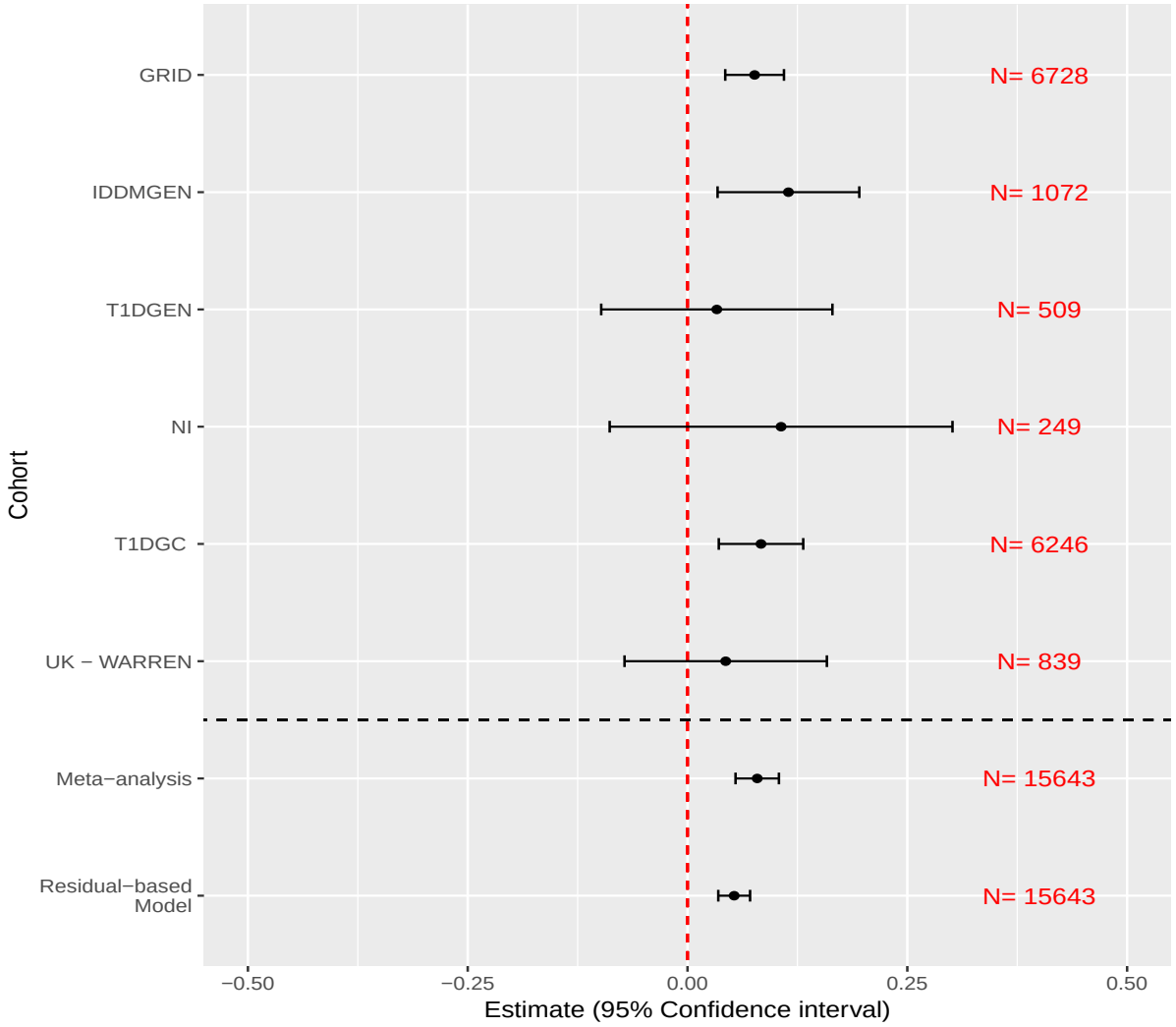
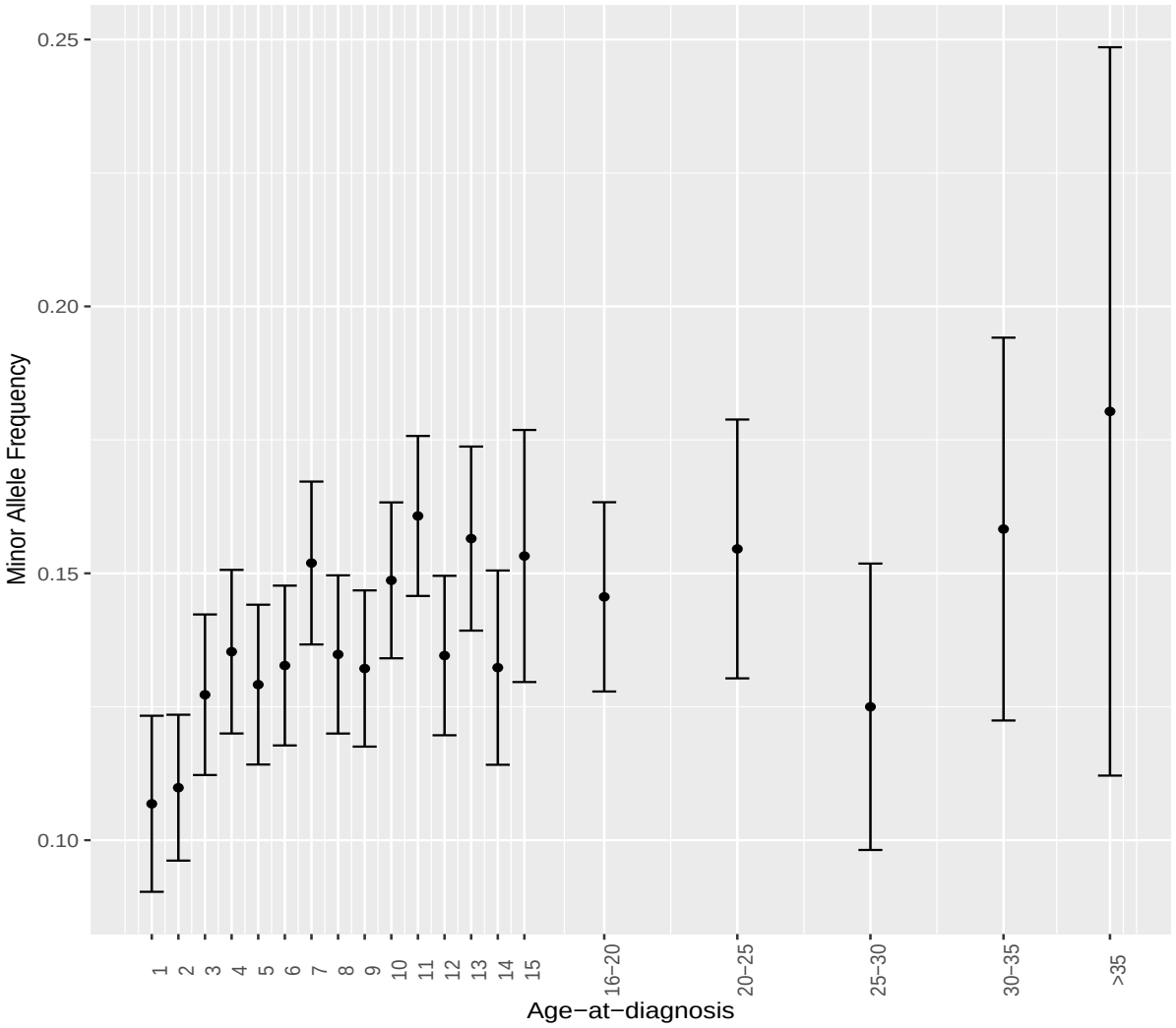


Figure 2.10: Minor allele frequency (95% confidence interval) of index variant in the chromosome 6q22.33 region, rs72975913, by T1D AAD.



2.3.3 Discovery analysis: residual-based method

After fitting the linear mixed model with random intercept terms for collection, country and family identifier, it was first important to examine the residuals to see the characteristics of the vector of which would later be used as the outcome variable for the present analysis and also for fine mapping. Figure 2.11 shows the theoretical normal distribution quantiles plotted against the quantiles of the residuals from the LMM. When including all individuals in the plot, the normal approximation looks reasonable but it is also clear that the data within collections do not follow the normal distribution particularly well, for example the GRID collection only recruited individuals up to 18 years of age and thus

the quantile-quantile plot reflects this, whereas collections that recruited a wider range of ages better fit the normal distribution. Figure 2.12 shows more residual plots, with a histogram of the residuals and also the predicted log_eAAD against the actual log_eAAD. It is clear that including only sex as a fixed effect and the three random effects does not lead to effective prediction of log_eAAD, which is expected. However, the correlation between the predicted values and the actual values is strong enough that I was comfortable continuing with this sensitivity analysis.

Figure 2.11: Quantile-quantile plots overall (left) and by cohort and country (right) of the residuals from the LMM regressing log_eAAD on sex and including family ID, collection and country as random intercept terms.

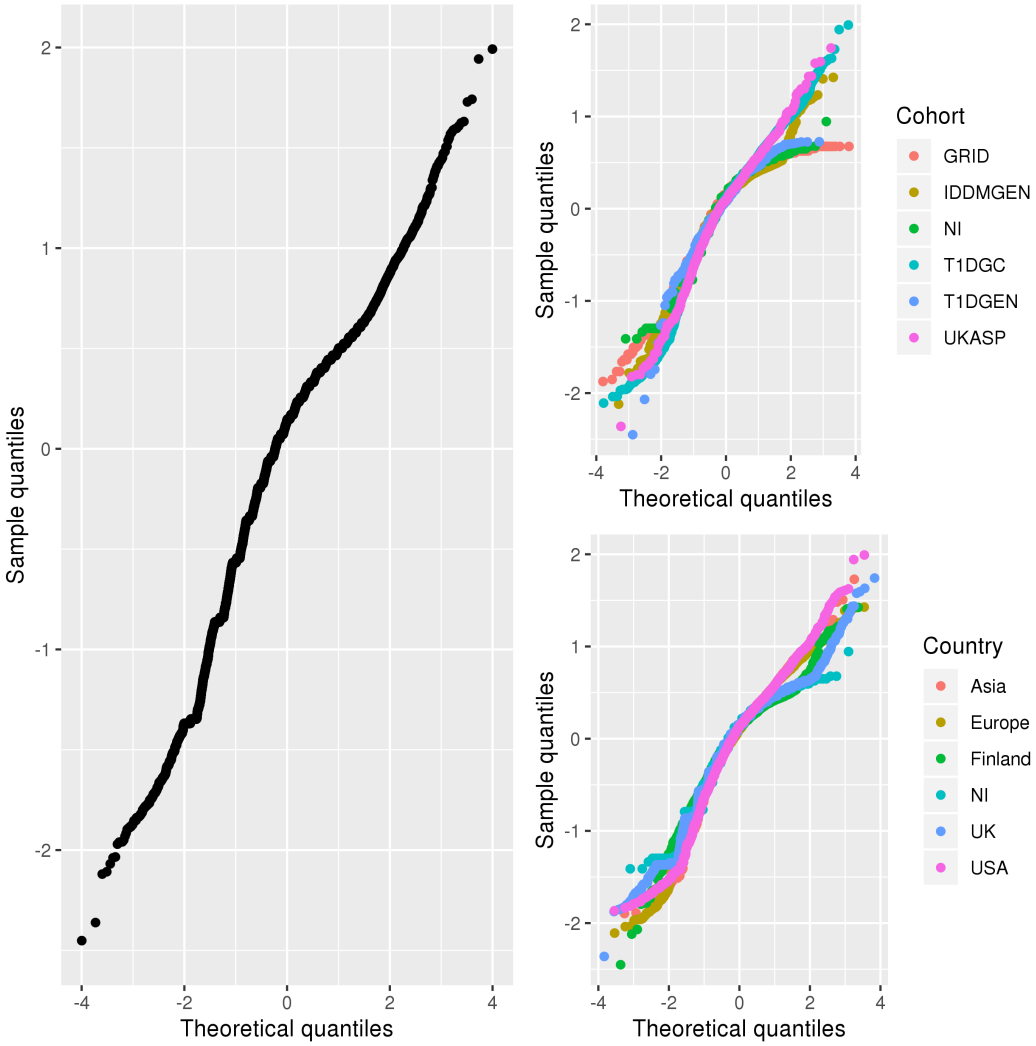
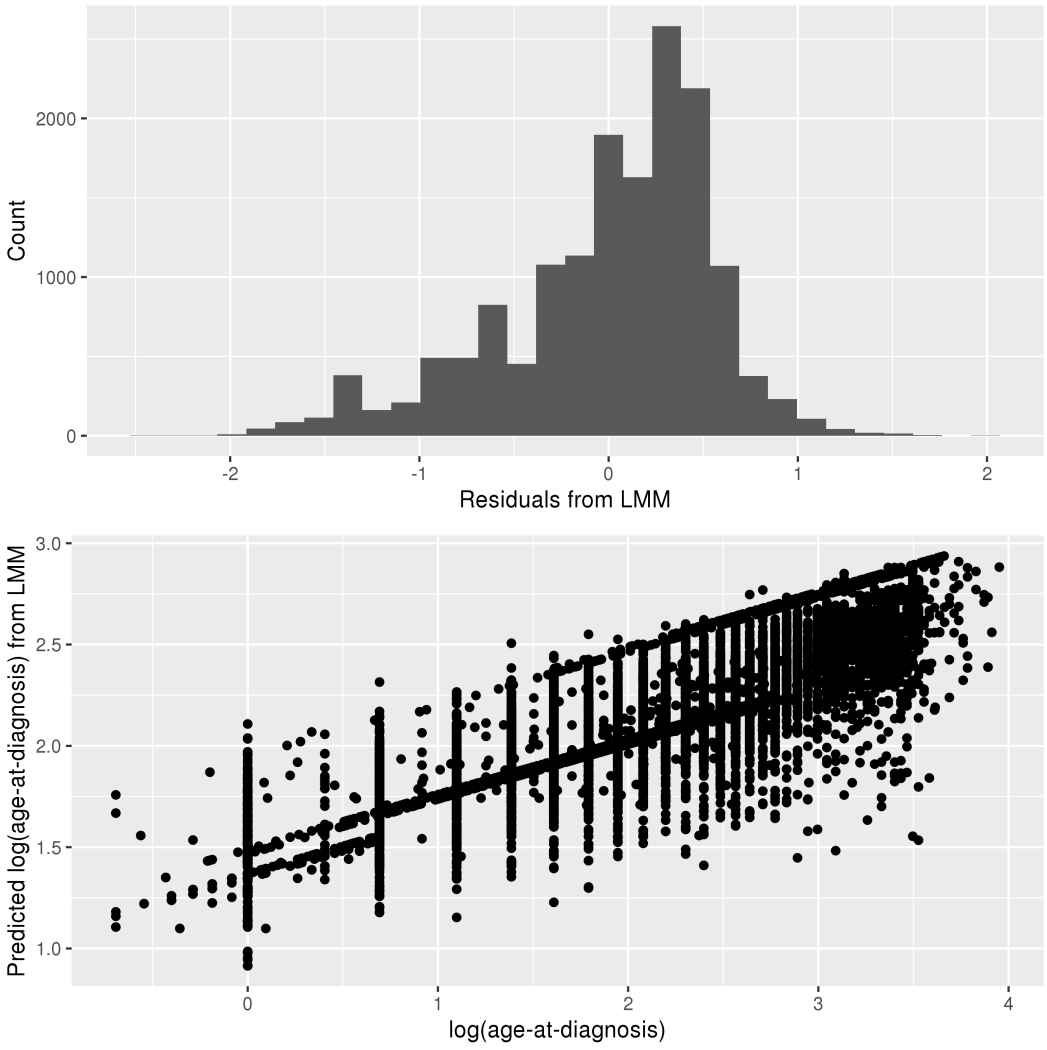


Figure 2.12: Top panel: Histogram of residuals from the LMM regressing $log_e AAD$ on sex and including family ID, collection and country as random intercept terms. Bottom panel: predicted values from the same LMM against actual $log_e AAD$ values.



The Manhattan plot for the residual-based method is shown in Figure 2.13 and shows the same two regions reaching genome-wide significance as in the meta-analysis, with a smaller inflation factor statistic once the HLA region was removed ($\lambda = 1.04$). This decrease in λ from this analysis implies that although the methods to handle population structure and relatedness were less sophisticated, the method is not producing a large number of false positive associations. The association with the variant near *TNFSF4*, rs10912265, which reached FDR significance in the meta-analysis, had $p = 1.24 \times 10^{-6}$. Generally, concordance between the two methods was good (Figure 2.14) but the methods in the meta-analysis for dealing with population structure and relatedness are generally more sophisticated than the simplistic LMM and thus I put more emphasis on the primary

analysis results than the residual-based method.

Figure 2.13: Results from residual-based method ImmunoChip-wide analysis. Top panel: Manhattan plot with each point showing the association of a genetic variant with T1D AAD and the x axis showing position along the genome. Blue line shows p value$< 1\times10^{-5}$ and red line shows p value $< 1 \times 10^{-8}$. Bottom panel: Quantile-quantile plot showing observed and expected $-log_{10}$ p values as well as the inflation factor, λ, for the analysis, calculated after excluding the HLA region.

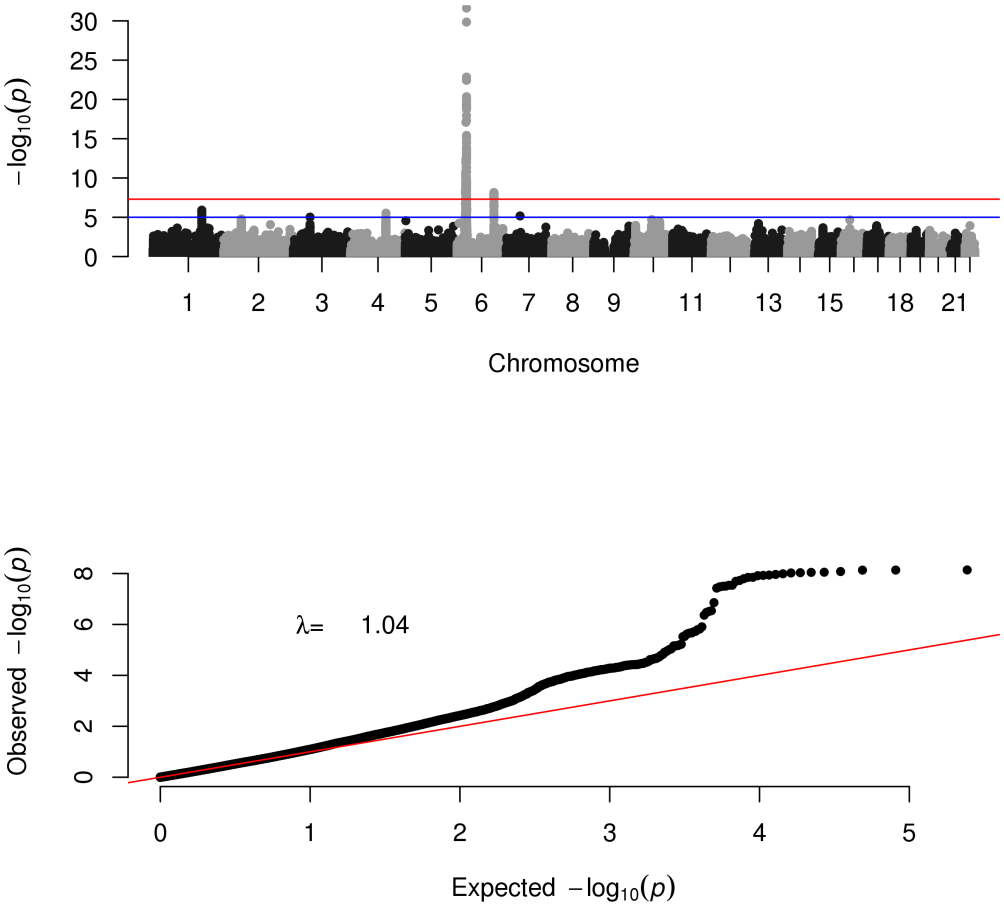
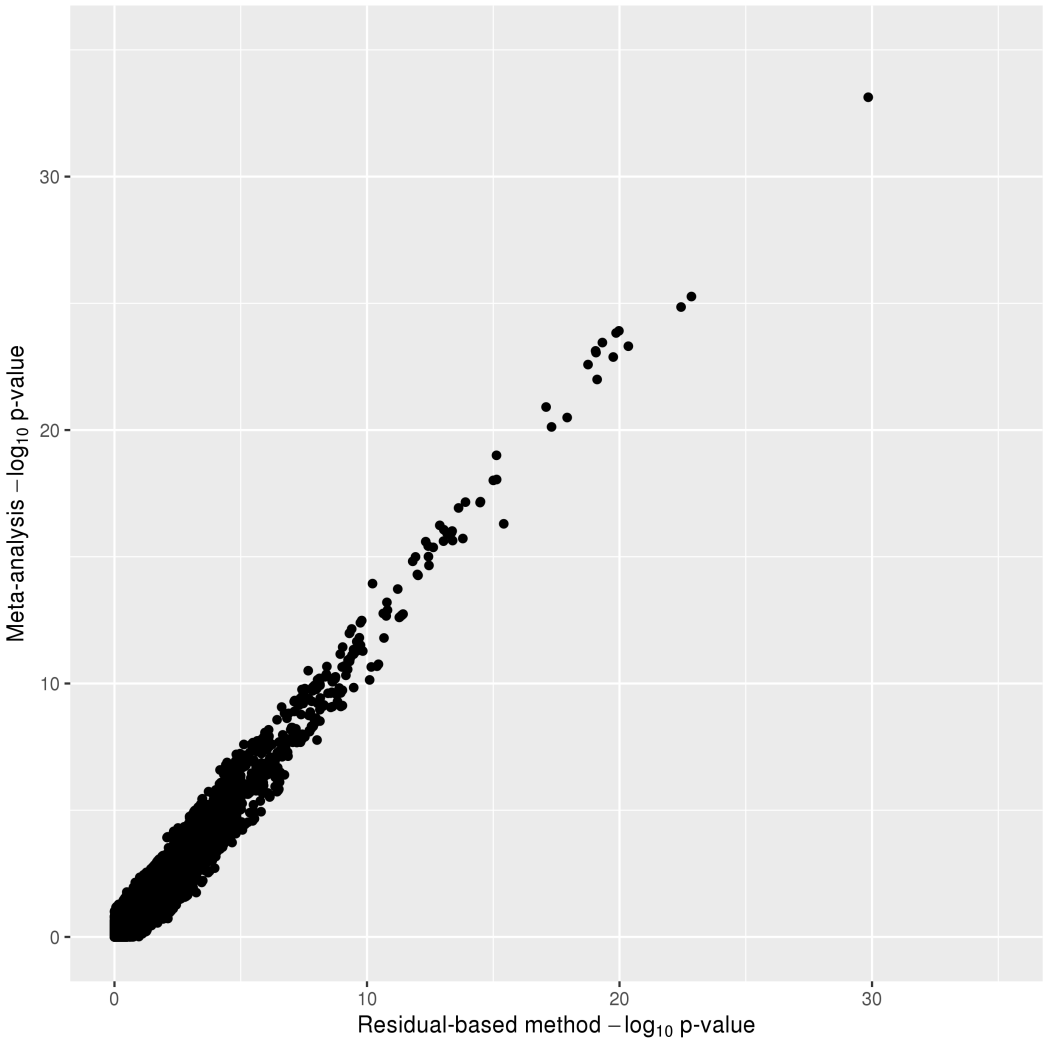


Figure 2.14: Comparison of association statistics of variants under the meta-analysis and the residual-based method.



The results from fitting a linear mixed model with the same fixed and random effects as the residual-based method does, but also including the variants of interest to obtain an effect size on the log scale rather than a residual scale, estimate that the addition of an A allele (the minor allele in controls) at the lead HLA SNP, rs9273363, is associated with an 11.8% decrease in AAD, which translates to a decrease of 1.11 years (13 months). The addition of the major C allele at the lead 6q22.33 SNP, rs72975913, is associated with a younger AAD by 0.7 years (8 months). If an individual is homozygous for the AAD risk allele at the lead SNP in both regions, they are estimated to be diagnosed 4.12 years younger than those who are homozygous for the non-risk allele at both loci.

Finally, for completeness and consistency between literature, I examined the association

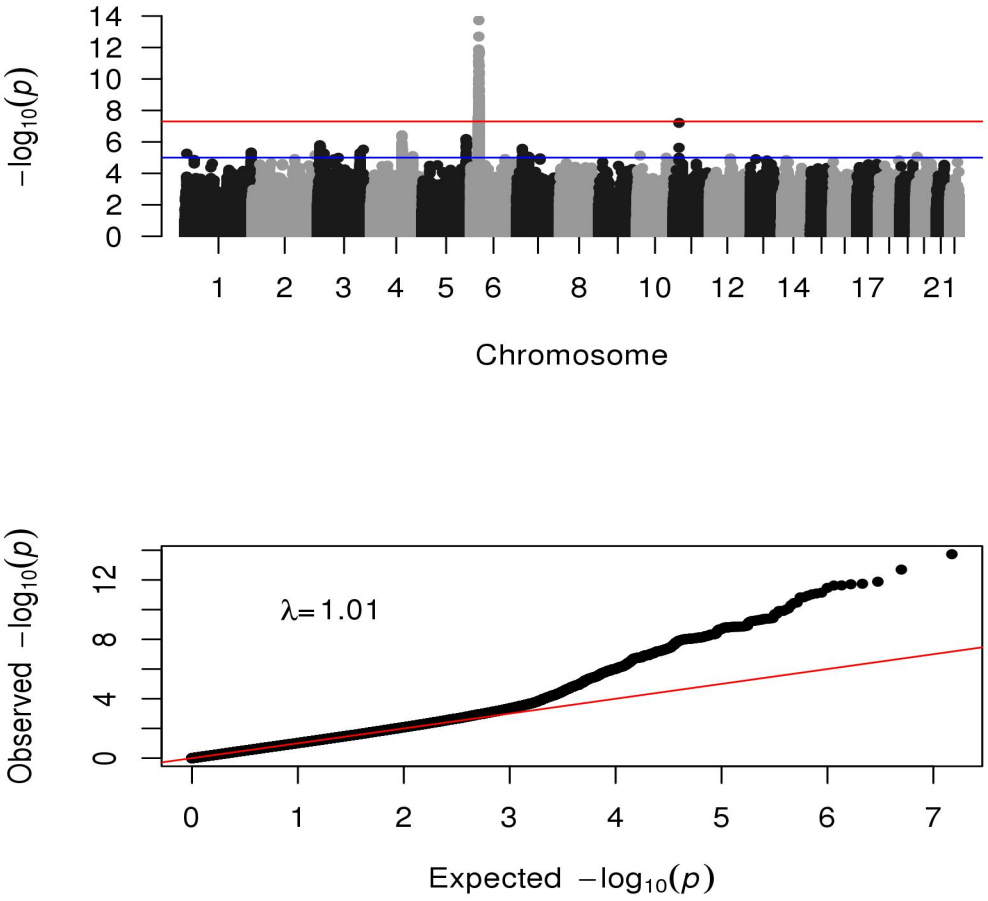
with AAD in the current analysis of the two previously identified AAD-associated SNPs near *RNLS*, rs10509540 (T>C) and *IL2/IL21*, rs2069763 (C>A). Both showed some association with AAD in this analysis, suggesting that they might be contributing to younger AAD but either there is insufficient statistical power to reject the null hypothesis at these variants to a genome-wide significance threshold or the age range covered in the current population is not of the appropriate distribution to detect the effects previously observed (rs10509540 minor C allele associated with 4.0% increase in AAD, $p = 7.37 \times 10^{-5}$, rs2069763 minor A allele associated with 3.4% decrease in AAD, $p = 1.87 \times 10^{-4}$). These effect estimates are reassuringly in the same direction as in [84].

2.3.4 Imputed GWAS analysis

Results from the imputed GWAS analysis showed only one region associated at genome-wide significance with AAD, the HLA region. The individuals included in this analysis are a subset of those included in the ImmunoChip analysis but the analysis has a higher coverage of SNPs given the GWAS platforms that were used for genotyping and the imputation carried out. Figure 2.15 shows the Manhattan and quantile-quantile plots for this analysis, with very little evidence of inflation of association statistics $\lambda = 1.01$ once removing the HLA region from this calculation.

Given the decreased sample size in this analysis ($N = 5601$ compared to $N = 15,696$ in the ImmunoChip analysis), it may simply be expected that the 6q22.33 region does not reach genome-wide significance, as it has a smaller estimated effect size than the HLA signal and only marginally reached significance in the ImmunoChip analysis. However, there is a hint of association at the index variant rs72975913, with a similar estimated effect size to in the ImmunoChip analysis ($p = 7.93 \times 10^{-5}$, $\beta = 0.109$). The other FDR associated variant, rs10912265, also shows a suggestive association with AAD here ($p = 6.26 \times 10^{-3}$, $\beta = 0.085$).

Figure 2.15: Results from imputed GWAS analysis including UK-based individuals only examining the effect of each genetic variant on T1D AAD. Top panel: Manhattan plot with each point showing the association of a genetic variant with T1D AAD and the x axis showing position along the genome. Blue line shows p value$< 1 \times 10^{-5}$ and red line shows p value $< 1 \times 10^{-8}$. Bottom panel: Quantile-quantile plot showing observed and expected $-log_{10}$ p values as well as the inflation factor, λ, for the analysis, calculated after excluding the HLA region.



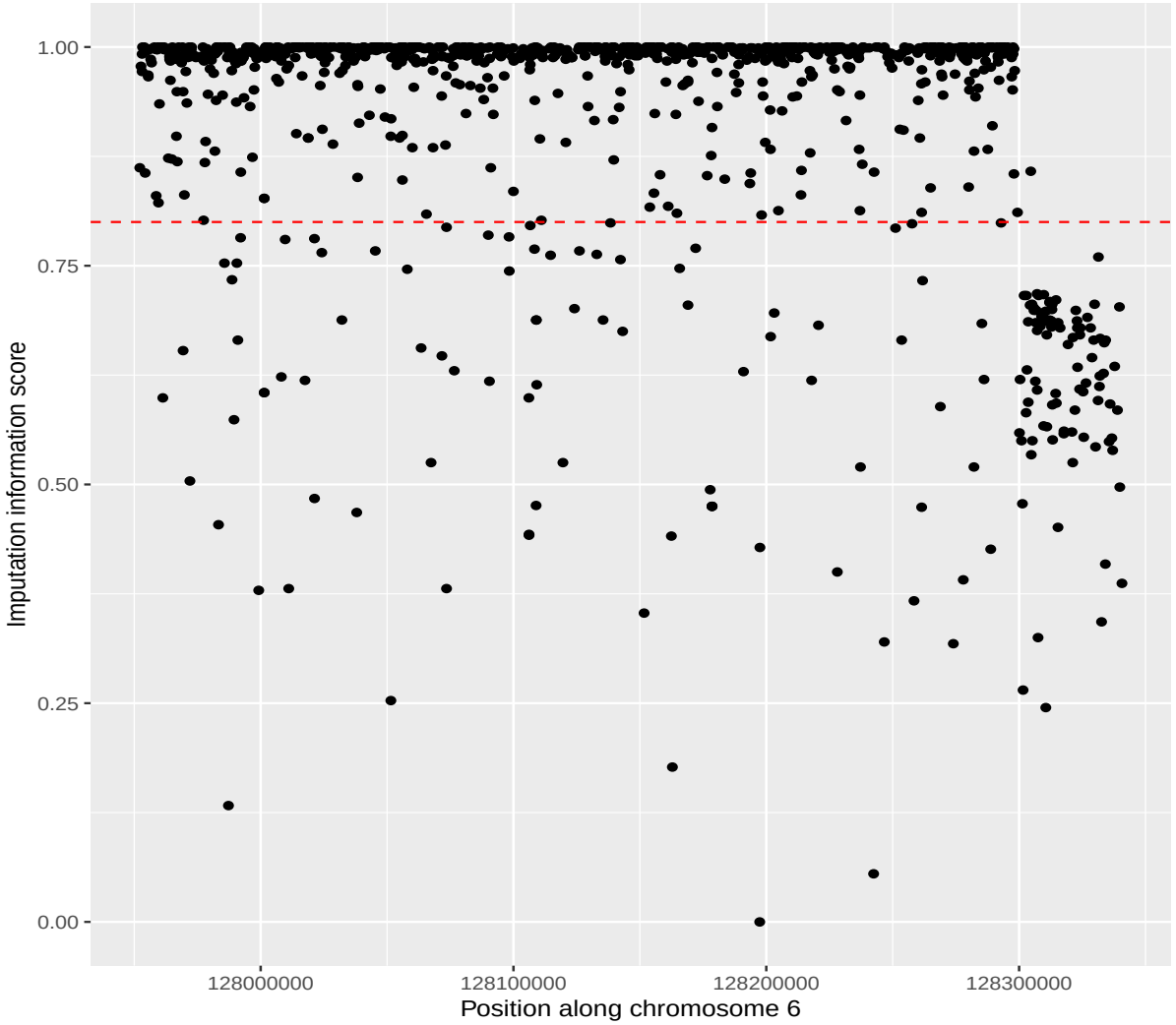
2.3.5 Fine mapping

Imputation

After imputation of the 6q22.33 region (between positions 127952182 and 128340790, genome build 37), 1071 variants were available for analysis. Of these, 186 had an information score of < 0.8 and were therefore removed (Figure 2.16). In addition, 99 variants were removed for either having a low MAF (< 0.01) or for being out of HWE (HWE

absolute z-score > 10), leaving 786 variants included in the fine mapping.

Figure 2.16: Imputation information scores in the 6q22.33 region after imputation with IMPUTE2 including all individuals.



Fine mapping, all individuals

Fine mapping the 6q22.33 region with the residuals from the LMM as the outcome and including all individuals in the analysis had a posterior distribution for the number of variants included in the model centred around two (Figure 2.17). Three groups were identified by GUESSFM, two of which were in LD with each other and therefore the algorithm could not determine which of the groups should be prioritised, but the LD was not so high between the groups that they could be combined. The index variants for the groups were rs11753289 (T>G), rs802719 (T>C) and rs802724 (C>T). The group indexed by rs11753289 contains the index variant from the discovery analysis, rs72975913

and it can be observed that in the majority of the most visited models in the stochastic search, a variant from this group was included in the model (Figure 2.18). In a proportion of models in those most visited, a variant from either the rs802719 or rs802724 group were included as well as a variant from the rs11753289 group, indicating there could be an independent second association, as suggested in the posterior distribution for the number of variants included in the model. There are no models in the most visited set that contain a variant from all three groups, implying there is most likely two signals, a variant from the rs11753289 and one from either of the rs802719 or rs802724 groups.

Figure 2.17: Prior and posterior probability of number of causal variants associated with AAD of T1D in the 6q22.33 region from analysis including all individuals.

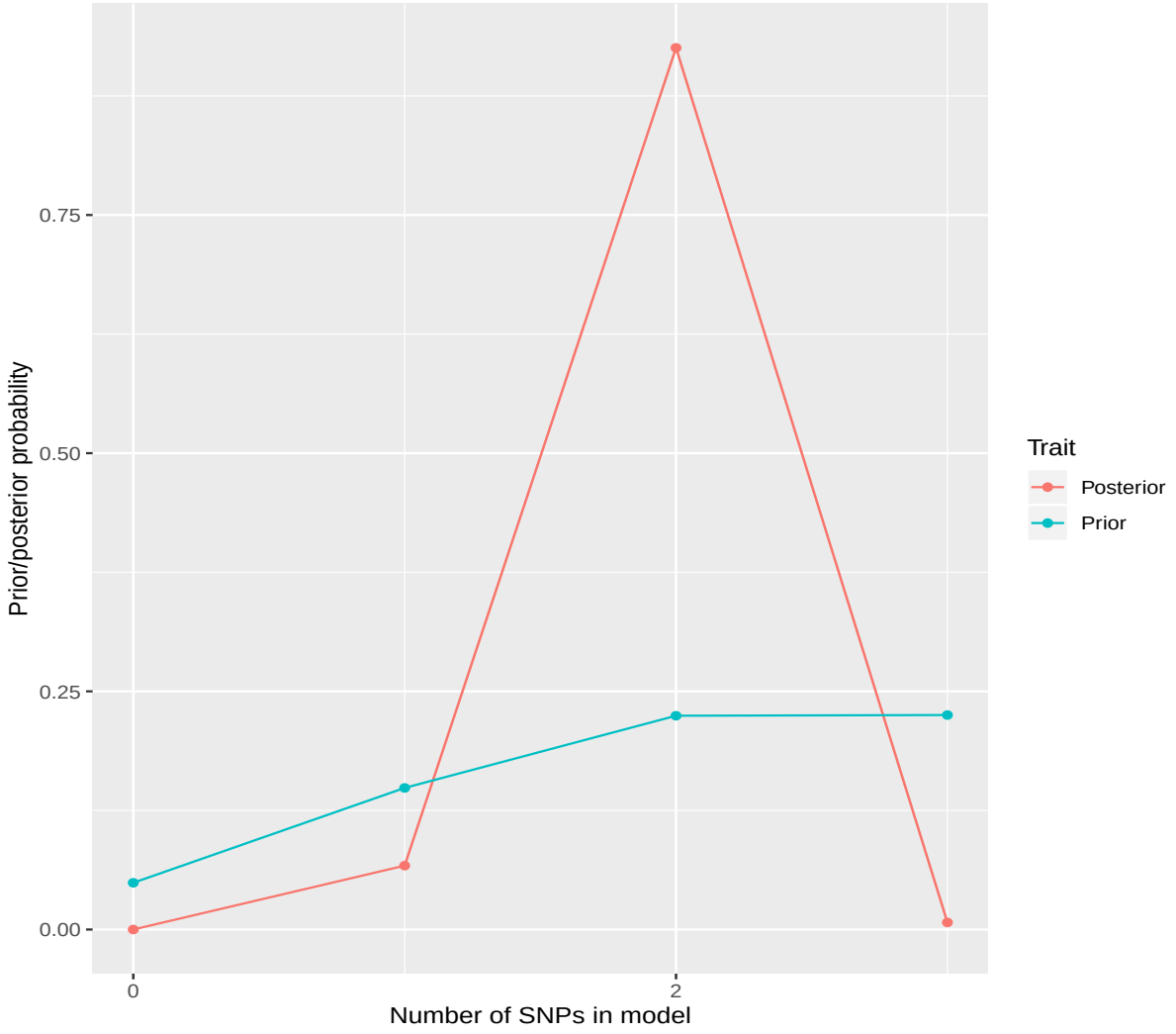


Figure 2.18: Graph showing which variants were included in models from the stochastic search including all individuals, suggesting most models contained one variant from the group indexed by SNP rs11753289 and some models which contains both rs11753289 and either rs802719 or rs802724. Shaded black area is the proportion of times the model includes a variant from the group that the column indexes in the stochastic search. Shaded black in more than one column indicates the proportion of times the stochastic search visits a model with one variant from each group.

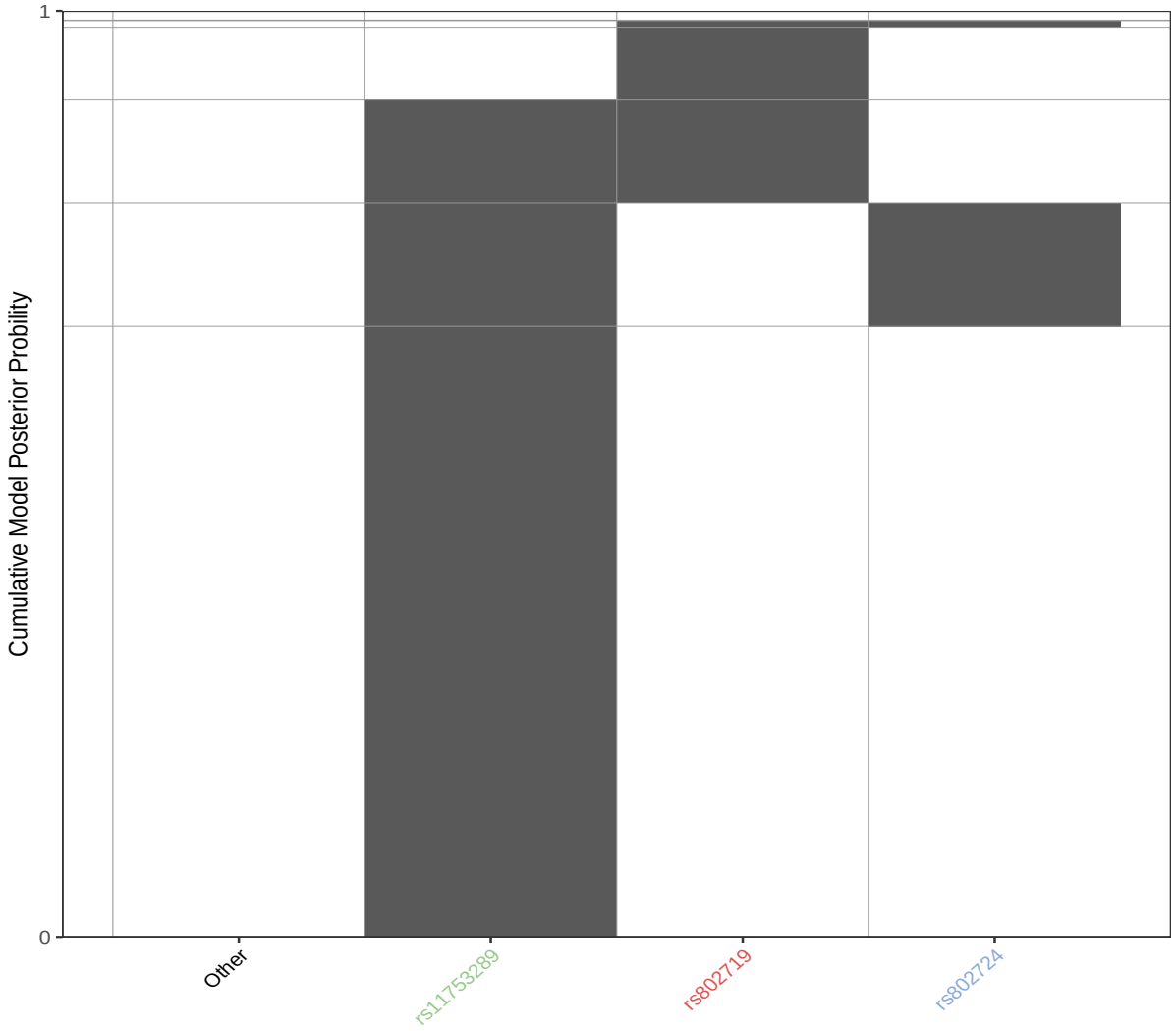
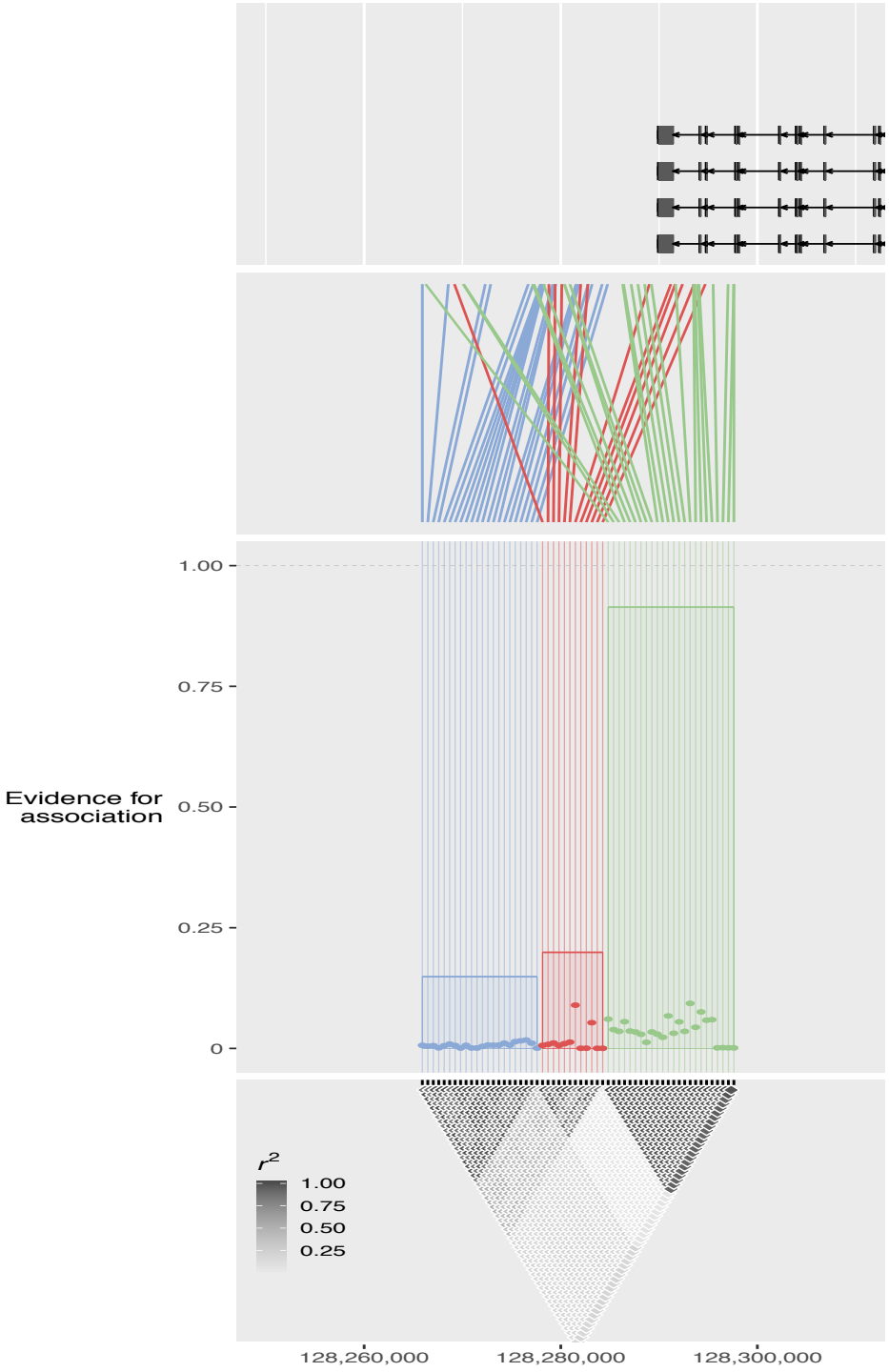


Figure 2.19 shows the locations of the most likely causally associated variants along the genome and also the LD within and between groups of variants prioritised. The group indexed by rs11753289 has the strongest evidence for association (group marginal posterior probability of inclusion= 0.91), and the variants are spread from near to the 3' untranscribed region of the PTPRK gene and towards the promoter region of the THEMIS gene.

Figure 2.19: Results from GUESSFM fine mapping the 6q22.33 region including all individuals. Bottom panel: LD r^2 between variants. Third panel: GUESSFM prioritised variants, dots show variant posterior probability of association, the shaded region shows the group posterior probability of association. Variants within the same group are shown next to each other to form group as opposed to their physical location along the chromosome. Second panel: maps the variants which are represented next to each other in panel 3 to their actual position along the genome (genome build 37). Top panel: PTPRK gene transcript locations.



Fine mapping, UK-based individuals

The UK-specific sensitivity analysis provided support for the primary results, with the same top signal detected. The posterior for the number of variants included in the model suggested one causal variant in this analysis (Figure 2.20), though it must be stressed that the sample size is smaller here and therefore the statistical power to detect the second signal is less. There is some evidence in the posterior distribution for a second independent association signal, it is just not large enough for a second group to be created in the GUESSFM algorithm. By visual inspection and comparison of the most visited models in this analysis compared to the analysis including all individuals, I observed that 8.4% of models contained a variant from the rs802719 or rs802724 groups in this stochastic search, implying that some of the models in the 'other' column in Figure 2.21 are a model including one of the two signals picked out from the primary analysis.

Figure 2.20: Prior and posterior probability of number of causal variants associated with AAD of T1D in the 6q22.33 region from analysis including UK-based individuals only.

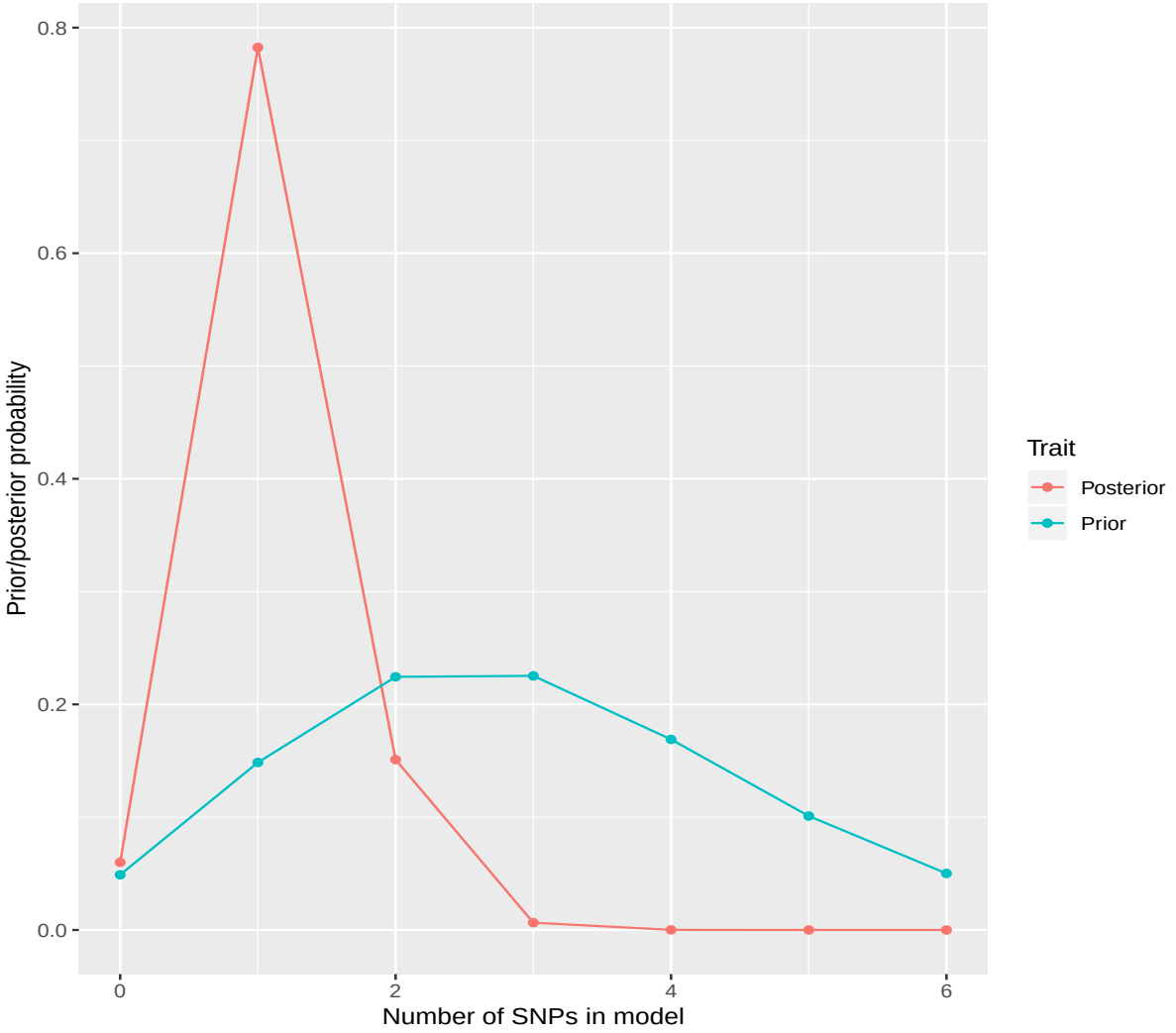
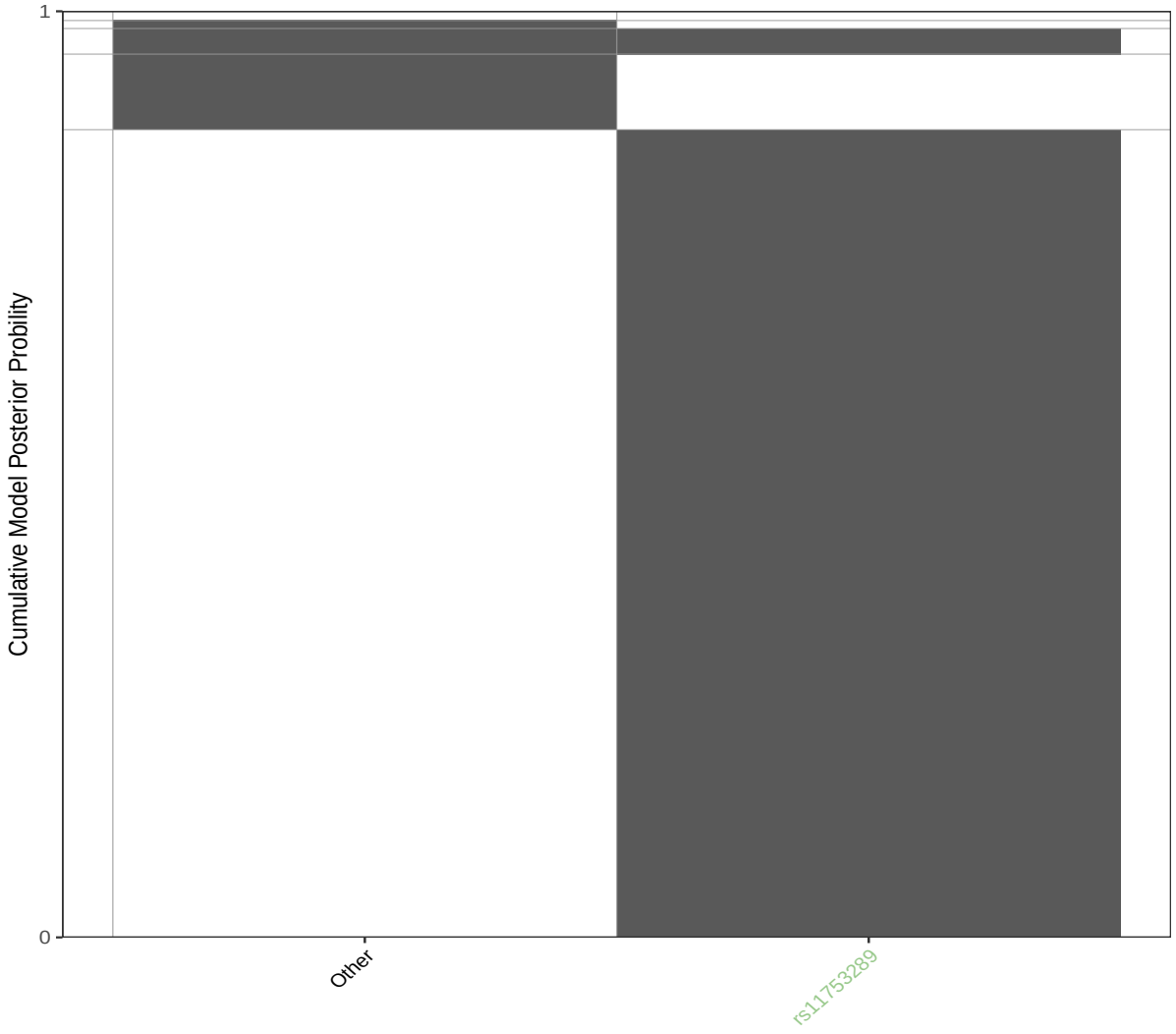


Figure 2.21: Graph showing which variants were included in models from the stochastic search including individuals from the UK only, suggesting most models contained one variant from the group indexed by SNP rs11753289. Shaded black area is the proportion of times the model includes a variant from the group that the column indexes in the stochastic search. Shaded black in more than one column indicates the proportion of times the stochastic search visits a model with one variant from each group.



Overall, there was enough concordance between the primary and secondary analyses to suggest there are likely two AAD-associated haplotypes on chromosome 6q22.33.

2.3.6 Haplotype analyses

Figure 2.22 shows the results from haplotype analyses including all individuals and only the variants with the highest marginal posterior probability of inclusion from each GUESSFM group. The effect size of each haplotype (with frequency > 0.01) is compared to the haplotype with all variants having the major allele. The third column is the only

column that shows the minor allele at rs11753289 and an increase in estimated log_eAAD can be observed. This column also shows that the minor allele at the third group indexed by rs802724, often appears with the minor allele at rs11753289 and therefore in univariable analyses the effect of the minor allele at rs802724 reducing AAD will be masked by the fact it sometimes appears with the minor allele at rs11753289, which is associated with increased AAD. The second column shows the effect of the second AAD-associated haplotype, indexed by rs802719 and the average decrease in estimated log_eAAD for the addition of a minor C allele. Again, the minor allele at the third GUESSFM group indexed by rs802724 often appears with the minor allele at rs802719, reflecting the LD between the two variants that was observed in the GUESSFM analysis. The fourth column shows the effect of the minor allele at the third group without either of the first two GUESSFM group minor alleles present. The effect estimate is similar to that of the second column (minor alleles at rs802719 and rs802724), but the standard error around the estimate is larger, since the frequency of this combination of alleles across the three variants is low. It is difficult to tell from this analysis whether the second causal signal is rs802719 or rs802724, reflecting the uncertainty observed in the model search by GUESSFM. Increased sample size would be the best way to elucidate the combination of alleles in the region that best explain the AAD phenotype, since there would be an increased number of haplotypes with a minor allele at one of rs802719 or rs802724 but not both.

The UK-specific sensitivity analysis (Figure 2.23) shows a similar pattern of association for each combination of alleles, indicating the secondary effect may also be present in this analysis, though with a smaller estimated effect size and larger standard error around the effect size estimate, as expected given the decrease in sample size.

Figure 2.22: Haplotype analysis of all individuals, including variants with highest posterior probability prioritised from the fine-mapping analysis.

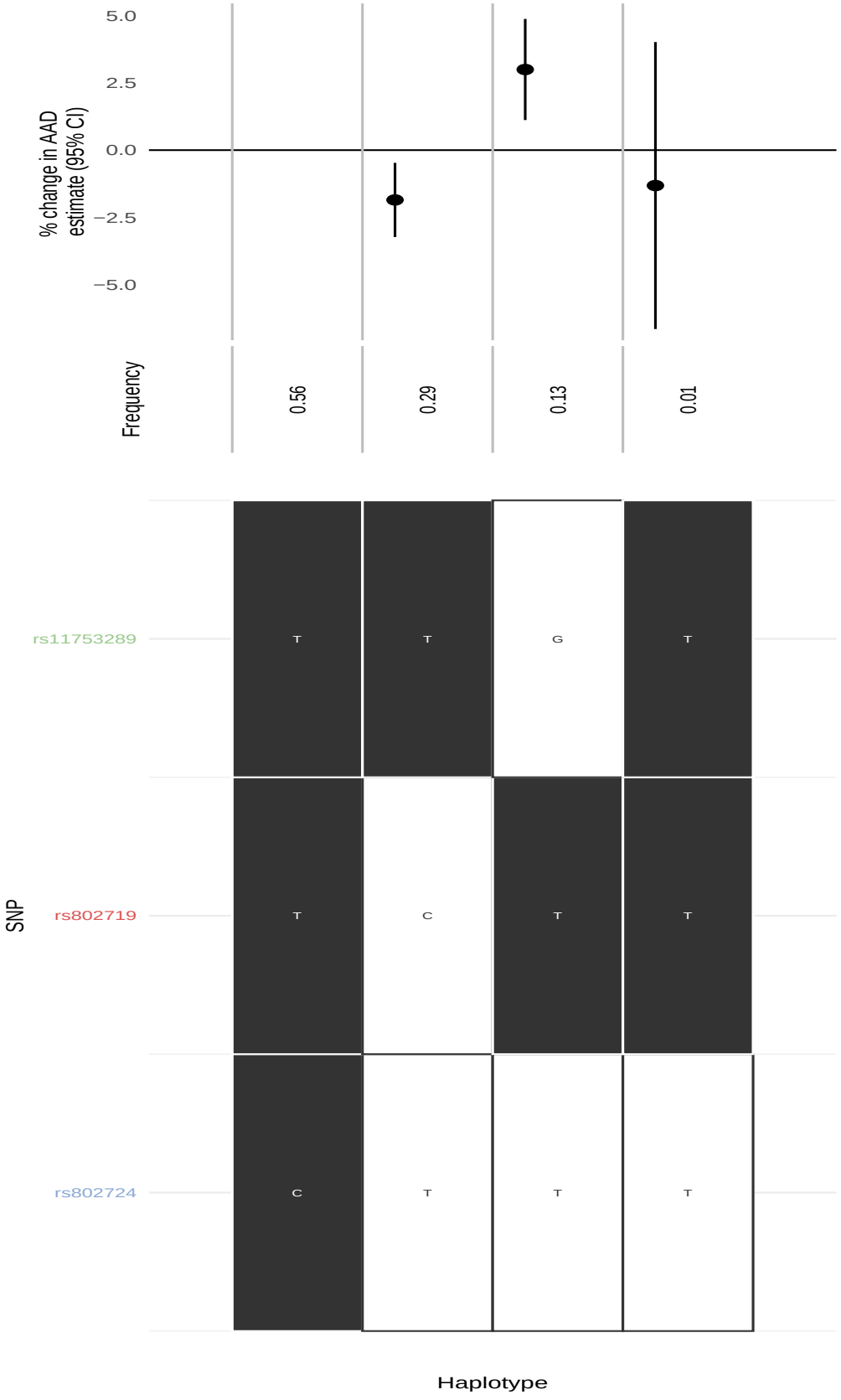
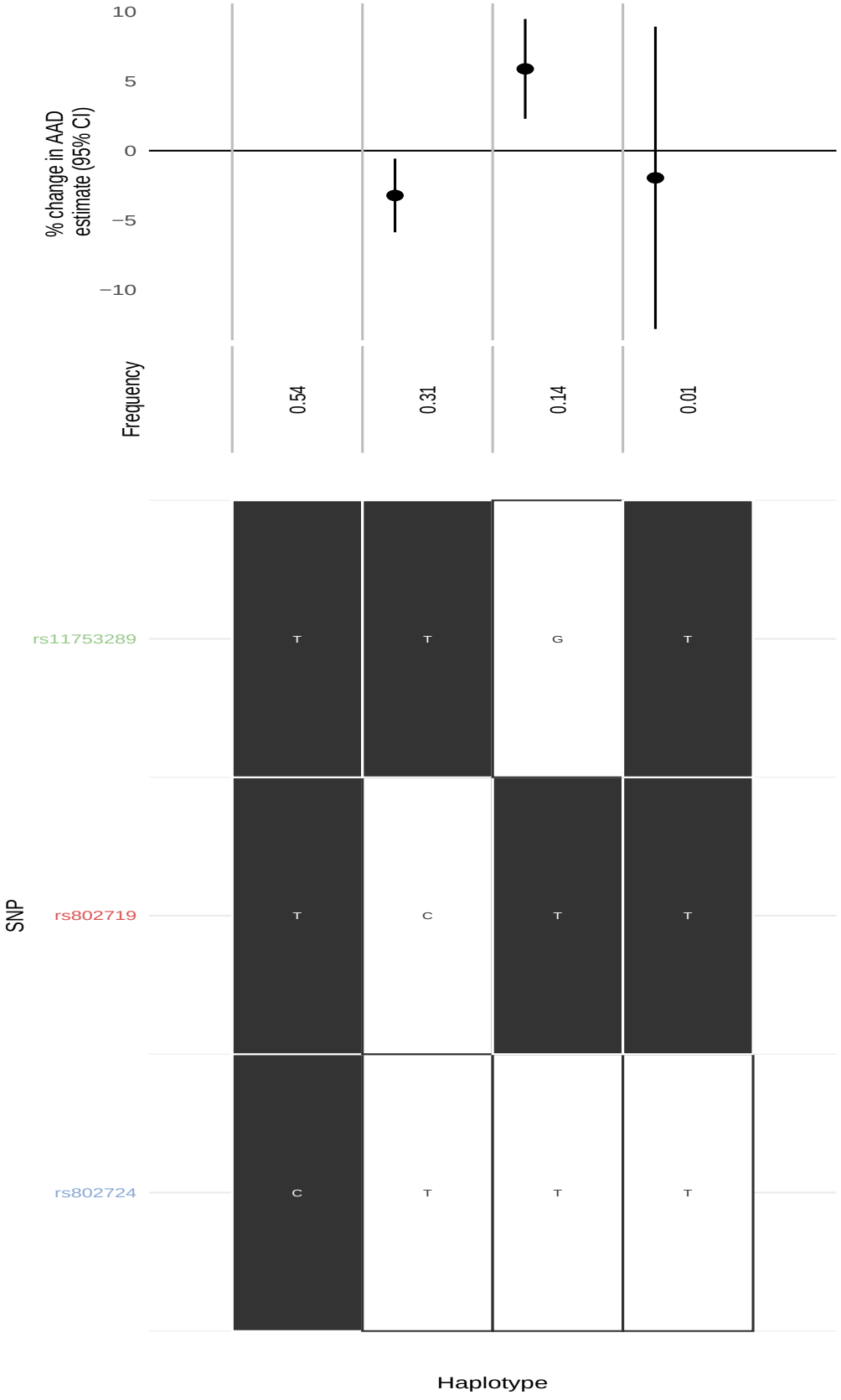


Figure 2.23: Haplotype analysis of UK-based individuals, including variants with highest posterior probability prioritised from the fine-mapping analysis.



2.3.7 Disease-association meta-analysis at various age-at-diagnosis groups

The effect size on disease risk of the HLA variant rs9273363 was larger in those diagnosed at < 5 years compared to all other AAD groups. Figure 2.24 shows a trend in effect size estimate whilst there still being a strong effect of the variant in the eldest AAD group. This contrasts to the rs72975913 (near *PTPRK*) and rs10912265 (near *TNFSF4*) variants (Figures 2.25 and 2.26), which have large effect sizes in the < 5 group (as well as reaching genome-wide significance) and show very small effect sizes at other AAD groups, indicating that the majority of the genetic risk at these loci is acting in individuals diagnosed at a young age.

Table 2.3 shows the effect size estimates for the < 5 analysis as well as the overall effect size estimate for T1D. For each SNP included in the Table, the effect size estimate for T1D diagnosed at < 5 years is larger than the overall T1D effect estimate.

Figure 2.24: T1D-risk meta-analysis by AAD-groups for the index variant in the HLA region, rs9273363. Effect size for the addition of the minor (in controls) A allele.

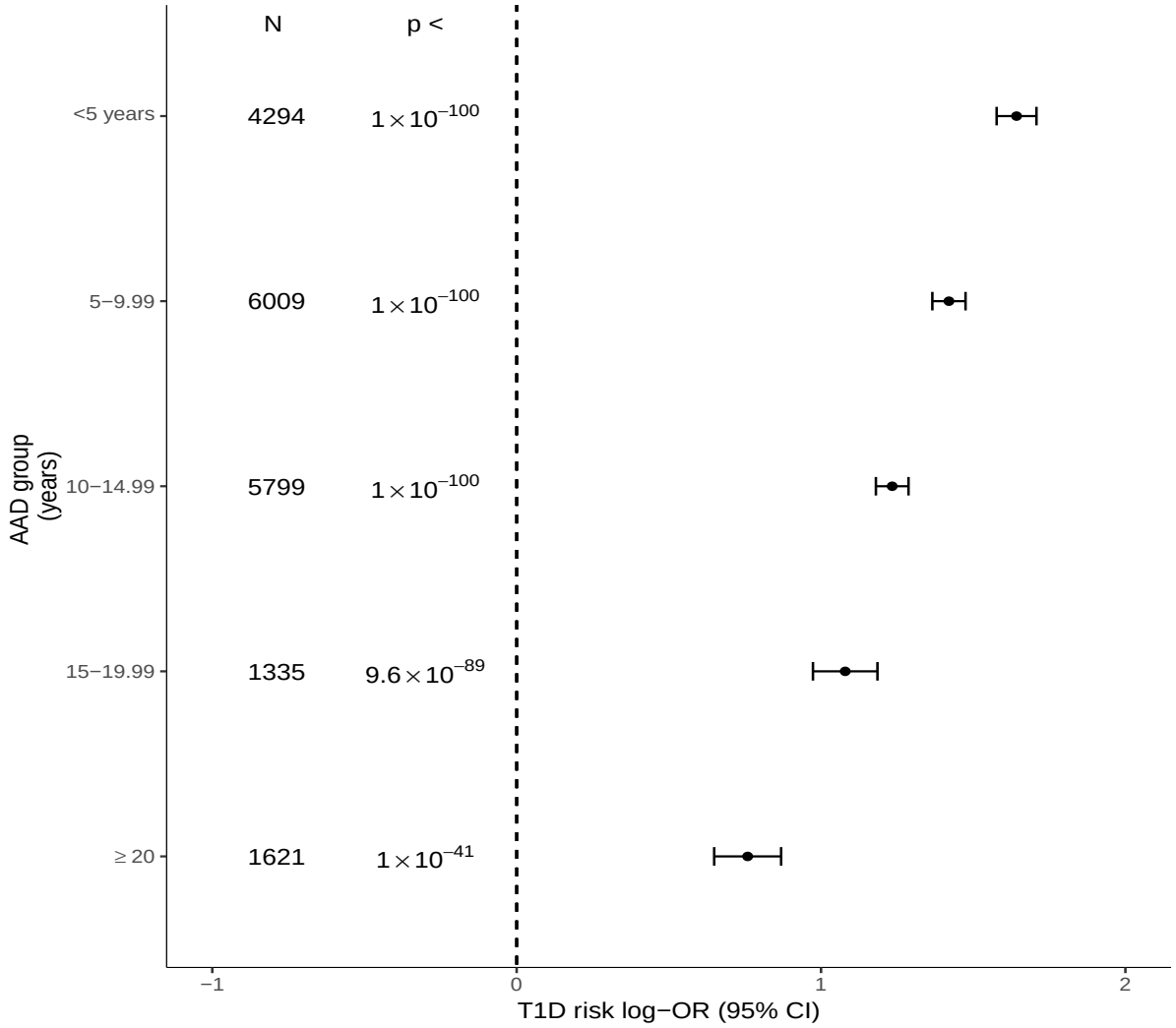


Figure 2.25: T1D-risk meta-analysis by AAD-groups for the index variant in the chromosome 6q22.33 region, rs72975913. Effect size for the addition of the minor A allele.

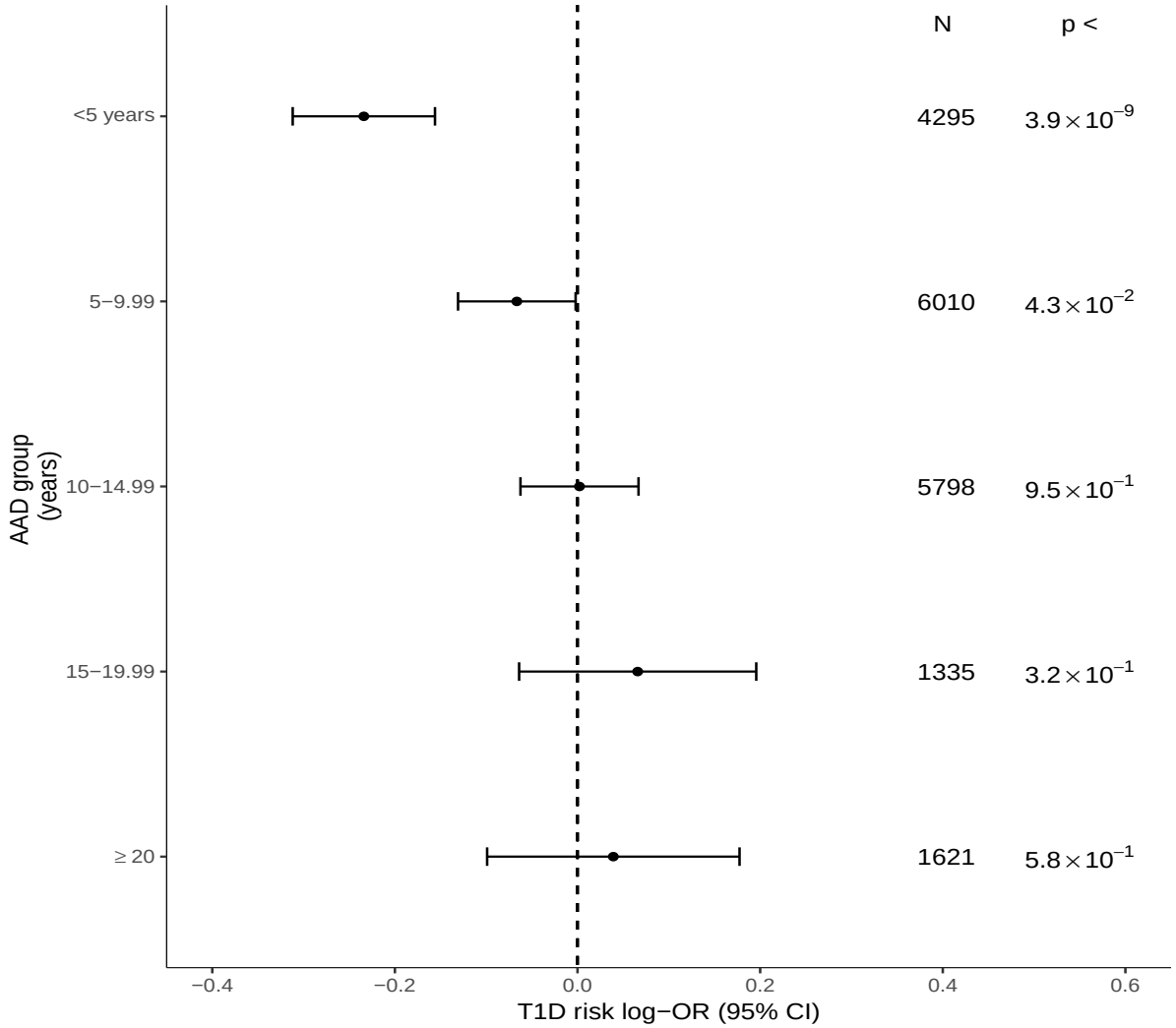
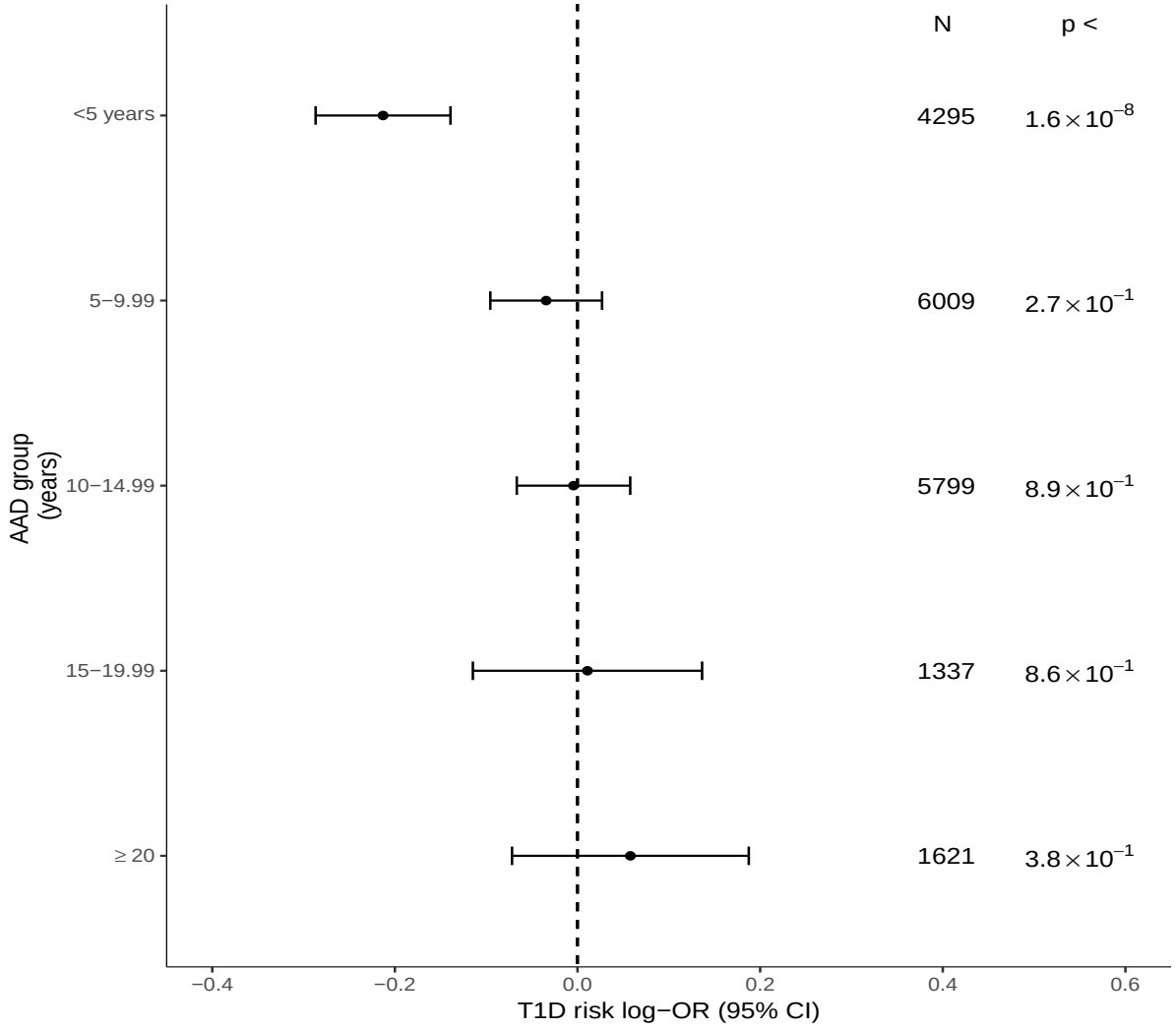


Table 2.3: Results for FDR associated regions showing summary statistics for associations with T1D overall and for T1D diagnosed at < 5 years.

SNP	Alleles	Gene	T1D OR (95% CI)	< 5 T1D OR (95% CI)
rs10912265	C>T	TNFSF4	0.94 (0.9, 0.98)	0.81 (0.75, 0.87)
rs72975913	C>A	PTPRK	0.93 (0.89, 0.98)	0.79 (0.73, 0.86)
rs802719	T>C	PTPRK	1.01 (0.98, 1.05)	1.14 (1.08, 1.2)
rs9273363	C>A	HLA-DQB1	3.61 (3.47, 3.74)	5.17 (4.84, 5.52)

Figure 2.26: T1D-risk meta-analysis by AAD-groups for the index variant in the chromosome 1q24.3 region, rs10912265. Effect size for the addition of the minor T allele.



Sensitivity to strata threshold

The results from altering the threshold for 'early-diagnosed' are shown in Table 2.4 and the associations remain despite the change in threshold, adding support to the hypothesis that these alleles are associated with AAD of T1D.

Table 2.4: Results for FDR associated regions showing summary statistics for T1D diagnosed at < 4 and < 6 years instead of < 5 years.

SNP	Alleles	<4 OR	<4 p value	<6 OR	<6 p value
rs10912265	C>T	0.804	4.3×10^{-7}	0.841	2.9×10^{-7}
rs72975913	C>A	0.757	1.5×10^{-9}	0.808	2.3×10^{-9}
rs802719	T>C	1.144	2.8×10^{-5}	1.123	5.9×10^{-6}
rs9273363	C>A	5.378	$< 5 \times 10^{-50}$	5.00	$< 5 \times 10^{-50}$

2.4 Discussion

Examination of AAD of T1D in an ImmunoChip-wide analysis lead to the discovery of a chromosome region that has not previously been associated with T1D risk that was found to be associated particularly with T1D diagnosed at a young age. This work has been published in a peer-reviewed journal [89].

The HLA region was also associated with AAD of T1D, which was expected given the previously-identified associations in this region [195, 135, 83]. The index variant in the HLA region, rs9273363, has been shown to tag the high risk T1D classical allele HLA-DQB1*03:02 (also known as DQ8) [137]. There is evidence that the HLA-DQB1*03:02 allele presents a repertoire of autoantigen-derived peptides that could cause the selection of autoreactive T cells [122]. This classical allele also predisposes to coeliac disease [51] and goes some way to explaining the increased prevalence of coeliac disease amongst T1D individuals compared to controls [51]. Individuals with T1D diagnosed at a younger age have a higher prevalence of coeliac disease [36, 51], which can also be explained, at least in part, from the T1D AAD-association with rs9273363 observed here.

The region on 6q22.33 contains two genes, *PTPRK* and *THEMIS*, both of which are expressed in the thymus. They have both been studied functionally and identified important roles in the thymus in selecting double positive thymocytes for differentiation to single positive thymocytes. Knocking out either gene reduces the number of mature $CD4^+$ cells entering circulation in mice [7, 93]. Moreover, knocking out both genes has a larger effect on number of mature $CD4^+$ cells in circulation than knocking out either separately, indicating that both genes are important in thymic selection [91]. Which of the two genes are acting to alter early-diagnosed T1D risk is yet to be understood, but this is explored further in chapter 3 where I examine statistical colocalisation with whole-blood eQTLs for both genes.

The 6q22.33 region shows association to three other autoimmune diseases, irritable bowel disease (IBD) [86], multiple sclerosis (MS) [166] and coeliac disease [193]. In coeliac disease, the secondary signal in the region is rs72975916 (C>T), which is in high LD with rs72975913 ($r^2 = 1$), and the effect of the minor T allele is in the same direction

as early-diagnosed T1D, where the T allele is protective [193]. It was also found that *THEMIS* expression was increased and *PTPRK* expression was decreased in the intestinal mucosa of active coeliac patients, but that neither transcripts were associated with the rs72975913 genotype [22]. In IBD, the effect is in the opposite direction to T1D, where the minor A allele at the lead AAD signal, rs72975913 is associated with susceptibility to Crohn's disease [86]. In rats, it was found that a mutation in the *THEMIS* gene lead to decreased levels of circulating $CD4^+$ T cells, decreased suppressive capacity of regulatory T cells and increased incidence of IBD [37]. In MS, the index variant, rs802734 (A>G) in the region is contained in the list of SNPs included in the GUESSFM secondary signal associated with AAD and is in the opposite direction [166], so the minor G allele is associated with protection from MS and susceptibility to early-diagnosed T1D.

There was no evidence of the region being AAD associated in either IBD or MS [86, 166], and this is difficult to assess in coeliac disease, given the common delay between onset and diagnosis [64]. However, the shared genetic signal between T1D AAD and coeliac disease may offer another genetic explanation for the fact that individuals diagnosed with T1D at a young age have higher prevalence of coeliac disease [36].

Although the 6q22.33 region has been identified as associated with T1D by examining AAD of T1D, it is conceptually unlikely to encounter a variant that decreases AAD but has no effect on disease risk itself. Instead, it seems likely that the variant does have an effect on disease risk, but this effect is only observed in young individuals. Once the individuals have aged into adolescence, the effect of this variant on T1D risk is no longer present. This is a subtly different conclusion to suggesting having this variant will decrease the AAD of T1D but not increase disease risk. The thymus undergoes involution (shrinking) with age [169], so more naive cells are in circulation in childhood than in adulthood, where the thymus becomes fatty tissue with minimal function. Given the variant identified in 6q22.33 is near two genes with known function in the thymus, and the immune system is reliant on the thymus for immune tolerance in the first years of life, it can be speculated that everyone with the major C allele is at increased risk of T1D in the first years of life, but once the immune system is more fully developed and

less reliant on thymic selection, that increased risk is lost.

The third AAD-associated region (FDR< 0.05) near *TNFSF4* was indexed by rs10912265 in the primary analysis since rs78037977 was removed from the analysis, but secondary analyses suggest rs78037977 is the more likely causal AAD-associated variant in the region, which is the same variant that was found to be associated with T1D [66]. This variant has also shown association with vitiligo, where the minor G allele is associated with increased risk of vitiligo, whereas the minor allele is associated with older AAD in this analysis and protection from T1D in [144]. There is weak evidence to suggest the minor G allele is also protective for asthma (`https://genetics.opentargets.org`).

There were a number of limitations in this analysis, some of which could be explored and others that could not be accounted for in this dataset. For example, the AAD distribution was mainly individuals diagnosed at < 20 years, so any conclusions about variants associated with AAD should be interpreted as a variant associated with younger childhood-diagnosed AAD. There is therefore some ascertainment bias in our collection, since it has been suggested that T1D incidence remains at a similar level throughout life but later diagnosed T1D is often misdiagnosed as type 2 diabetes [188]. Further analyses would need to be carried out to explore whether any variants are associated with childhood-diagnosed disease compared to adult-diagnosed (> 20 years), which this analysis cannot answer. Indeed, in a recent analysis that examined the genetic determinant of AAD of T1D using an older collection of T1D individuals, the association in the chromosome 6q22.33 region failed to replicate, almost certainly owing to the small number of individuals in the cohort diagnosed at < 5 years [130]. In addition, one publication has explored the similarity in childhood-diagnosed and adult-diagnosed T1D, but only examined 20 known T1D-risk loci and did not look genome-wide for differences. They found evidence that nine of the 20 regions associated with paediatric T1D were also associated with adult-diagnosed T1D [85].

A second potential limitation of the analysis is that the inflation factor, λ, is quite high in the primary analysis at 1.11. This is high despite the conservative attempts to handle population structure. An examination of the inflation factor statistics of each collection

separately showed this inflation was caused mainly by the GRID cohort, which is the largest and most ancestry-homogeneous collection in the analysis. Had the inflation have been driven by the T1DGC collection, which contained individuals from multiple ancestries, I would have been more concerned by the inflation factor as it may have been reflecting association driven by population structure, but the fact it was highest in the GRID collection was less concerning and may just reflect increased statistical power in that collection. In addition, the inflation factor was lower in the residual-based model, and the same two regions were found to be associated with AAD to a genome-wide threshold, providing evidence that these associations were likely not due to population structure in the dataset alone.

In conclusion, by analysing AAD as a continuous covariate, a region in the genome not-previously associated with T1D has been discovered to be associated specifically with early-diagnosed T1D, which is the most aggressive form of the disease and accompanied by the highest prevalence of complications and largest decrease in life expectancy. This potentially important discovery may point to a deeper understanding of early-diagnosed T1D, and may be of benefit when examining future therapeutic targets.

Chapter 3

Type 1 diabetes variants acting primarily in early-diagnosed individuals

3.1 Introduction

The genome-wide analysis of AAD as a continuous outcome in chapter 2 identified a novel T1D-associated locus using this approach. A potential limitation of this analysis was that by examining regions from across the genome using the ImmunoChip, there is a multiple comparisons burden, given the number of variants tested for their association with AAD. However, the discovery that this locus is associated with T1D diagnosed at < 5 years old opens up the possibility that there could be an AAD threshold delineating different clinical and phenotypic features, which may have subtle differences in underlying genetic susceptibility. By examining only previously identified T1D susceptibility regions in the genome to investigate this hypothesis, rather than examining every variant on the ImmunoChip, alleviates the multiple testing burden and potentially increases sensitivity to detect variants more strongly associated with early-diagnosed T1D over later-diagnosed T1D.

In an epidemiological study characterising mortality risk for T1D individuals diagnosed

at various ages, it was observed that those diagnosed at < 10 years had an approximately 15 year shorter life expectancy than matched controls, whereas for individuals diagnosed at > 15 years, the difference in life expectancy compared to matched controls was approximately 10 years [157]. There was also a greater risk of cardiovascular disease and cardiovascular related mortality in the early-diagnosed compared to later diagnosed individuals [157]. A number of other studies have attempted to dichotomise individuals into AAD groups based on phenotypic differences. It has been shown that T1D individuals diagnosed at < 5 years have faster C-peptide decline [14], a higher coincidence of T1D and coeliac disease [147] as well as higher rates of diabetic ketoacidosis [90] compared to individuals diagnosed at ≥ 5 years. A decline in cognitive function has also been linked with T1D diagnosed at < 5 years [75]. There is evidence that levels of particular phospholipids in cord blood are altered in individuals who go on to be diagnosed with T1D at < 4 years compared to those diagnosed at ≥ 4 years [107].

A potentially important observation was made when a B-cell infiltration signature was identified coincident with T-cell infiltration in the pancreatic islets of newly-diagnosed T1D individuals aged < 7 years. Those newly-diagnosed and aged > 13 years had a similar T-cell infiltrate but the ratio of $CD20^+$ B cells to $CD4^+$ T cells was markedly lower than in those diagnosed at < 7 years [109]. This observation could, in part, be down to the altered composition of immune cell subsets with age. Two publications have suggested a B-cell signature in poor responders to T1D therapy, which reduce or disappear once adjusting for age in their analysis [56, 116], suggesting the B-cell signature observed is reflecting the known effect of younger-diagnosed individuals losing their C-peptide faster. Regardless, the striking difference in $CD20^+$ B cell to $CD4^+$ T cell ratio in the target tissue for T1D may reflect some heterogeneity in underlying disease aetiology between early-diagnosed and later-diagnosed T1D.

Given the array of clinical and phenotypic differences between early-diagnosed and later-diagnosed T1D, defined by various AAD thresholds, it would be of interest to examine all the known T1D-associated genetic regions and see whether any have markedly heterogeneous effects between early-diagnosed and later-diagnosed T1D cases. A genetic

explanation for the heterogeneity in phenotypic characteristics would be of interest, since the individuals diagnosed at a young age represent the subset with the highest susceptibility to later in life complications and early mortality, and therefore identification of any therapeutic targets tailored towards for this subset would be of greatest overall benefit.

3.2 Methods

The aim was to investigate whether there were any T1D-associated loci that had a higher estimated effect size in those individuals diagnosed at < 7 years compared to those diagnosed at ≥ 13 years, whilst also examining the effect in those diagnosed at $7 - 13$ years; if the effect size of this intermediate group fell between the < 7 and ≥ 13 groups, it would add weight to the hypothesis that the locus under consideration has a stronger effect in early-diagnosed individuals.

3.2.1 Individuals included in analysis

ImmunoChip data from $39,623$ samples were used, which is the same dataset before QC as was used in chapter 2. Some samples were duplicates and others were removed due to low quality genotyping throughout the genome. Samples were first separated into ancestry-homogeneous groups by combining the UK case-control, the Northern Irish case-control and the Warren affected sibling pair collections together to form a UK collection, the T1DGEN and IDDMGEN Finnish cohorts were combined to form the Finnish collection and the T1DGC families were considered a separate collection. Variant QC was performed, where variants with a low MAF < 0.005, low call rate < 0.95 or an absolute HWE exact test p value (in controls) of $< 5 \times 10^{-5}$ were removed.

Sample QC was performed in a similar way to in chapter 2 but with an additional step subsequently to remove related individuals. As in chapter 2, variants were pruned to an LD $r^2 < 0.2$ and the HLA region was removed (chromosome 6, positions 25,000,000 - 35,000,000, genome build 36) so a relatively independent set of variants were used to determine relatedness between individuals. The same formula to estimate relatedness as in chapter 2 was used (Equation 2.2) and the KING software was used to carry out these analyses [125]. Individuals related to 2nd degree or higher were removed from the analysis. Specifically, one individual in a pair (i, j) where $\hat{\phi}_{i,j} > 0.0884$ was removed from the analysis at random.

Since reported sex data were missing for a proportion of individuals, I inferred sex for these individuals by assuming any with an X-chromosome homozygosity rate between

0.65 and 0.85 were female and any with a homozygosity rate of > 0.93 were male. This increased sample size since sex was included as a covariate in all analyses.

After sample QC, there were 27,071 individuals included in the analysis: 18,485 controls, 3121 diagnosed at < 7 years, 3757 diagnosed at $7 - 13$ years, and 1708 diagnosed at ≥ 13 years. Table 3.1 shows the baseline characteristics of individuals included, by AAD group.

The same pruned set of genotype data was used to calculate PCs from the matrix of genotype calls, using standard PCA as all individuals were unrelated.

Table 3.1: Baseline characteristics of individuals included in the analysis.

	Controls	< 7	$7 - 13$	≥ 13
N	18485	3121	3757	1708
Mean age-at-diagnosis	NA	3.67	9.54	17.95
Female	9771 (52.9%)	1528 (49%)	1968 (52.4%)	743 (43.5%)
Asia-Pacific	925 (5%)	28 (0.9%)	26 (0.7%)	52 (3%)
European	1681 (9.1%)	52 (1.7%)	63 (1.7%)	156 (9.1%)
Finland	2824 (15.3%)	100 (3.2%)	152 (4%)	434 (25.4%)
Northern Ireland	478 (2.6%)	222 (7.1%)	248 (6.6%)	35 (2%)
UK	10593 (57.3%)	2662 (85.3%)	3206 (85.3%)	923 (54%)
USA	1984 (10.7%)	57 (1.8%)	62 (1.7%)	108 (6.3%)

3.2.2 Loci studied

I examined each HLA haplotype/classical allele and non-HLA variant that has previously shown association to T1D. The HLA alleles most strongly associated with T1D are defined in [141] and Table 3.2 summarises the findings from this paper and shows how many individuals in each AAD group have these class II haplotypes or class I classical alleles. Due to the strength of association between the DR3-DQ2/DR4-DQ8 diplotype and T1D, when examining the effect of the other class II haplotypes, I excluded all cases and controls with this diplotype to remove any confounding effects; the numbers in Table 3.2 are the numbers after removing these individuals (in all except the row containing DR3-DQ2/DR4-DQ8).

In addition to the HLA alleles, there were, at the time, 55 regions on the ImmunoChip outside the HLA region that had been associated with T1D. Using various publications,

Table 3.2: HLA allele definitions and number of individuals included in the analysis, with each HLA haplotype/classical allele. HLA class II haplotypes exclude all individuals with the DR3-DQ2-DR4-DQ8 diplotype.

Haplotype/Diplotype	Definition	N control	N<7	N $7-13$	N≥ 13
DR3-DQ2	DRB1*03:01-DQB1*02:01	3920	566	687	319
DR4-DQ8	DRB1*04:01-DQB1*03:02	3025	569	775	417
	DRB1*04:02-DQB1*03:02				
	DRB1*04:04-DQB1*03:02				
	DRB1*04:05-DQB1*03:02				
	DRB1*04:08-DQB1*03:02				
	DRB1*04:01-DQB1*03:04				
	DRB1*04:02-DQB1*03:04				
	DRB1*04:04-DQB1*03:04				
	DRB1*04:05-DQB1*03:04				
	DRB1*04:08-DQB1*03:04				
	DRB1*04:01-DQB1*02:02				
	DRB1*04:02-DQB1*02:02				
	DRB1*04:04-DQB1*02:02				
	DRB1*04:05-DQB1*02:02				
	DRB1*04:08-DQB1*02:02				
DR3-DQ2/DR4-DQ8	Row 1 and 2 combination	752	1163	1156	404
DRB1*13:03-DQB1*03:01	DRB1*13:03-DQB1*03:01	137	4	4	4
DRB1*11:04-DQB1*03:01	DRB1*11:04-DQB1*03:01	153	4	1	0
DRB1*15:01-DQB1*06:02	DRB1*15:01-DQB1*06:02	3212	18	35	34
DRB1*07:01-DQB1*03:03	DRB1*07:01-DQB1*03:03	817	6	13	16
DRB1*14:01-DQB1*05:03	DRB1*14:01-DQB1*05:03	587	15	5	7
A*02:01	A*02:01	7196	1290	1624	766
A*24:02	A*24:02	1862	448	455	165
A*11:01	A*11:01	1103	114	147	64
A*32:01	A*32:01	648	79	96	42
DPB1*03:01	DPB1*03:01	3315	811	917	386
DPB1*04:02	DPB1*04:02	3299	275	339	172
B*18:01	B*18:01	1321	352	391	173
B*39:06	B*39:06	223	194	136	32
B*44:03	B*44:03	1173	107	132	68

I collated these associations and, where fine-mapping studies had identified more than one causal variant in a region, I included each of the signals. The nomenclature for the locus names begins with the name of a nearby or candidate gene, and if there is more than one disease-associated signal in the region, is followed by a number indicating which signal. The variants listed are from the most recent publication identifying these regions and their association with T1D, most of which are from Onengut-Gumuscu [144]. The interleukin 2 (*IL2*)/interleukin 21 (*IL21*) region has two signals, identified in [31], the *THEMIS* region was identified in chapter 2 and published in [89]. Finally, the interleukin 2 receptor (*IL2RA*) region uses the four signals identified in a fine-mapping analysis in [202]. Table 3.3 shows all non-HLA loci examined in this analysis.

Table 3.3: Non-HLA loci studied in the analysis.

Variant name	Chr	Position (gr 37)	Publication identified	Locus name in this thesis
rs2476601	1	114377568	Onengut-Gumuscu	PTPN22
rs78037977	1	172681031	Fortune	TNFSF4
rs6691977	1	200814959	Onengut-Gumuscu	CAMSAP2
rs3024505	1	206939904	Onengut-Gumuscu	IL10
rs4849135	1	111615079	Onengut-Gumuscu	ACOXL
rs13415583	2	100764087	Onengut-Gumuscu	AFF3
rs2111485	2	163110536	Onengut-Gumuscu	IFIH1 (1)
rs35667974	2	163124637	Onengut-Gumuscu	IFIH1 (2)
rs72871627	2	163136942	Onengut-Gumuscu	IFIH1 (3)
rs3087243	2	204738919	Onengut-Gumuscu	CTLA4
rs113010081	3	46457412	Onengut-Gumuscu	CCR5
rs6819058	4	123114622	Burren	IL2/IL21 (1)
rs67797421	4	123116177	Burren	IL2/IL21 (2)
rs2611215	4	166574267	Onengut-Gumuscu	CPE
rs11954020	5	35883251	Onengut-Gumuscu	IL7R
rs72975913	6	128293932	Inshaw	THEMIS

rs72928038	6	90976768	Onengut-Gumuscu	BACH2
rs1538171	6	126752884	Onengut-Gumuscu	CENPW
rs62447205	7	50465830	Onengut-Gumuscu	IKZF1
rs10277986	7	51028987	Onengut-Gumuscu	COBL
rs6476839	9	4290823	Onengut-Gumuscu	GLIS3
rs61839660	10	6094697	Wallace	IL2RA (1)
rs11594656	10	6122009	Wallace	IL2RA (2)
rs6602437	10	6130077	Wallace	IL2RA (3)
rs41295121	10	6129643	Wallace	IL2RA (4)
rs12416116	10	90035654	Onengut-Gumuscu	PTEN
rs689	11	2182224	Onengut-Gumuscu	INS (1)
rs72853903	11	2198665	Onengut-Gumuscu	INS (2)
rs917911	12	9905851	Onengut-Gumuscu	CD69
rs705705	12	56435504	Onengut-Gumuscu	IKZF4
rs653178	12	112007756	Onengut-Gumuscu	SH2B3
rs9585056	13	100081766	Onengut-Gumuscu	GPR183
rs1456988	14	98488007	Onengut-Gumuscu	LINC01550
rs56994090	14	101306447	Onengut-Gumuscu	MEG3
rs72727394	15	38847022	Onengut-Gumuscu	RASGRP1
rs34593439	15	79234957	Onengut-Gumuscu	CTSH
rs151234	16	28505660	Onengut-Gumuscu	IL27
rs12927355	16	11194771	Onengut-Gumuscu	DEXI (1)
rs193778	16	11351211	Onengut-Gumuscu	DEXI (2)
rs8056814	16	75252327	Onengut-Gumuscu	CTRB1
rs12453507	17	38053207	Onengut-Gumuscu	IKZF3
rs757411	17	38775150	Onengut-Gumuscu	CCR7
rs1052553	17	44073889	Onengut-Gumuscu	MAPT
rs1893217	18	12809340	Onengut-Gumuscu	PTPN2 (1)
rs12971201	18	12830538	Onengut-Gumuscu	PTPN2 (2)

rs1615504	18	67526644	Onengut-Gumuscu	CD226
rs34536443	19	10463118	Onengut-Gumuscu	TYK2 (1)
rs12720356	19	10469975	Onengut-Gumuscu	TYK2 (2)
rs402072	19	47219122	Onengut-Gumuscu	PRKD2
rs516246	19	49206172	Onengut-Gumuscu	FUT2
rs6043409	20	1616206	Onengut-Gumuscu	SIRPG
rs11203202	21	43825357	Onengut-Gumuscu	UBASH3A
rs6518350	21	45621817	Onengut-Gumuscu	ICOSLG
rs4820830	22	30531091	Onengut-Gumuscu	HORMAD2
rs229533	22	37587111	Onengut-Gumuscu	C1QTNF6

3.2.3 Logistic regression

Logistic regression is introduced briefly here from the perspective of a genetic case-control study with a binary outcome, $\mathbf{y}$, N individuals and K predictors potentially associated with the probability of being a case, π or a control $(1-\pi)$. Logistic regression models the probability of being case relative to the probability of not being a case, using linear predictors, $\mathbf{X}$, and their estimated effects on the log-odds of being a case, $\boldsymbol{\theta}$:

$$log\left(\frac{\boldsymbol{\pi}}{1-\boldsymbol{\pi}}\right) = \mathbf{X}\boldsymbol{\theta}. \tag{3.1}$$

Here, $\boldsymbol{\theta}$ is $K \times 1$ vector of effect estimates of K predictors on the log-odds of being a case and $\mathbf{X}$ is an $N \times K$ matrix of dependent variable values including an $N \times 1$ dummy vector, $\mathbf{1}$, so the intercept, or baseline odds, can be calculated.

This leads to the following definition of case probability:

$$\boldsymbol{\pi} = \frac{e^{\mathbf{X}\boldsymbol{\theta}}}{1+e^{\mathbf{X}\boldsymbol{\theta}}}. \tag{3.2}$$

The probability of the observed outcome, y_i, for individual i, given the value of their covariates, $\mathbf{x_i}$, is modelled as a Bernoulli function where the value of 1 has probability π

and the value 0 has probability $(1 - \pi)$:

$$p(y_i|\mathbf{x_i}, \boldsymbol{\theta}) = \pi_i^{y_i}(1 - \pi_i)^{1-y_i}. \tag{3.3}$$

The probability across all N individuals is the product of all these probabilities in Equation 3.3:

$$p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}) = \prod_{i=1}^{N} \pi_i^{y_i}(1 - \pi_i)^{1-y_i}. \tag{3.4}$$

Substituting Equation (3.2) into Equation (3.3) and instead of examining the probability of the outcome, calculating the probability of the data given the unknown parameters, otherwise known as the likelihood function, gives:

$$L(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X}) = \prod_{i=1}^{N} \left(\frac{e^{\mathbf{x_i}\boldsymbol{\theta}}}{1 + e^{\mathbf{x_i}\boldsymbol{\theta}}} \right)^{y_i} \left(1 - \frac{e^{\mathbf{x_i}\boldsymbol{\theta}}}{1 + e^{\mathbf{x_i}\boldsymbol{\theta}}} \right)^{1-y_i}. \tag{3.5}$$

The likelihood is then maximised with respect to $\boldsymbol{\theta}$. This is done by differentiating the likelihood function with respect to $\boldsymbol{\theta}$ and equating this to zero, which is easier to do once taking the log of the likelihood function:

$$logL(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X}) = \sum_{i=1}^{N} y_i log \left(\frac{e^{\mathbf{x_i}\boldsymbol{\theta}}}{1 + e^{\mathbf{x_i}\boldsymbol{\theta}}} \right) + (1 - y_i) log \left(1 - \frac{e^{\mathbf{x_i}\boldsymbol{\theta}}}{1 + e^{\mathbf{x_i}\boldsymbol{\theta}}} \right). \tag{3.6}$$

This can be partially differentiated with respect to each element of $\boldsymbol{\theta}$, and the values of $\boldsymbol{\theta}$ that jointly maximise the likelihood are obtained using Newton's method [215] with the resulting partial derivatives. Typically, in genetic case-control studies, the genetic variant under consideration is included in the model as a predictor to obtain the log-odds ratio estimate of the genetic variant on risk of T1D, with the remaining coefficients containing the estimated effects of the other coefficients included in the model, for example sex or genetic PCs.

3.2.4 Multinomial logistic regression

Multinomial logistic regression can be thought of as an extension to logistic regression where there are more than two possible outcomes. In general, T different outcomes 0, 1, 2, ... , T can be considered. Multinomial logistic regression is similar to fitting $T-1$ logistic regressions, all compared to a control group (group 0), but the denominator of the likelihood is such that the overall probability of an individuals being in category 0, 1 ... T sums to 1. The likelihood function for multinomial data is:

$$L(\boldsymbol{\theta}|\mathbf{y},\mathbf{X}) = \prod_{i=1}^{N} \pi_{i0}^{y_{i0}} \pi_{i1}^{y_{i1}} ... \pi_{iT}^{y_{iT}}. \tag{3.7}$$

However, the probability of being in each category is of most interest when in relative terms to a baseline category and therefore the likelihood is re-expressed, but the baseline category is derived as 1− the probability of being in all the other categories $\pi_0 = 1 - \sum_{t=1}^{T} \pi_t$. The likelihood function is then:

$$L(\boldsymbol{\theta}|\mathbf{y},\mathbf{X}) = \prod_{i=1}^{N} (1 - \sum_{t=1}^{T} \pi_{it})^{1-\sum_{t=1}^{T} y_{it}} \pi_{i1}^{y_{i1}} \pi_{i2}^{y_{i2}} ... \pi_{iT}^{y_{iT}} \tag{3.8}$$

By similar re-arrangement of the likelihood function as the logistic regression case, the probability of being in each group is now:

$$\begin{cases} \pi_{it} = \frac{e^{\mathbf{x_i}\theta_{it}}}{1+\sum_{t=1}^{T} e^{\mathbf{x_i}\theta_{it}}}, & \text{if } y_{it} > 0 \\ \pi_{i0} = \frac{1}{1+\sum_{t=1}^{T} e^{\mathbf{x_i}\boldsymbol{\theta_{it}}}}, & \text{if } y_{it} = 0. \end{cases} \tag{3.9}$$

Here, group 0 is the group everything else is compared to, and the likelihood function is now maximised in the same way as logistic regression with respect to K parameters for each $\boldsymbol{\pi_t}$.

In the present analysis, $T = 3$, as I divided the cases into AAD groups, those diagnosed at < 7, $7 - 13$ and ≥ 13 years old. I examined the effect of genetic variants on the probability of having T1D diagnosed at each of these age groups compared to controls. It would be of particular interest if the prevalence of the T1D risk allele at the variant

under consideration was higher in T1D individuals diagnosed at < 7 years compared to the older AAD groups.

3.2.5 Likelihood ratio test for heterogeneity in effect size between AAD groups

It is possible that the strongest effect of the genetic variant can be in the < 7 group due to chance, so a formal test was adopted from [133] to examine the likelihood of the model under a standard multinomial logistic regression where the effect size of the genetic variant in each of the AAD groups is unconstrained, and compare this to the likelihood of a multinomial logistic regression where the parameter estimate for the genetic variant in the < 7 and ≥ 13 groups were constrained to equal each other. The difference in the log-likelihoods between these models is χ^2 distributed with $K - P$ parameters under the null hypothesis of equal effect sizes between the two groups, where K is the number of parameters fitted in the unconstrained model and P is the number of parameters fitted in the constrained model. In each analysis therefore, I compared the difference in log-likelihood between the two models to a χ^2 distribution with 1 degree of freedom, since one less parameter was estimated in the second model compared to the first. I will refer to this test as the heterogeneity test.

Each multinomial regression was adjusted for sex and the ten largest PCs as calculated from the genotype data from individuals included in the analysis. Sex was included for the same reasons as in chapter 2, that there is a younger mean AAD in females compared to males in this dataset (Mean AAD females = 8.60, males = 8.99).

For analyses examining the HLA class II haplotypes and the HLA class I classical alleles, spurious associations could be identified because there is a large and complex LD structure in the region and the effect sizes of the associated alleles are large. Therefore alleles correlated with the associated HLA haplotypes/classical alleles could be artificially associated with disease risk. I attempted to account for this by adjusting for as many other T1D-associated classical HLA haplotypes/classical alleles as possible for the logistic regression algorithm to converge to a solution, namely the DR3-DQ2/DR4-DQ8 diplotype,

the DR3-DQ2 hapotype, the DR4-DQ8 haplotype, A*24:02 (when not examining HLA-A classical alleles) and B*39:06 (when not examining HLA-B classical alleles).

The strongest genetic association with T1D in the genome, the DR3-DQ2/DR4-DQ8 diplotype, is by definition correlated with both the DR3-DQ2 and DR4-DQ8 haplotypes. Therefore when examining these haplotypes, as well as the other class II haplotypes, in order to guard against spurious associations, all individuals with the DR3-DQ2/DR4-DQ8 diplotype were removed from the analyses so, for example, the effect of the DR3-DQ2 haplotype was examined by comparing the prevalence of the DR3-DQ2 haplotype in non DR3-DQ2/DR4-DQ8 T1D cases to non DR3-DQ2/DR4-DQ8 controls.

Since I examined 55 SNPs for heterogeneity in effect size, any SNPs with BH-FDR [19] < 0.1 were noted as being 'differentially associated' between the < 7 and ≥ 13 groups.

3.2.6 HLA imputation

To impute classical HLA alleles, the HiBag software was used [216], which implements a Random Forest algorithm. First a subset of individuals were selected from a dataset that had both ImmunoChip and classical HLA typing data available, and then a random subset of SNPs were selected from the HLA region. From these selection of individuals and SNPs, the optimal subset of SNPs associated with the classical allele of interest were identified. This was done hundreds of times and eventually the SNPs that best explain the HLA classical allele type were selected. These SNPs were then used to predict the classical HLA calls for individuals not included in the training dataset. A subset of the individuals in the present analysis had been directly genotyped for their HLA classical alleles. These individuals were used to assess the quality of the imputation by comparing the concordance of 4 digit directly genotyped HLA allele with the 4 digit imputed HLA prediction from the HiBag software. I used the same classifiers as were calculated from the original publication, as opposed to carrying out the Random Forest analysis myself [216].

3.2.7 Sensitivity analyses

Since the dataset consisted of individuals from different collections and different regions of the world (ancestry), and the distribution of individuals from each region of the world in each AAD group was not even, spurious differences in effect size between AAD groups might easily have been detected. Although I attempted to account for population structure in the data by adjusting for the ten largest genetic PCs, I performed some sensitivity analyses to improve confidence in the results at the non-HLA loci.

In the first sensitivity analysis, I restricted the analysis to only include individuals from the UK and Northern Ireland. This had the advantage of examining a relatively homogenous sample with regards to ancestry and therefore decreased the probability of observing spurious associations due to ancestry alone, but also decreased the statistical power to detect heterogeneity, since the number of individuals in the dataset from the UK and Northern Ireland diagnosed at ≥ 13 was just 958. PCs were calculated using the LD-pruned ImmunoChip data of the individuals included in the analysis, and models were adjusted for sex and the five largest PCs rather than ten as in the primary analysis, since the population was selected to be relatively homogeneous with respect to ancestry.

Secondly, to ensure that none of the loci were declared heterogeneous due to confounding effects of the highest risk HLA alleles, I repeated the analysis including all individuals, adjusting additionally for the most associated classical HLA alleles. These are class II haplotypes DR3-DQ2, DR4-DQ8, their diplotype, DR3-DQ2/DR4-DQ8, and class I classical alleles A*24:02 and B*39:06.

Thirdly, I tested the sensitivity of the results to the AAD cutoff value of 7 years, and instead compared effect sizes between those people diagnosed at < 6 years to those diagnosed at ≥ 13 years and also those diagnosed at < 5 years compared to those diagnosed at ≥ 13 years. This was to see whether or not loci were appearing differentially-associated between AAD groups by chance as a result of what threshold I defined the AAD groups at.

Finally, in order to test whether there were particular subsets of individuals in the dataset driving the significance of the heterogeneity tests at certain loci, I randomly divided cases

and controls from each ancestry group into two sets and performed the heterogeneity test at each locus on each set separately. I did this procedure 100 times and therefore performed 200 heterogeneity tests for each SNP included in the analysis. Given that the statistical power in each of these tests was substantially lower than the power of the primary analysis, instead of requiring that the genetic variant under consideration needed to pass FDR significance, I instead counted the proportion of times that the variant reached nominal ($p < 0.05$) significance, with a low proportion of nominally significant results at any of the loci indicating that the association in the primary analysis may have been due to a chance distribution of alleles creating a false positive heterogeneity test value between AAD groups.

3.2.8 Fine mapping

For regions that showed evidence of differential-association between AAD groups, I fine mapped the most likely causal variants underlying T1D risk diagnosed at < 7 years by comparing T1D cases diagnosed at < 7 to controls. I used the GUESSFM software to identify the most likely combinations of variants that best explain the phenotype, as introduced in chapter 2. To reduce the probability of identifying false positive variants associated with T1D diagnosed at < 7 years, I restricted this analysis to individuals from the UK or Northern Ireland in order to carry out fine mapping on an ancestry-homogeneous population. The difference here is that the outcome variable is binary, but the outcome is assumed to be Gaussian, which is an approximation that is reasonable with large sample sizes with minimal population structure [39]. Specifically, I initially fitted a linear model with T1D status (diagnosed < 7 years) as the outcome, adjusting for sex and the five largest genetic PCs, then took the residuals from this model, using them as the outcome for the stochastic search. The same approximation that the binary outcome is linear was assumed in the post-processing step, where approximate Bayes factors (ABFs) are calculated from the bayesian information criterions (BICs): the outcome was taken as the residuals from a linear regression including sex and the top five PCs, with T1D as the outcome. This is done since the number of models to fit can be very large,

and logistic regression models are slower to fit than linear regression models. However, to compare the BICs, the top 100 ranking models fitted using linear regression were compared to the top ranking 100 models fitted using logistic regression models to assess the reliability of the approximation. Similar BICs between the two methods implies the linear approximation is reasonable, since the ABFs are calculated from the BIC and used to generate downstream posterior probabilities.

To obtain a more dense set of variants for which to carry out the fine mapping on, and thereby increase the probability of including the true (unknown) causal variant in the analysis, I imputed a 0.5 Mb region around the index variant using IMPUTE2, as in chapter 2, using the 1000 Genomes Project data as a reference panel of haplotypes. Then, in order to guard against identifying false positive disease-associated variants owing to poor imputation quality, a stringent post-imputation quality control procedure was performed. Firstly, variants with an overall imputation score of < 0.8, an overall minor allele frequency of < 0.005 or an absolute HWE exact test z-score in controls only of > 8 were removed. Next, the association of each variant with T1D diagnosed at < 7 years was calculated using imputed dosages, run using SNPTEST [127], which provides the information score for each variant amongst cases and controls only. Using these information scores, I removed variants where the difference in information score between cases and controls was $> 5\%$, since this could drive an artificial association with these variants.

Before carrying out the stochastic search, variants were pruned to an r^2 of 0.99. The top 10,000 most visited models were saved for downstream analysis. The prior for the number of variants associated with the phenotype was a binomial prior centred at 3 causal variants.

3.2.9 Colocalisation

In regions that had evidence of differential-association between AAD groups, I tested each variant that passed imputation QC for an association with T1D diagnosed at < 7 years compared to controls using logistic regression. I then took association summary statistics

for these variants from an eQTL analysis in whole blood of over $30,000$ individuals [199]. Summary statistics from both association studies (disease risk at < 7 years and eQTL) were converted to ABFs. ABFs were calculated with the following formula, first derived in [201].

$$ABF = \sqrt{1-r} \times e^{\frac{Z^2}{2} \times r}, \tag{3.10}$$

where Z the association z-score with T1D diagnosed at < 7 years and r is the ratio of the prior variance to the total variance, where the total variance, V, is $\frac{1}{(2Nf(1-f))}$ for quantitative traits and V is $\frac{1}{(2Nf(1-f)s(1-s))}$ for binary traits, with f the MAF of the variant under consideration, N the number of samples and s the proportion of samples that are cases in the dataset. The prior variance is taken to be 0.15^2 for the eQTL analysis and 0.2^2 for the disease variance, thus $r = \frac{0.15^2}{0.15^2+V}$ for eQTL associations and $r = \frac{0.2^2}{0.2^2+V}$ for disease associations. These priors are suggested by the authors [72] and are consistent with priors used in historical Bayesian analyses of genetic case-control data [182].

Values proportional to posterior probabilities are then calculated for each variant for each of the following scenarios, by multiplying the prior by the ABF under:

i) association with disease only

ii) association with eQTL only

iii) association with disease and eQTL but distinct signals or

iv) shared disease and eQTL signal in the region.

This follows methods introduced in [72]. For each variant, the prior $\times$ the ABF is calculated for each scenario, then the product across all variants is taken for each scenario to obtain a value proportional to the posterior for each scenario. Finally, to obtain the posterior probability for each scenario, the product (across all variants) of the prior $\times$ the ABF for each scenario is divided by the normalising factor, which is the product of the prior $\times$ the ABF for each variant across all scenarios. For clarity, in a region containing M variants (using the logarithm of each value so the sum can be taken rather than the product), first the value which is proportional to the posterior probability is calculated (called val), for example for scenario 1:

$$Posterior(scenario_1) \propto \Sigma_{i=1}^{M}\ log(prior_{scenario\ 1}) + log(ABF_i) = val_1 \tag{3.11}$$

Then, this value is divided by the sum across all scenarios to obtain the posterior probability for each scenario.

$$Posterior(scenario_1) = \frac{val_1}{\Sigma_{j=0}^{4} val_j} \tag{3.12}$$

The main focus is on scenario iv), colocalisation of disease and eQTL signal. I set the prior probability of eQTL and disease association at 0.0001, and the prior probability of colocalisation of 0.00001. This reflects a prior belief that approximately one in ten variants that are associated with trait 1 will also be associated with trait 2. The 'coloc' R package was used to carry out this analysis.

For both the fine mapping and colocalisation analyses, only individuals from the UK or Northern Ireland were included in order to use a relatively ancestry-homogeneous population, since both these analyses are sensitive to heterogeneity in ancestry. In addition, only regions with evidence of a single causal variant in the region were examined for colocalisation, since this is an underlying assumption of the colocalisation method.

3.2.10 Heritability analyses by age-at-diagnosis group

Genetic heritability analyses were originally conducted using monozygotic (MZ, which share approximately 100% of their variable DNA) and dizygotic (DZ, which share approximately 50% of their variable DNA) twins. If MZ twins have a higher correlation of a particular trait (r_{mz}) than DZ twins (r_{dz}), this implies more of the trait can be explained by genetics rather than the environment [23], although caution is required in interpretation of these results, since MZ twins may also share more environmental factors than DZ twins. The heritability estimate from twin studies is defined as:

$$h^2_{twin} = 2 \times (r_{mz} - r_{dz}) \tag{3.13}$$

for quantitative traits. For binary traits, the binary outcome is transformed onto the

liability scale and a similar analysis is carried out [60]. In T1D, twin heritability estimates have been estimated as 88% [87], with higher concordance amongst MZ twins if one of the twins was diagnosed in childhood rather than after 25 years [159, 158].

However, as with many other traits, the heritability estimates using case-control data, as opposed to family or twin data, for T1D are much lower at approximately 32% [108, 180], but has been reported at much higher levels for paediatric T1D [112]. Here, I employ the case-control heritability method genomic relatedness restricted maximum likelihood (GREML), employed using the genome-wide complex trait analysis (GCTA) software [212], which can estimate the heritability explained by SNP data using unrelated individuals. To examine whether there is a higher heritability in individuals diagnosed at a young age as opposed to older, I stratified the cases as in the multinomial logistic regression analyses and fitted models including only cases diagnosed at the AAD group of interest compared to all controls. I repeated the analysis but excluded the HLA region, since this contains variants with very large effect sizes and are in LD with many variants; examining heritability estimates excluding this region would provide evidence that any difference in estimated heritability by AAD group is not driven solely by HLA associations in T1D.

Genomic relatedness restricted maximum likelihood (GREML)

The GRM between individuals is first calculated as in equation 2.34, since the GRM is required to be positive semi-definite in downsteam analysis, which this calculation ensures. I did not prune the matrix prior to calculation of the GRM as recommended by the authors of the original GCTA publication and follow-up paper [211], since the heritability estimate of interest is that of all SNPs in combination, hence accounting for LD.

The premise of using a linear mixed model to explain heritability is similar to that of the twin studies in as much that phenotypic similarity (concordance) is assessed against similarity in genotype (100% or 50% similar for MZ and DZ, respectively). Here, instead of twins there are a set of unrelated cases and controls and the aim is to examine phe-

notypic similarity between these individuals and see whether this corresponds to overall genetic similarity between the individuals.

The linear model fitted is the following:

$$\mathbf{y} = \mathbf{u} + \mathbf{e} \tag{3.14}$$

Where $\mathbf{y}$ is the T1D status, modelled as a combination of risk from all variants on the ImmunoChip, $\mathbf{u} \sim N(\mathbf{0}, \mathbf{R}\sigma_g^2)$, otherwise known as the genetic liability, and any additional variance, $\mathbf{e} \sim N(0, \mathbf{I}\sigma_e^2)$. Here, $\mathbf{R}$ is the estimated GRM between all individuals which is assumed fixed once estimated using equation 2.34.

However, this model is generated on the observed scale, where T1D status is either 0 (no) or 1 (yes). Modelling in this way and taking σ_g^2 as the heritability estimate would lead to misleading estimates, as our case-control design does not reflect the true prevalence of T1D in the population. Instead the design is such that there is a large over-sampling of cases in our analysis (termed 'ascertainment bias') and thus heritability estimation without taking this into account is inappropriate.

The outcome is transformed so that instead of having a binary outcome, each individual has an unobserved continuous risk of T1D, termed a liability of disease. The theory is that if an individual's liability exceeds a particular threshold, t, they are a case, if it is less than t, they are a control. Viewing the outcome in this way and transforming the genotype effects to also reflect their effect on the liability scale leads to a correction factor applied to the heritability estimate from Equation 3.14 [108]:

$$h_l^2 = \frac{\sigma_g^2 K(1-K)}{z^2}, \tag{3.15}$$

where K is the prevalence of the disease under consideration, which is approximately 0.4% for T1D, and z is the proportion of individuals with a liability threshold $> t$.

In the present analysis, I included fixed effects of the ten largest genetic PCs and sex in the model in equation 3.14 to account for any additional variance of these factors. In addition, I examined heritability estimates under different T1D prevalence assumptions,

since estimates of prevalence, particularly by AAD, can be imprecise.

3.2.11 Prediction

To examine whether or not the loci most associated wth early-diagnosed T1D could predict early-diagnosis, I took all FDR < 0.1 differentially-associated loci (both within and outside the HLA) and built a genetic risk score (GRS) using the following equation, for each variant, i, and for each individual j, with genotype g_{ij}:

$$GRS_j = \Sigma_i \hat{\beta}_i \times g_{ij}. \tag{3.16}$$

Here, $\hat{\beta}_i$ is the maximum likelihood estimate of the T1D log-odds ratio from the following logistic regression model (parameter β_{12}), including all individuals in the primary analysis:

$$log\left(\frac{\boldsymbol{\pi}}{1-\boldsymbol{\pi}}\right) = \alpha + \beta_1 \mathbf{PC1} + ... + \beta_{10}\mathbf{PC10} + \beta_{11}\mathbf{sex} + \beta_{12}.\mathbf{g_i} \tag{3.17}$$

I refer to this GRS as GRS_{fdr}. I built two other GRS's to compare to GRS_{fdr}, the first including all 55 T1D-associated variants as well as all associated HLA class II haplotypes and HLA class I classical alleles, which I refer to as GRS_{full} and the other including all T1D-associated variants, HLA class II haplotypes and HLA class I classical alleles that had the largest effect in the < 7 group, intermediate effect in the $7 - 13$ group and the smallest effect in the ≥ 13 group, referred to as GRS_{ord}.

Once the three GRS's were calculated for each individual, I fitted three logistic regression models including T1D cases only and the outcome was equal to 0 if the individual was diagnosed at < 7 years and 1 if the individual was diagnosed at ≥ 7 years. The same covariates were included in the model as in Equation 3.17 but without the genetic variants as covariates and including GRS_{fdr} in the first model, GRS_{ord} in the second model and GRS_{full} in the third model.

The fitted values from logistic regression models are bounded $(0, 1)$ and a Receiver Operator Characteristic (ROC) curve can be produced which shows how many individuals would be predicted to be cases or controls based on the threshold used between 0 and 1 to

declare an individual a case. The ROC curve shows the specificity and sensitivity of the classifcations based on varying the threshold from 0 to 1. For clarity, when the threshold t is 0, almost all individuals will have a predicted value of $> t$ and thus the sensitivity will be very high (all cases will be declared cases) but the specificity will be very low (also all controls will be declared cases), increasing the threshold t to be 1 would have the opposite effect, only cases with a predicted value of 1 will be declared cases, likely making the specificity very high, but many cases will be classified as controls, meaning the sensitivity will be very low.

The logistic regression models with the best ability to predict cases from controls will have high specificity and high sensitivity across all values of the threshold, t, so the predicted values of the controls will be close to 0 and the predicted values for the cases will be close to 1. Plotting the sensitivity against $1-$specificity at each value of t gives an idea of the overall predictive ability of the logistic regression, which is possible to calculate since the case and control status of each individual is known, as are the predicted values from the regression. The area under the curve (AUC) is a useful metric to summarise the ability of the model to determine cases from controls, with an AUC close to 1 indicating a good ability to classify cases from controls and a value of 0.5 indicating the model does no better than a random allocation of cases and controls.

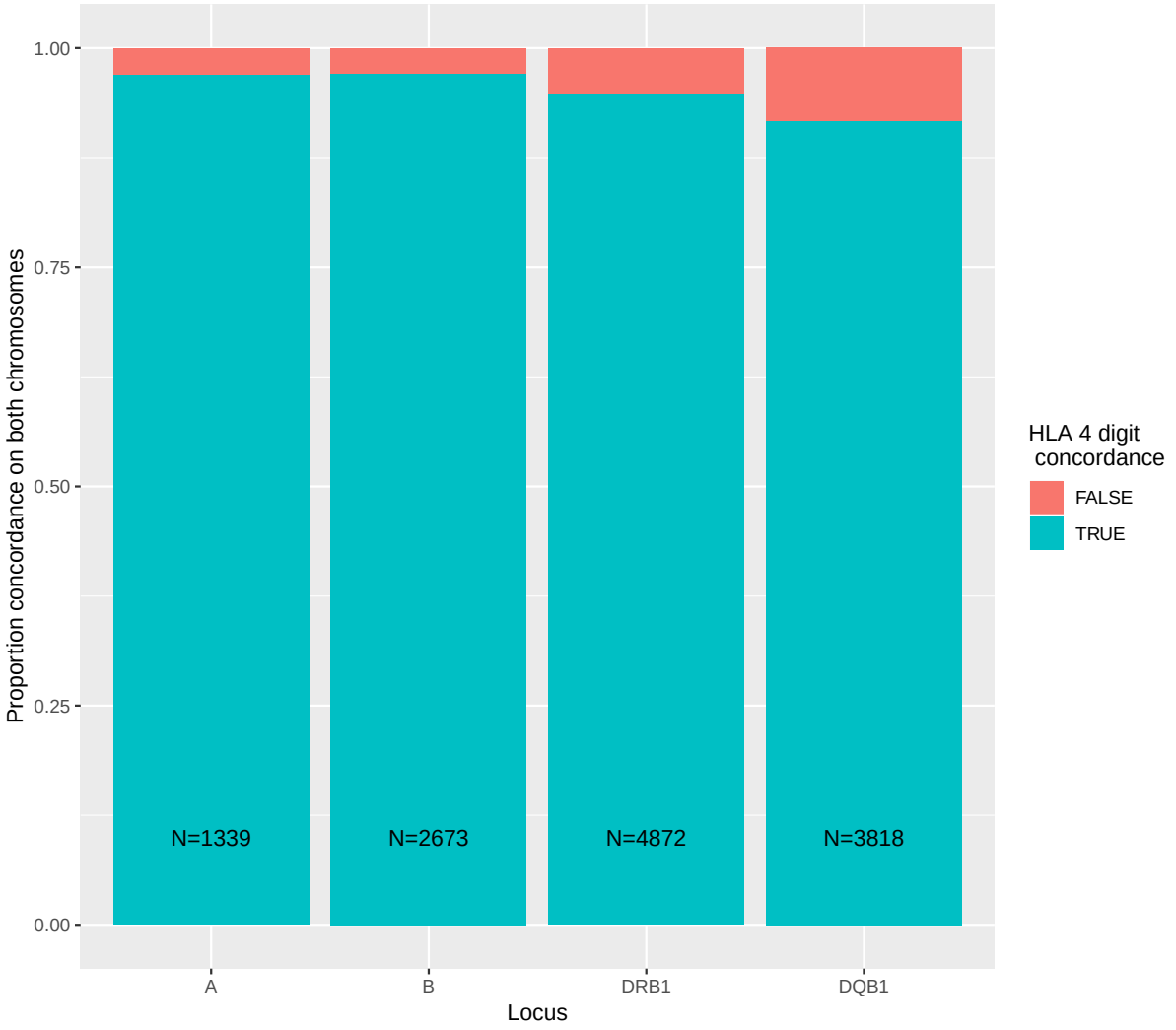
I compared the ROC curve AUC from logistic regression models including i) GRS_{fdr} ii) GRS_{ord} and iii) GRS_{full} as covariates to see how well the models predicted T1D diagnosed at < 7 years compared to ≥ 7 years.

3.3 Results

3.3.1 HLA results

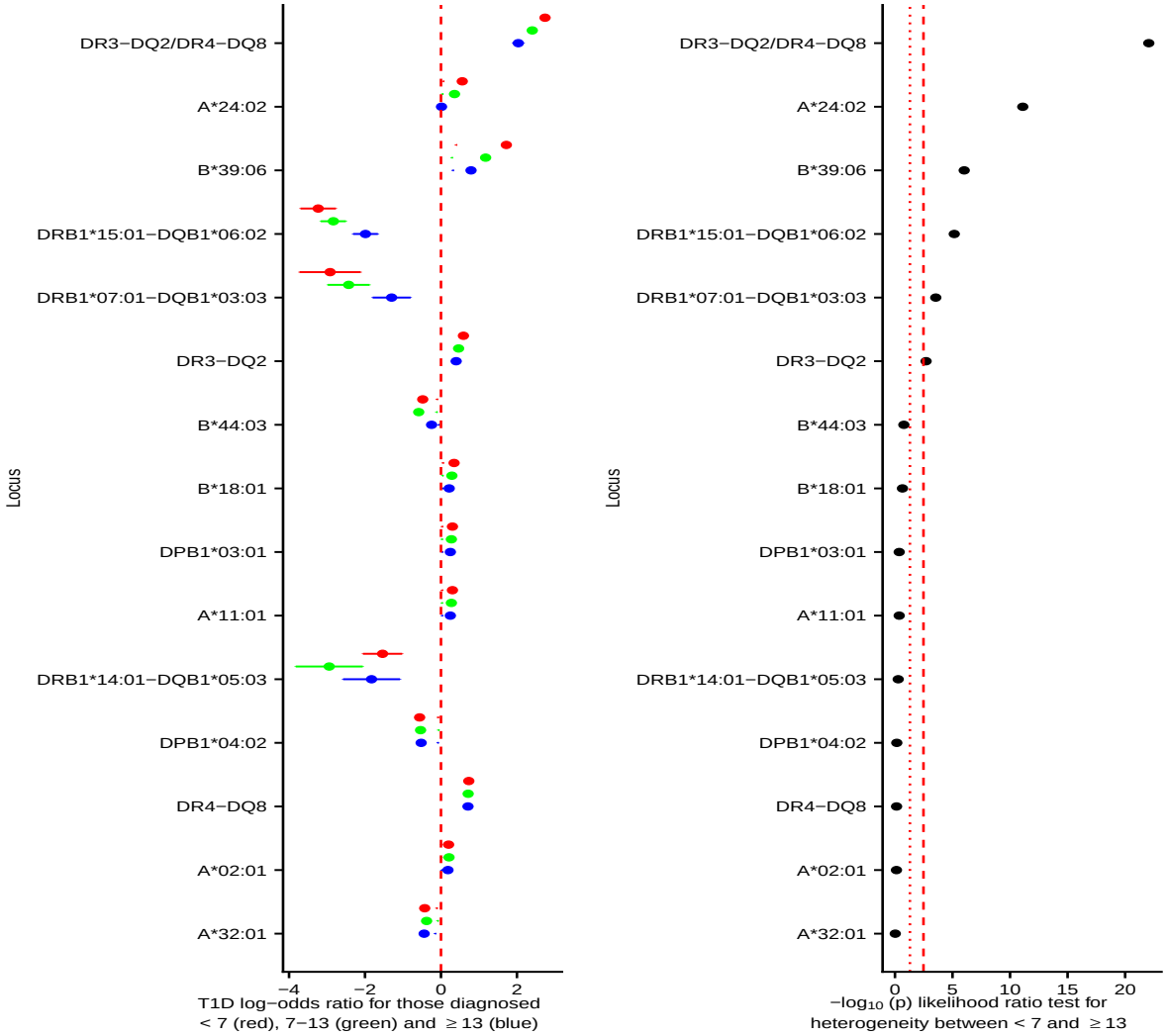
Prior to examining heterogeneity in T1D susceptibility effect size between the < 7 and ≥ 13 groups at each disease-associated haplotype/classical allele, it was important to check the quality of the HLA imputation. As a QC measure, classical alleles called with less than 0.5 posterior probability were treated as missing, then remaining alleles called with ≥ 0.5 posterior probability were compared against directly genotyped alleles in the proportion of individuals where the direct genotyping data was available. Exact concordance to 4 digit accuracy of over 90% was observed in the four genes used to form the classical class II haplotypes and the disease-associated class I alleles (Figure 3.1).

Figure 3.1: 4 digit HLA concordance between HiBag imputed classical alleles and directly genotyped for a proportion of individuals.



As has been shown previously, the strongest T1D-predisposing diplotype, DR3-DQ2/DR4-DQ8 was differentially-associated between the < 7 group and ≥ 13 group, despite the large effect size in the ≥ 13 group. This implies that carrying this diplotype increases risk of T1D at any age, but it is particularly prevalent in individuals diagnosed with T1D at a young age. The DR4-DQ8 haplotype had a similar effect across AAD groups, but the other class II haplotypes tested, the susceptible DR3-DQ2 and the protective DRB1*15:01-DQB1*06:02 and DRB1*07:01-DQB1*03:03 all had heterogeneous effects between AAD groups, and specifically stronger effects in the < 7 group compared to the ≥ 13 group. The observation that the two protective class II haplotypes have stronger effect sizes in the < 7 group compared to the ≥ 13 group has not been reported before. In addition, two class I classical alleles, A*24:02 and B*39:06, had stronger effects in the < 7 group compared to the ≥ 13 group (FDR< 0.1), which is expected as it replicates previous findings (Figure 3.2).

Figure 3.2: All T1D-associated classical HLA haplotypes/alleles examined for heterogeneity in effect size between the < 7 and ≥ 13 groups. Left panel: log-ORs for each AAD group, red dashed line indicating a log-OR of 0. Right panel: $-log_{10}$p value from heterogeneity test, red dotted line indicates a p value of 0.05 and the red dashed line indicates Bonferroni significance at the 5% level*.



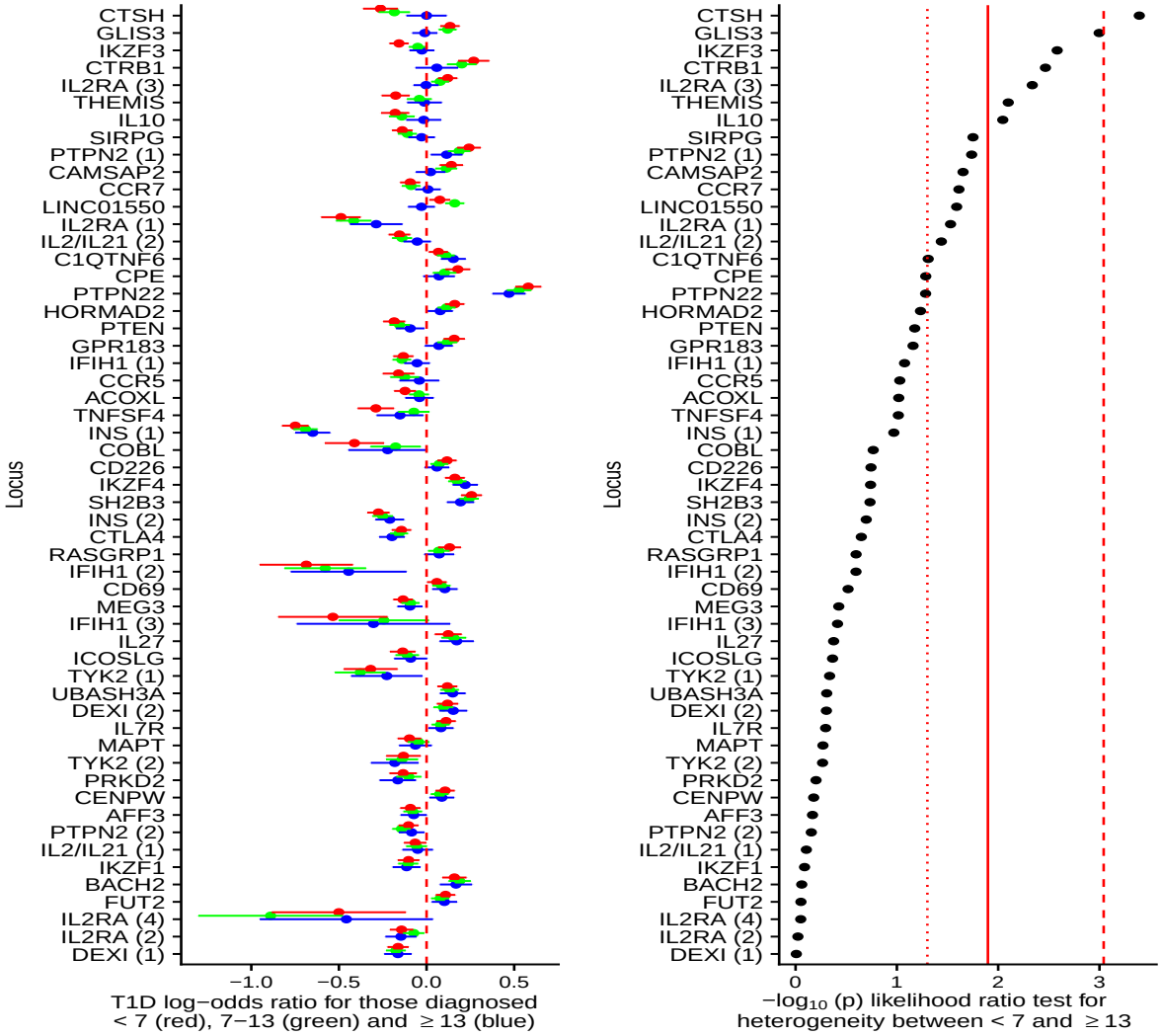
3.3.2 Non-HLA results

Outside the HLA, seven regions were differentially-associated between the < 7 and ≥ 13 groups (FDR$<$ 0.1), near Cathepsin H (*CTSH*), GLIS family zinc finger 3 (*GLIS3*), Ikaros family zinc finger 3 (*IKZF3*), Chymotrypsinogen B1 (*CTRB1*), the third index variant at interleukin 2 receptor alpha (*IL2RA*), *PTPRK/THEMIS* and Interleukin-10 (*IL10*) (Figure 3.3). One of these, *CTSH*, survived Bonferroni correction ($p < 0.05/55 = 0.00091$). At each locus associated with FDR$<$ 0.1, the $7-13$ group had a larger effect size than the ≥ 13 group and smaller than the < 7 group. Given the ≥ 13 group comprises

*Figure published and rights owned by Diabetes Care [88].

just 1,708 individuals, it is probable that with increased sample size and hence statistical power, other T1D risk loci with sizeable estimated effect size differences between groups might reach statistical significance with regards to heterogeneity. Of the 48 variants not satisfying an FDR< 0.1, 21 had the strongest signal in < 7 group, weakest in ≥ 13 group and intermediate in $7-13$ group. Given there are six possible orderings of log-odds ratios between the three groups, this implies if there were no difference in effect size between groups at any of these loci, $48/6 = 8$ loci would have the largest effect size in the < 7 group, intermediate effect size in the $7-13$ group and largest effect size in the ≥ 13 group by chance. The probability of observing 21 variants with this ordering given consistent effect sizes between the groups was 9.74×10^{-6}, as calculated from an exact binomial test, suggesting the presence of substantial additional signal in variants that were not declared as showing evidence individually.

Figure 3.3: All T1D-associated variants examined for heterogeneity in effect size between the < 7 and ≥ 13 groups. Left panel: log-ORs for each AAD group, red dashed line indicating a log-OR of 0. Right panel: $-log_{10}$p value from heterogeneity test, red dotted line indicates a p value of 0.05, red solid line indicates the FDR significance of 0.1, red dashed line indicates Bonferroni significance at the 5% level.

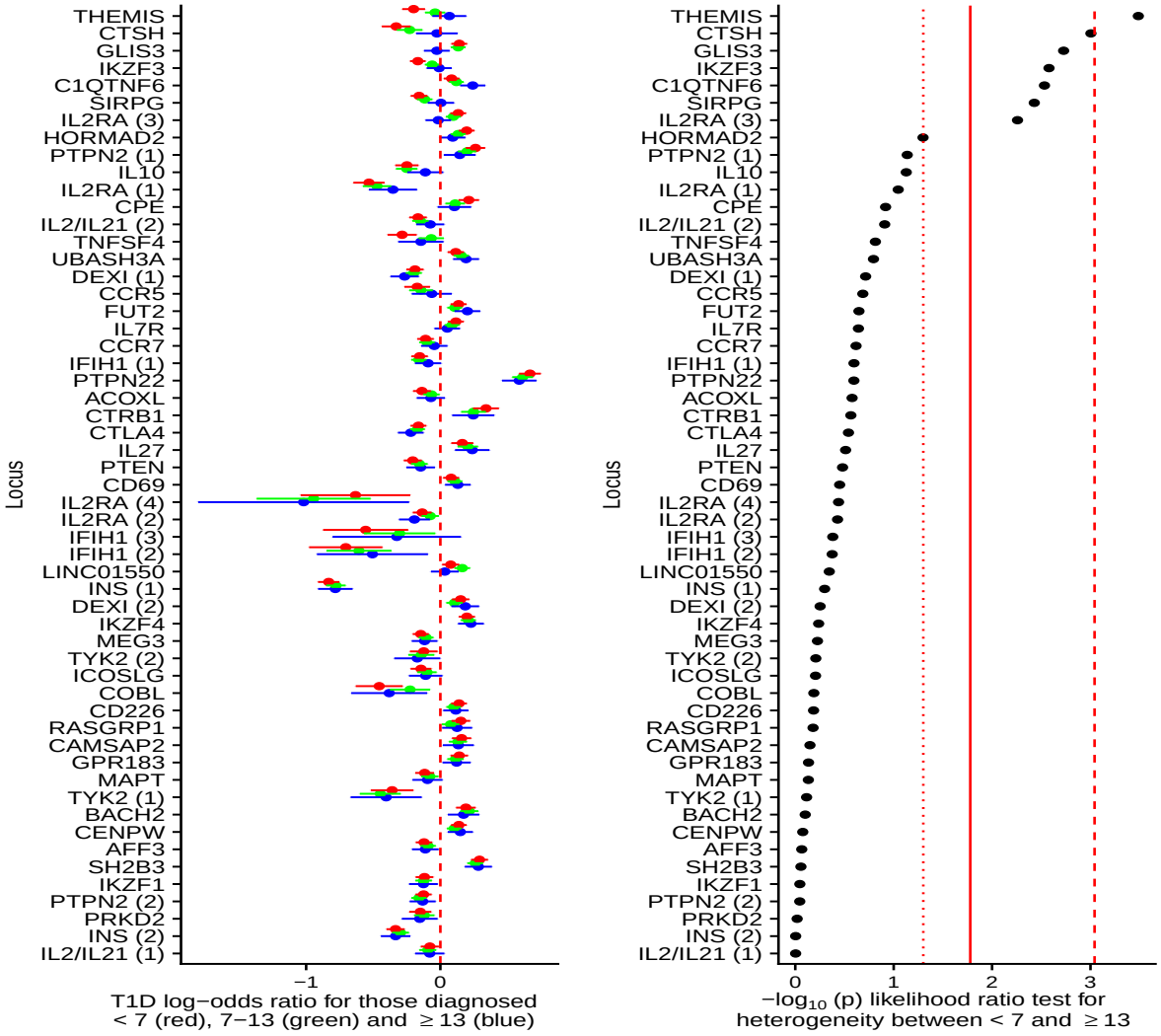


3.3.3 Sensitivity analyses

In the UK-specific analysis, five of the seven loci associated in the primary analyses had FDR< 0.1 (Figure 3.4). The *IL10* locus showed some evidence of association ($p = 0.08$) and so I kept this locus in the set of differentially-associated loci, given the decrease in sample size in this analysis. However, the *CTRB1* locus showed no evidence of association in this UK-based sensitivity analysis ($p = 0.272$) and so was removed from the set of differentially-associated loci. In addition, the region near Complement C1q Tumor Necrosis Factor-Related Protein 6 (*C1QTNF6*) showed a suggestion of heterogeneity be-

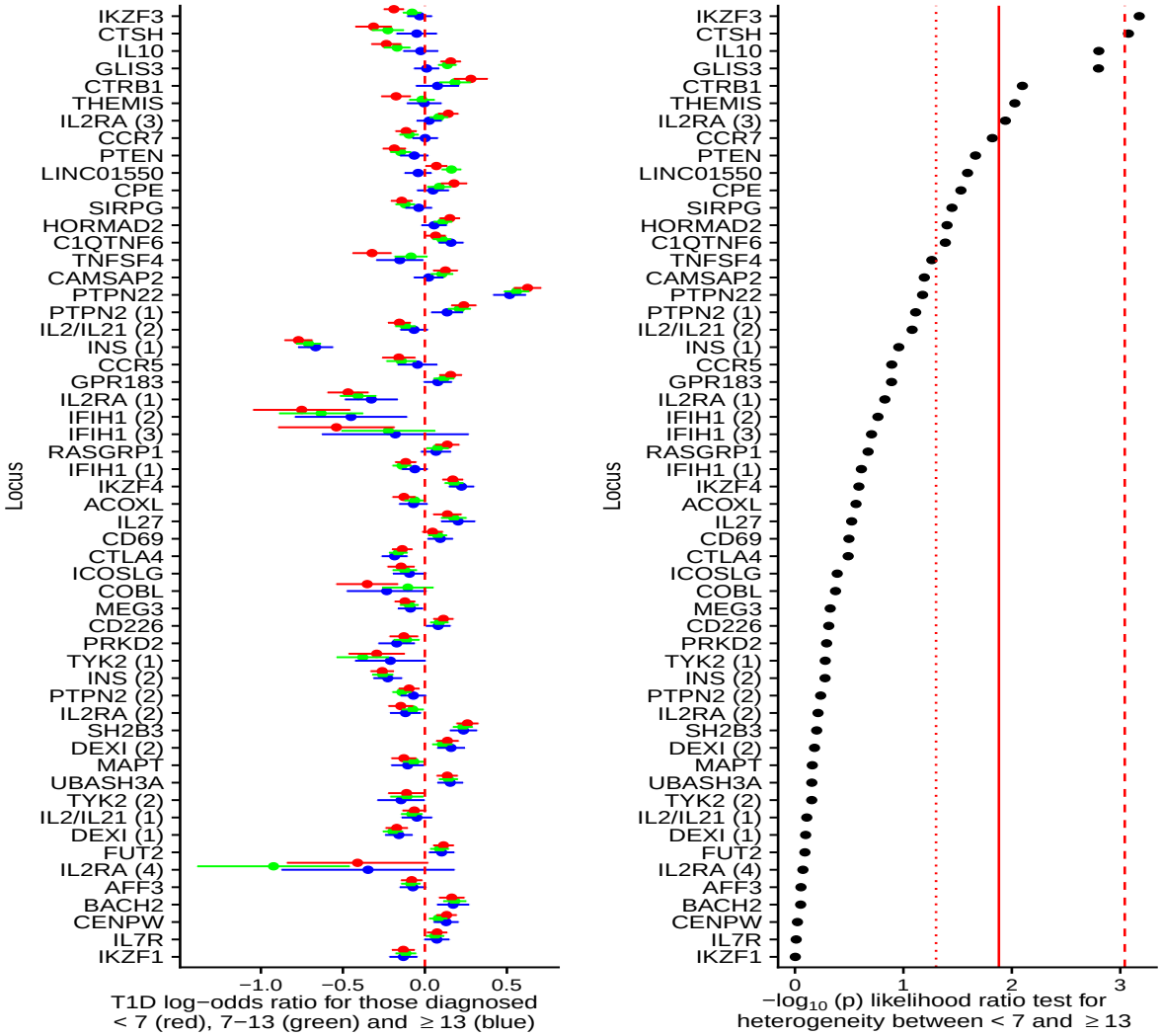
tween AAD groups but with a stronger effect in the ≥ 13 group. Since this is a sensitivity analysis, I did not consider it a robust finding and thus the *C1QTNF6* region was not added to the set of differentially-associated regions, although the effect direction is the same in the primary analysis, so it may be that with increased sample size this could be replicated.

Figure 3.4: All T1D-associated variants examined for heterogeneity in effect size between the < 7 and ≥ 13 groups examining UK-based individuals only. Left panel: log-ORs for each AAD group, red dashed line indicating a log-OR of 0. Right panel: $-log_{10}$p value from heterogeneity test, red dotted line indicates a p value of 0.05, red solid line indicates the FDR significance of 0.1, red dashed line indicates Bonferroni significance at the 5% level.



When adjusting for the most T1D-associated HLA haplotypes, all six of the differentially-associated loci from the primary and UK-specific analyses remained differentially-associated (FDR< 0.1), (Figure 3.5).

Figure 3.5: All T1D-associated variants examined for heterogeneity in effect size between the < 7 and ≥ 13 groups, adjusting additionally for the most T1D-associated HLA haplotypes/classical alleles. Left panel: log-ORs for each AAD group, red dashed line indicating a log-OR of 0. Right panel: $-log_{10}$p value from heterogeneity test, red dotted line indicates a p value of 0.05, red solid line indicates the FDR significance of 0.1, red dashed line indicates Bonferroni significance at the 5% level.



Likewise, when changing the AAD threshold to define early-diagnosed to < 6 (Figure 3.6) or < 5 (Figure 3.7), each of the six loci remained heterogeneous between AAD groups to FDR< 0.1.

Figure 3.6: All T1D-associated variants examined for heterogeneity in effect size between the < 6 and ≥ 13 groups. Left panel: log-ORs for each AAD group, red dashed line indicating a log-OR of 0. Right panel: $-log_{10}$p value from heterogeneity test, red dotted line indicates a p value of 0.05, red solid line indicates the FDR significance of 0.1, red dashed line indicates Bonferroni significance at the 5% level.

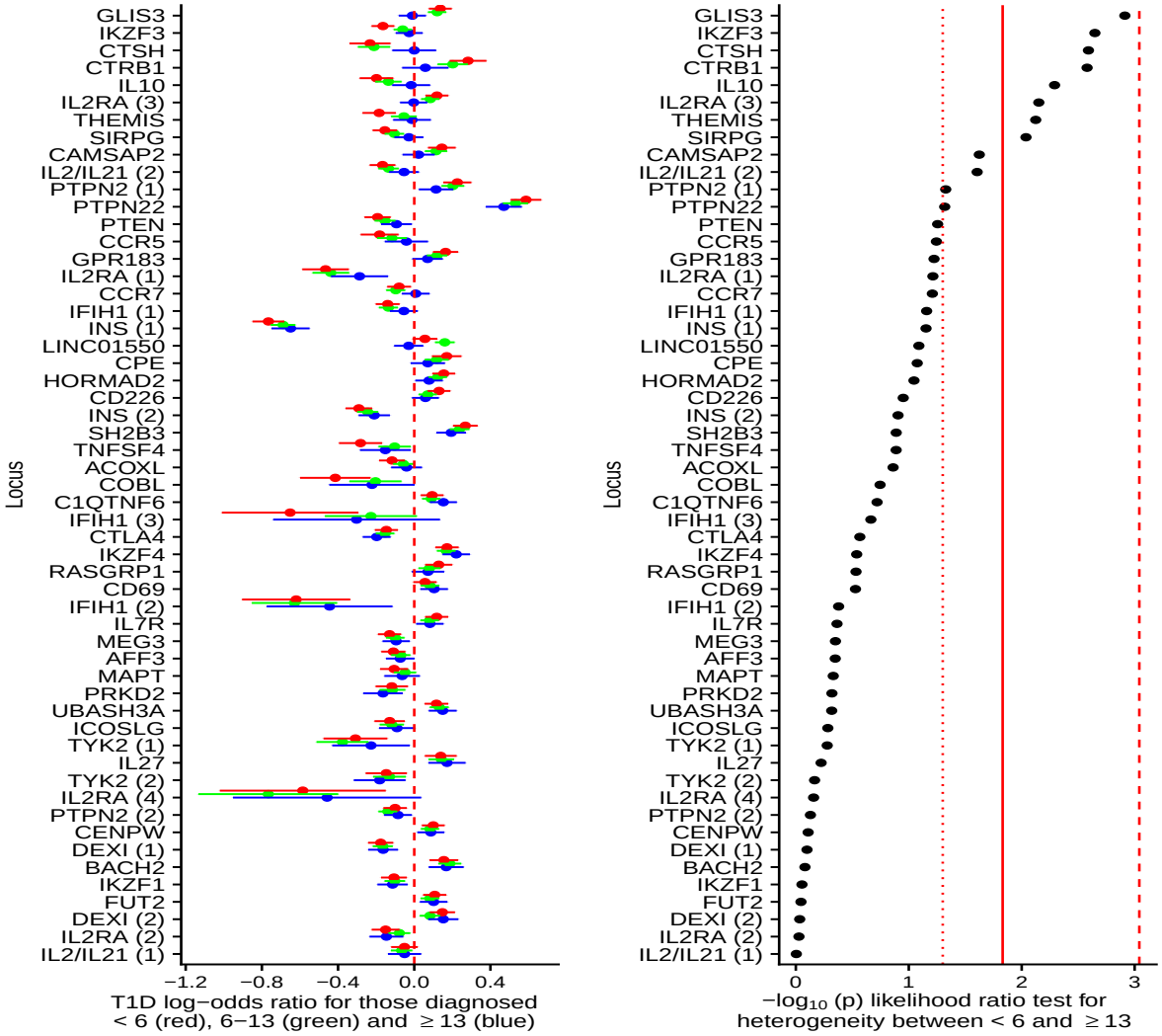
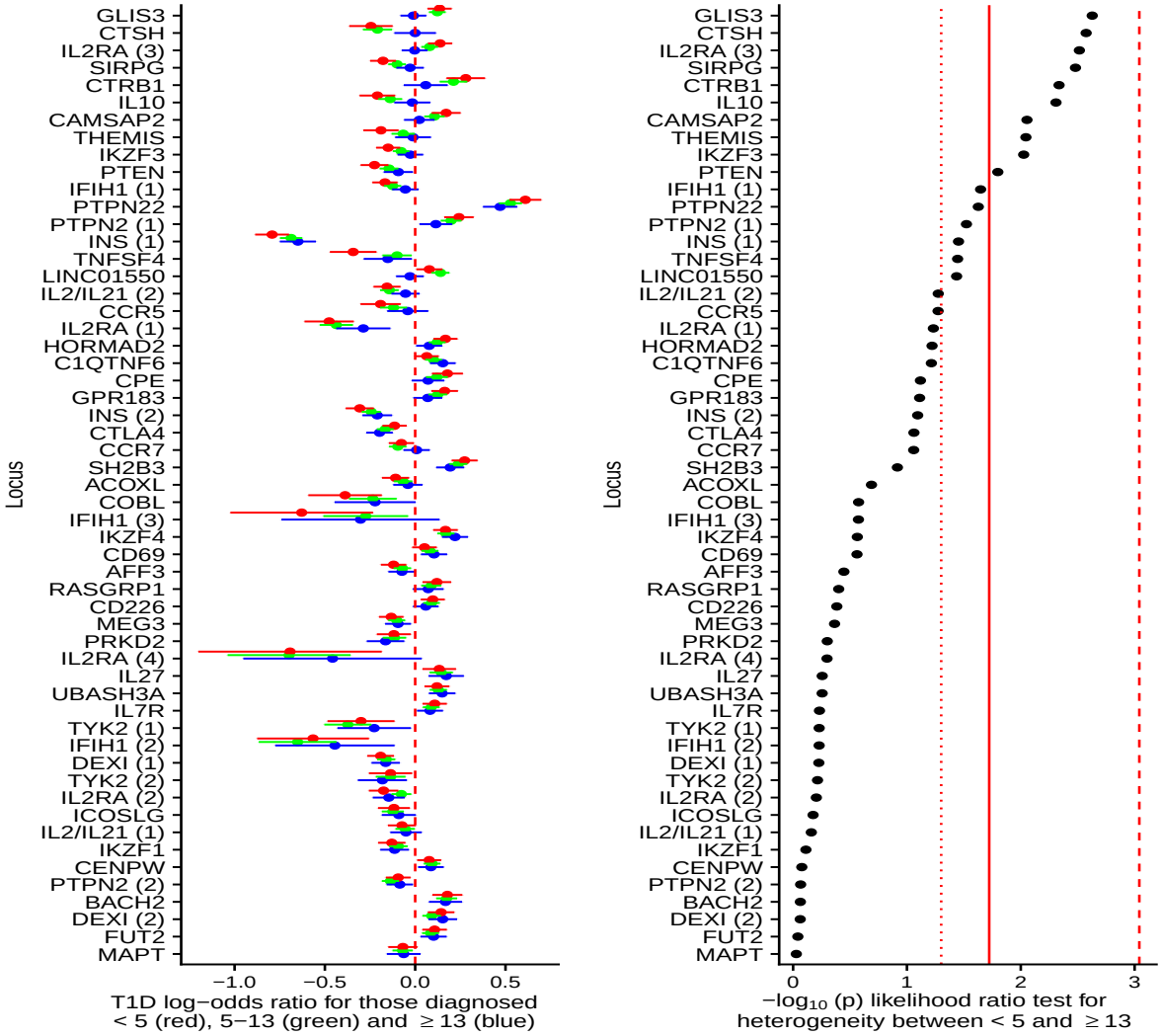


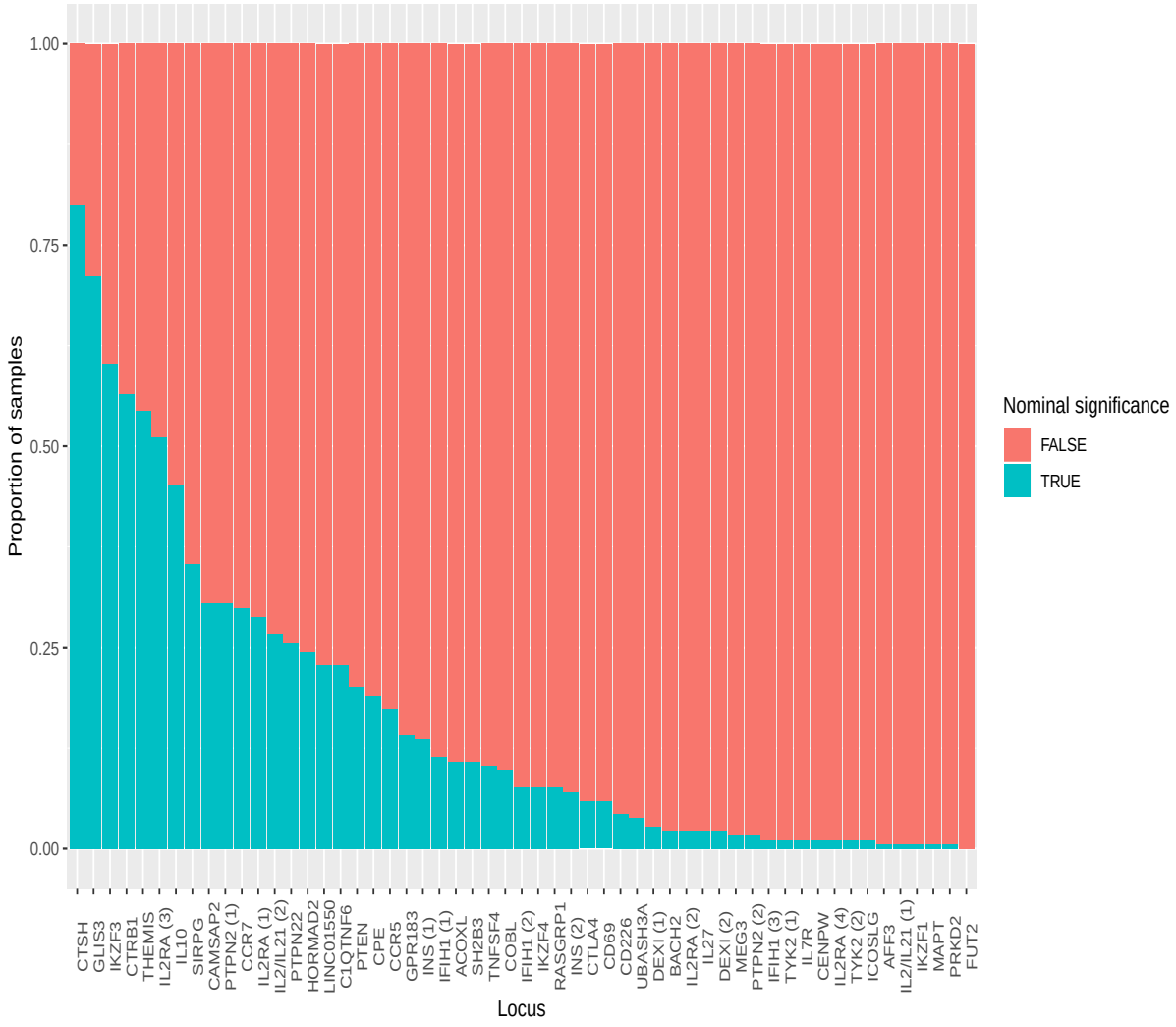
Figure 3.7: All T1D-associated variants examined for heterogeneity in effect size between the < 5 and ≥ 13 groups. Left panel: log-ORs for each AAD group, red dashed line indicating a log-OR of 0. Right panel: $-log_{10}$p value from heterogeneity test, red dotted line indicates a p value of 0.05, red solid line indicates the FDR significance of 0.1, red dashed line indicates Bonferroni significance at the 5% level.



Finally, when randomly splitting into sets of 50% cases and controls from each ancestry group, the nominal significance in the heterogeneity tests was observed $> 45\%$ of the time for each of the six loci with FDR significant heterogeneity in the primary analysis as well as evidence of heterogeneity in all subsequent sensitivity analyses (Figure 3.8). Although 45% would not be considered as particularly strong evidence by itself, there is less statistical power in each of the analyses here. The fact that nominal significance is observed over 50% of the time for five out of six differentially-associated loci provides reassurance that there were no loci with significant heterogeneity between groups due

solely to a chance distribution of alleles across the AAD groups in a subset of individuals driving the association. This array of sensitivity analyses provides confidence that the detected heterogeneity in the primary analysis are robust.

Figure 3.8: Proportion of nominally associated heterogeneity tests from sampling 50% of cases and controls from each ancestry group 100 times.



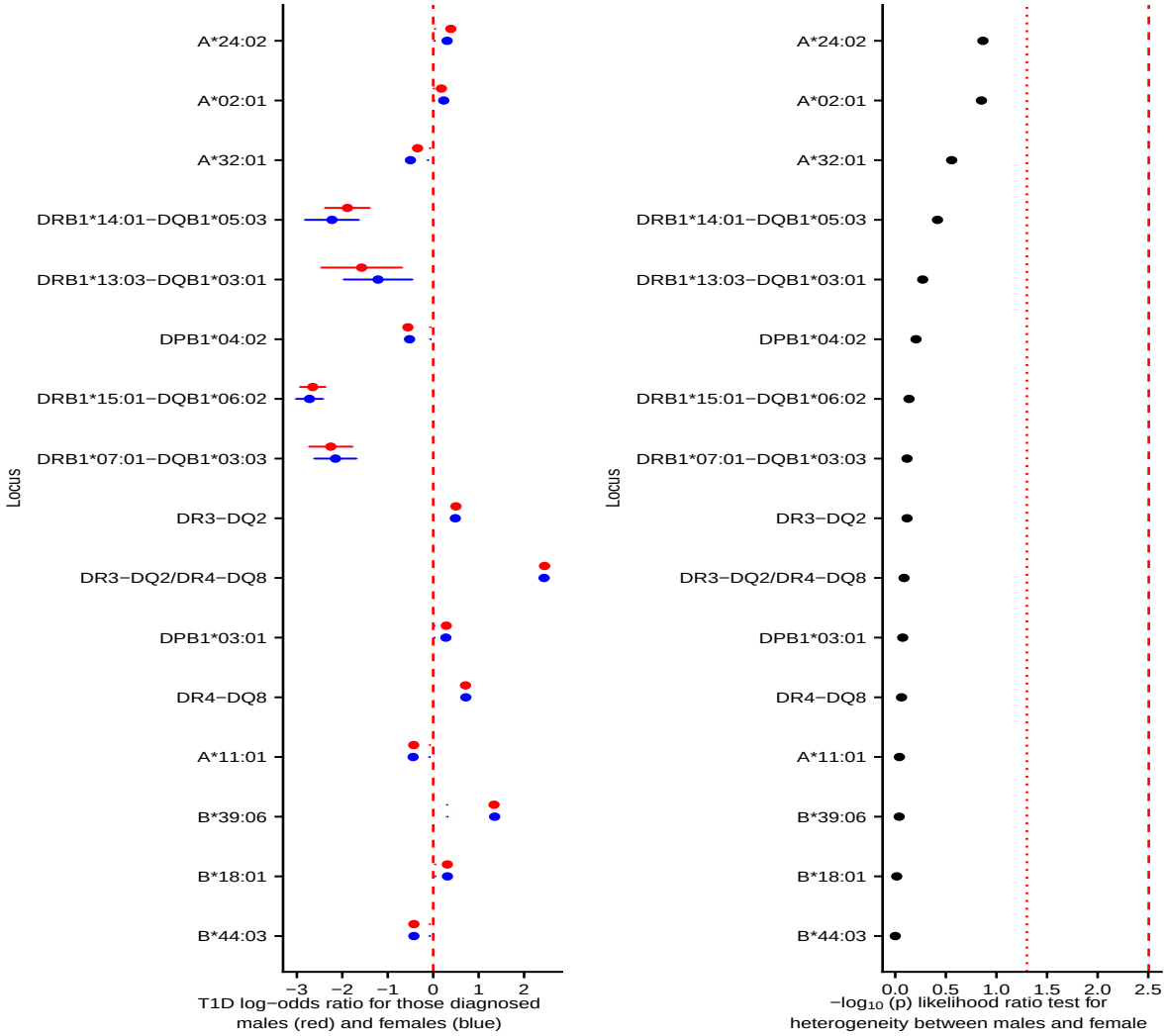
3.3.4 Differential-association between males and females

In order to examine whether or not any loci (HLA and non-HLA) were differentially-associated between male and female T1D individuals, I repeated the primary analysis above but with the outcome groups as controls, male T1D cases and female T1D cases, ignoring AAD in this analysis. This was primarily to examine this separate question, but also to gauge how often the heterogeneity test reaches significance using a phenotype

that intuitively I expected to display similar genetic effect sizes. Minimal evidence of heterogeneity for this phenotype would further increase confidence that the differential-associations observed in the AAD-group analysis were genuine.

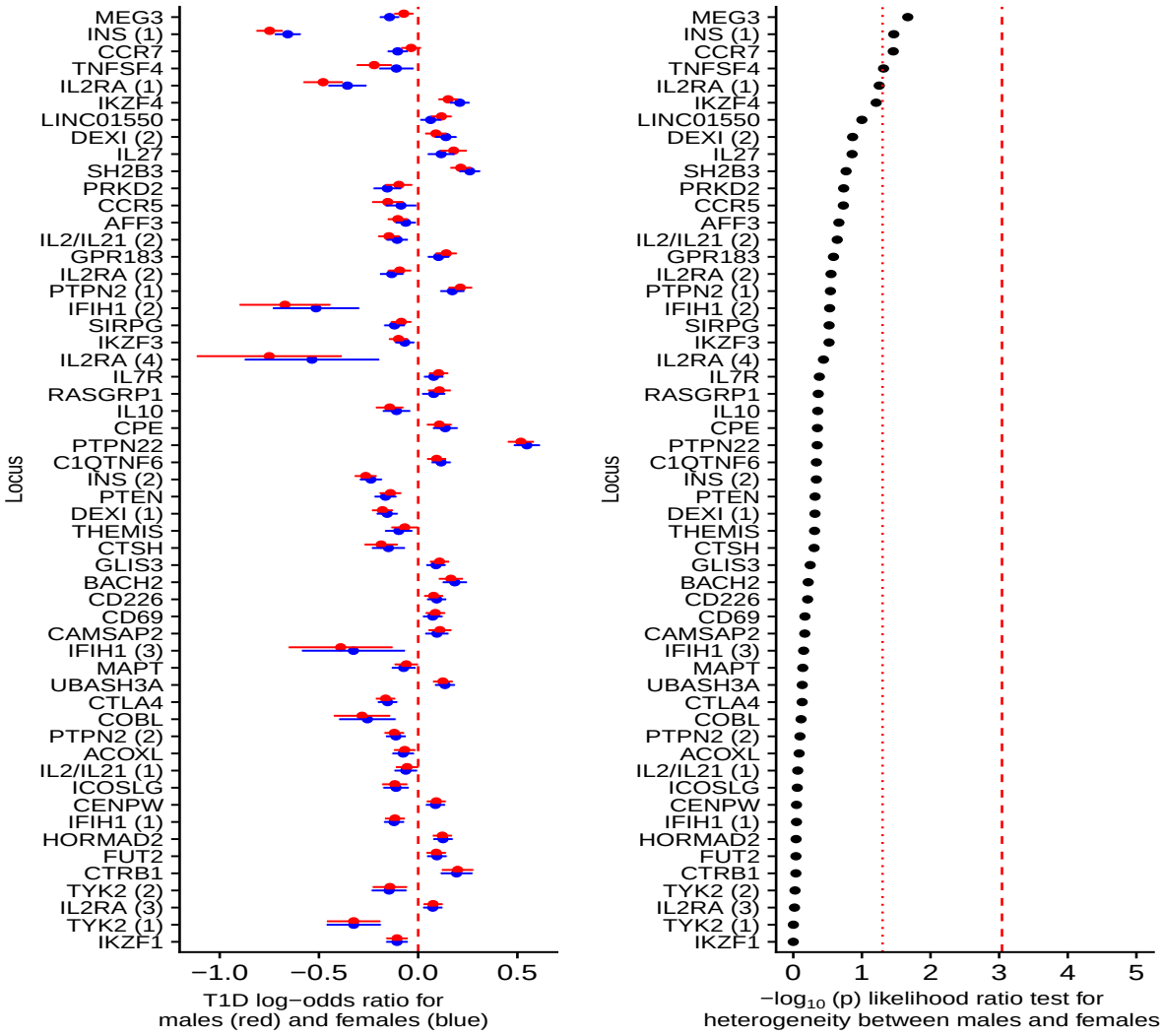
When examining the HLA haplotypes/classical alleles, there were no loci that reached nominal significance, with a strikingly similar effect size between males and females (Figure 3.9). The similarity in log-odds ratios between males and females, as well as the small standard errors around these estimates, which can be visualised by examining the size of the 95% confidence intervals, provides high confidence that the HLA associations are consistent between males and females.

Figure 3.9: All T1D-associated classical HLA haplotypes/alleles examined for heterogeneity in effect size between males and females. Left panel: log-ORs for each sex, red dashed line indicating a log-OR of 0. Right panel: $-log_{10}$p value from heterogeneity test, red dotted line indicates a p value of 0.05, red dashed line indicates Bonferroni significance at the 5% level.



Similarly, when examining the non-HLA alleles, no locus was associated to FDR< 0.1. The variants near Maternally expressed 3 (*MEG3*), the first index variant near Insulin (*INS*) and Chemokine CC Motif Receptor 7 (*CCR7*) showed nominally significant heterogeneity between males and females ($p < 0.05$), but these are not robust findings since they do not survive multiple comparison correction and thus the conclusion is that there is no evidence for any T1D-associated loci having a stronger effect in one sex compared to the other (Figure 3.10). Overall these results provide more certainty that the AAD results are genuine, since the same tests were performed here without any trend towards either sex in the same way as the trend towards larger effect sizes in the < 7 group.

Figure 3.10: All T1D-associated variants examined for heterogeneity in effect size between males and females. Left panel: log-ORs for each sex, red dashed line indicating a log-OR of 0. Right panel: $-log_{10}$p value from heterogeneity test, red dotted line indicates a p value of 0.05, red dashed line indicates Bonferroni significance at the 5% level.

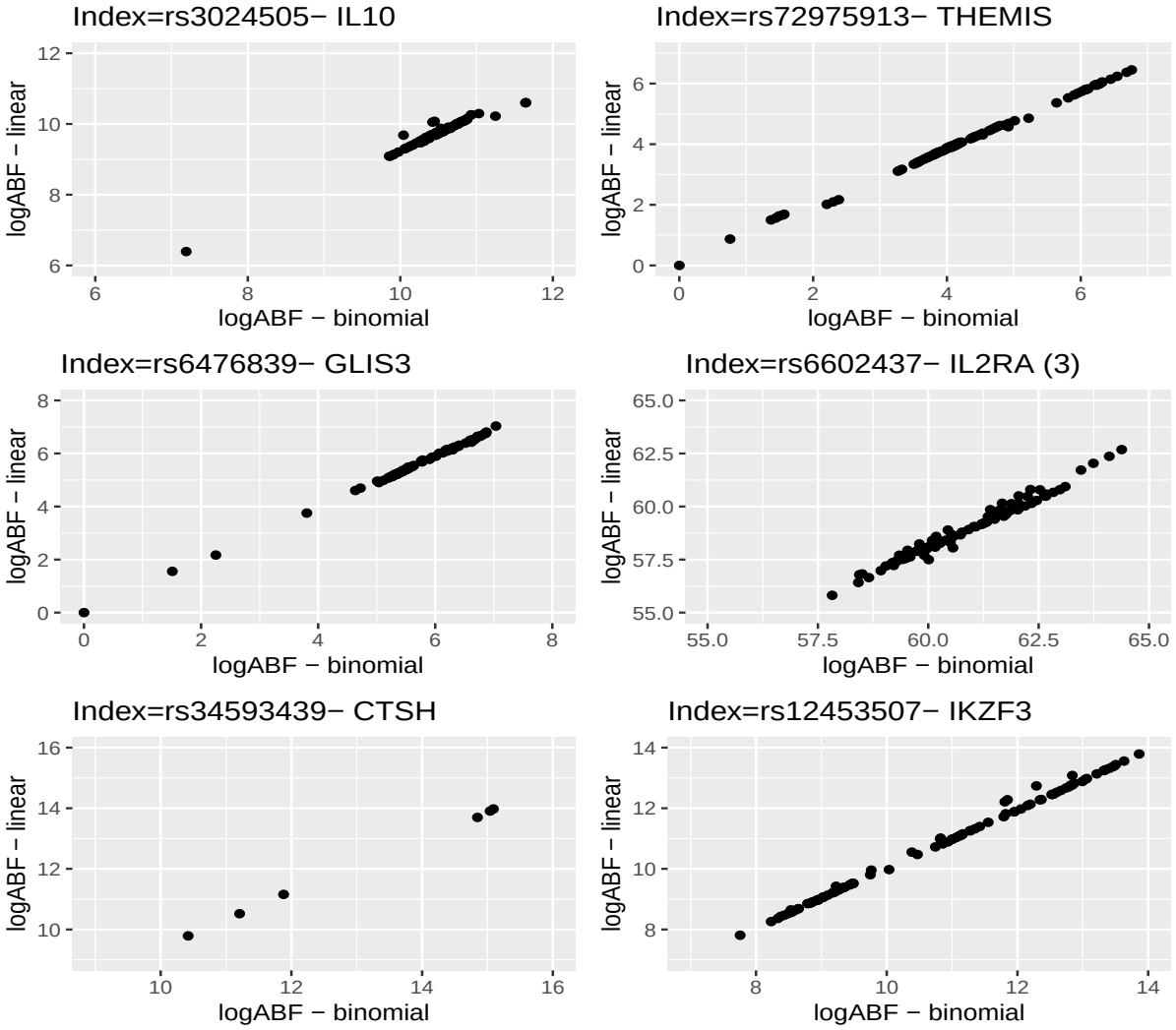


3.3.5 Fine mapping and colocalisation

I fine mapped the 0.5 Mb region around each of the index variants found to be differentially-associated between AAD groups. After filtering for imputation QC, the number of variants under consideration increased by a median (IQR) of 188 (151, 356) as a result of imputation.

Since the outcome in the analysis was binary but the models fitted in the stochastic search assume a Gaussian outcome, I calculated the ABF from the 100 most visited models from the GUESS stochastic search fitting linear regression models and logistic regression models. The ABFs from the models where the outcome was assumed linear were similar to those modelling the outcome as binary Figure 3.11, suggesting the linear approximation in the outcome is reasonable.

Figure 3.11: Comparison of approximate Bayes factors from the top 100 most visited models from the GUESS stochastic search at each of the six loci with evidence of differential-association between the modelling of a linear outcome versus the modelling of a binary outcome.



Five out of the six regions had posterior probability > 0.8 of a single causal variant in the region. The *IL2RA* region had highest posterior probability of three causal variants (posterior probability of three causal variants=0.566, posterior probability of four causal variants=0.387), which is consistent with current literature [202], including variants from groups A, C and E from [202] and not F, likely due to the low MAF of this variant.

I focussed on the five regions with highest posterior probability of a single causal variant in the region, which have the following number of variants in the group with highest group posterior probability of association *CTSH* (6), *GLIS3* (24), *IKZF3* (103), *THEMIS* (26) and *IL10* (4). Though some variants were removed following imputation for QC reasons,

so this is not an exhaustive set of variants that could be causal, it instead represents the most likely set of causal variants of those that passed QC.

Three of the five regions with high posterior probability of a single casual variant in the region had evidence of colocalisation with whole-blood eQTLs. The minor T allele at one if the most likely causal variants in the *CTSH* region, rs2289702 (C>T) is associated with decreased risk of early-diagnosed T1D and decreased *CTSH* expression (colocalisation probability with *CTSH* mRNA expression= 0.998) (Figure 3.12). The *IKZF3* region contains a large haplotype with over 100 variants that affect transcription of at least three genes in whole blood, *IKZF3*, ORMDL Sphingolipid Biosynthesis Regulator 3 (*ORMDL3*) and Gasdermin B (*GSDMB*). The minor T allele at a variant in the set of most likely causal variants, rs921649 (C>T), is associated with decreased risk of early-diagnosed T1D, decreases *IKZF3* expression and increases expression of *ORMDL3* and *GSDMB* (colocalisation probability with *IKZF3* mRNA expression= 0.982, *ORMDL3* mRNA expression=0.981, *GSDMB* mRNA expression= 0.981) (Figure 3.13). Finally, the minor A allele at one of the most likely causal variants in the *THEMIS* region, rs9491893 (G>A), is associated with decreased risk of early-diagnosed T1D and decreased *THEMIS* expression (colocalisation probability with *THEMIS* mRNA expression= 0.953). This region was first associated with T1D in chapter 2 with two named candidate genes in the region, *PTPRK* and *THEMIS*, the posterior probability of colocalisation between disease signal and *PTPRK* expression is much lower (colocalisation probability with PTPRK mRNA expression= 0.023) providing evidence that the more likely causal variant in the region is *THEMIS* (Figure 3.14).

Figure 3.12: Results from colocalisation/fine mapping the *CTSH* region. 1: Absolute z-scores from whole-blood eQTL study examining variant effects on *CTSH* expression, coloured by LD r^2 to the index eQTL. 2: Absolute z-scores from logistic regression examining variant associations with T1D risk at < 7 years, coloured by LD r^2 to the index *CTSH* eQTL. 3: Gene positions; arrows indicating direction of transcription. 4: Early-diagnosed T1D log-ORs and 95% CIs for each GUESSFM prioritised variant. 5: $\log_e$(absolute eQTL z-score) if z-score> 0 and $-\log_e$(absolute eQTL z-score) if z-score< 0, for eQTLs with a p value of $< 5 \times 10^{-50}$, coloured red if contained in the set of most likely causal variants, as produced by GUESSFM; shape corresponds to the gene that the variant is affecting transcription of.

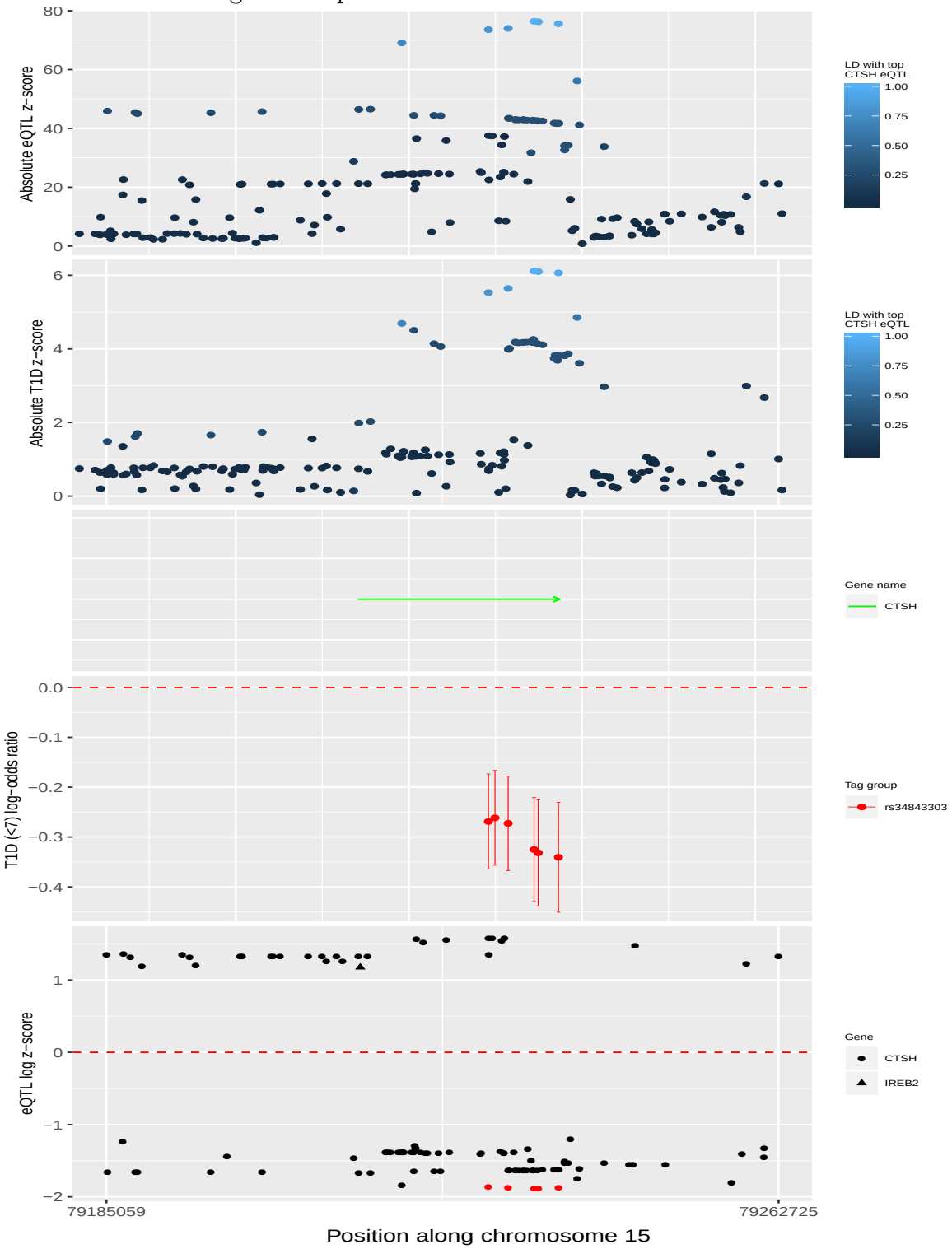


Figure 3.13: Results from colocalisation/fine mapping the *IKZF3* region. 1: Absolute z-scores from whole-blood eQTL study examining variant effects on *IKZF3* expression, coloured by LD r^2 to the index eQTL. 2: Absolute z-scores from logistic regression examining variant associations with T1D risk at < 7 years, coloured by LD r^2 to the index *IKZF3* eQTL. 3: Gene positions; arrows indicating direction of transcription. 4: Early-diagnosed T1D log-ORs and 95% CIs for each GUESSFM prioritised variant. 5: $\log_e$(absolute eQTL z-score) if z-score> 0 and $-\log_e$(absolute eQTL z-score) if z-score< 0, for eQTLs with a p value of $< 5 \times 10^{-150}$, coloured red if contained in the set of most likely causal variants, as produced by GUESSFM; shape corresponds to the gene that the variant is affecting transcription of.

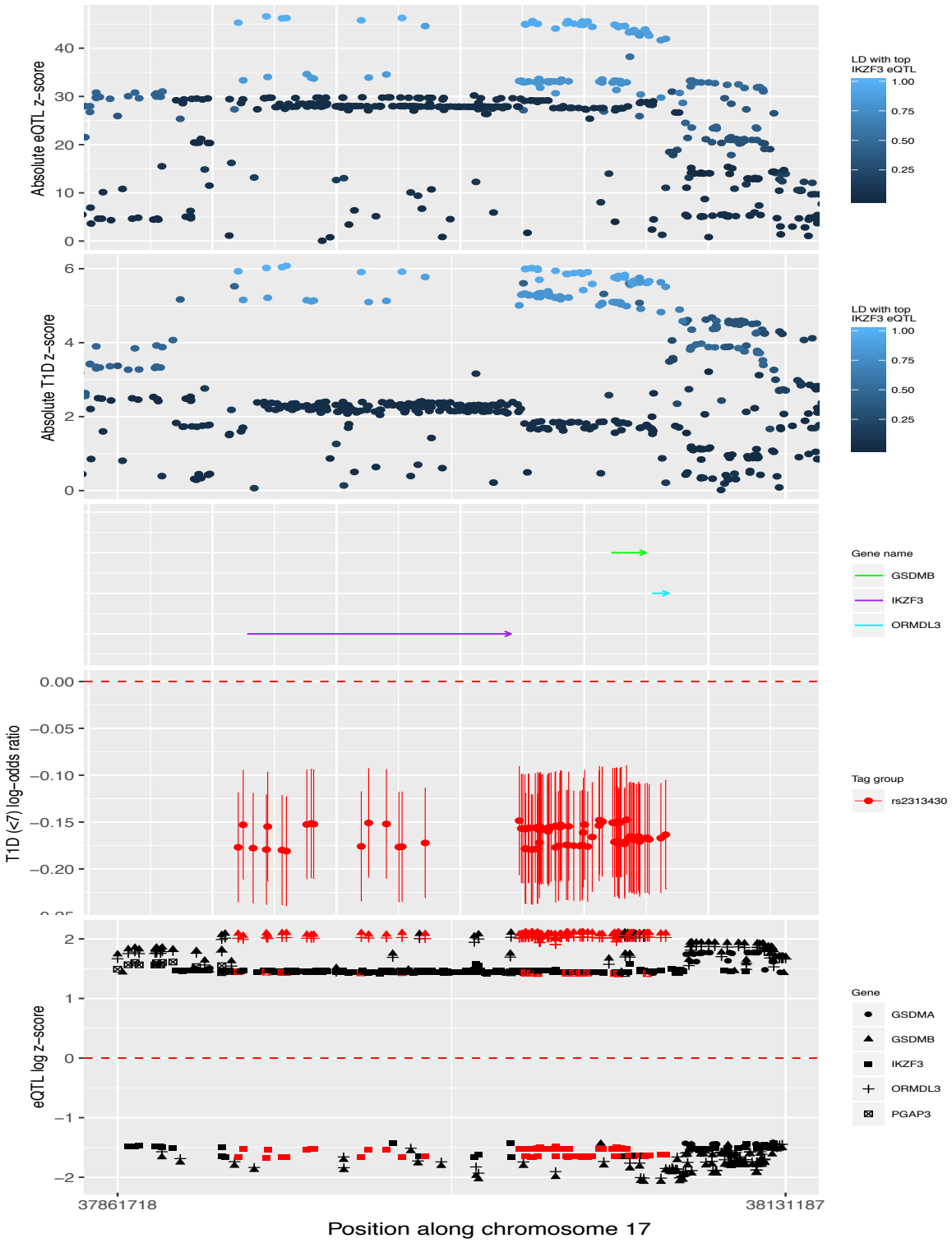
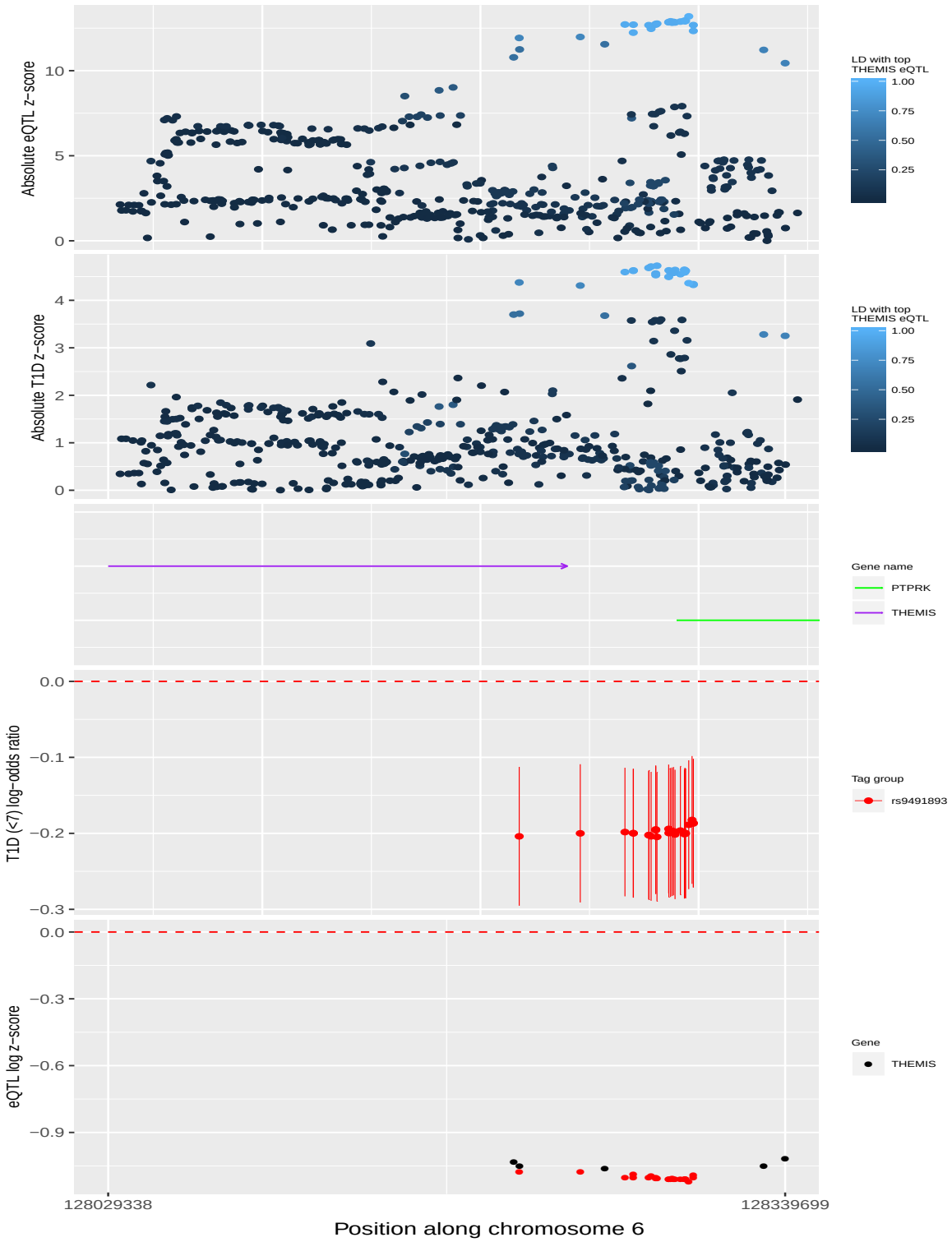


Figure 3.14: Results from colocalisation/fine mapping the *THEMIS* region. 1: Absolute z-scores from whole-blood eQTL study examining variant effects on *THEMIS* expression, coloured by LD r^2 to the index eQTL. 2: Absolute z-scores from logistic regression examining variant associations with T1D risk at < 7 years, coloured by LD r^2 to the index *THEMIS* eQTL. 3: Gene positions; arrows indicating direction of transcription. 4: Early-diagnosed T1D log-ORs and 95% CIs for each GUESSFM prioritised variant. 5: $\log_e$(absolute eQTL z-score) if z-score> 0 and $-\log_e$(absolute eQTL z-score) if z-score< 0, for eQTLs with a p value of $< 5 \times 10^{-25}$, coloured red if contained in the set of most likely causal variants, as produced by GUESSFM; shape corresponds to the gene that the variant is affecting transcription of.



Finally, neither of the *GLIS3* or *IL10* loci showed evidence of colocalisation with any whole-blood *cis* eQTLs (posterior probability of colocalisation < 0.02), three proximal genes had whole-blood eQTLs in the *IL10* locus, *IL10*, Interleukin-24 (*IL24*) and Fas Apoptotic Inhibitory Molecule (*FAIM*), all of which showed no evidence of colocalisation. The manhattan plots of the *GLIS3* and *IL10* whole-blood eQTLs and T1D risk are shown in Figures 3.15 and 3.16.

Figure 3.15: Absolute z-scores from *GLIS3* whole-blood eQTL (top panel) and T1D risk diagnosed at < 7 years, coloured by LD (r^2) to the index eQTL (bottom panel).

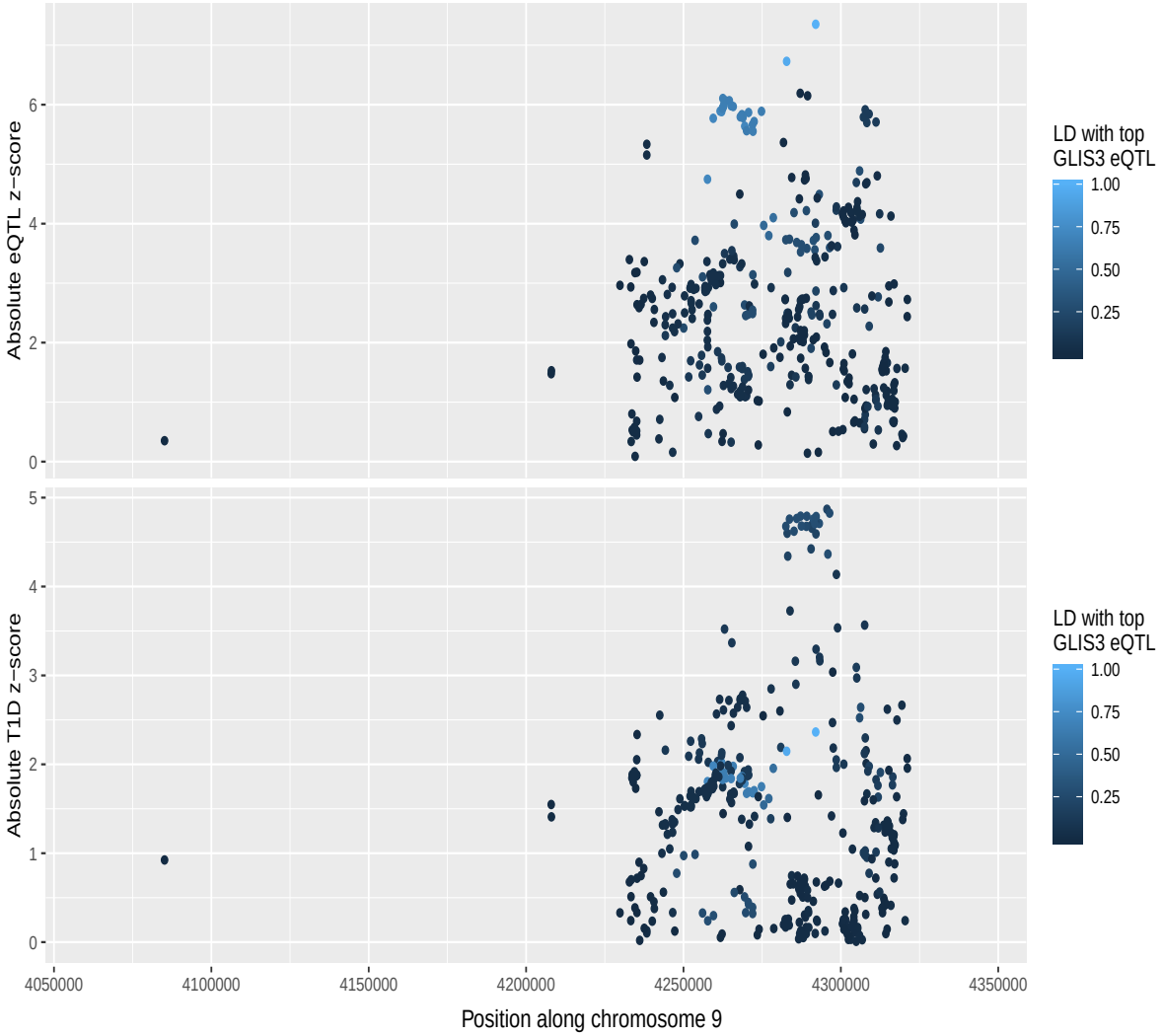
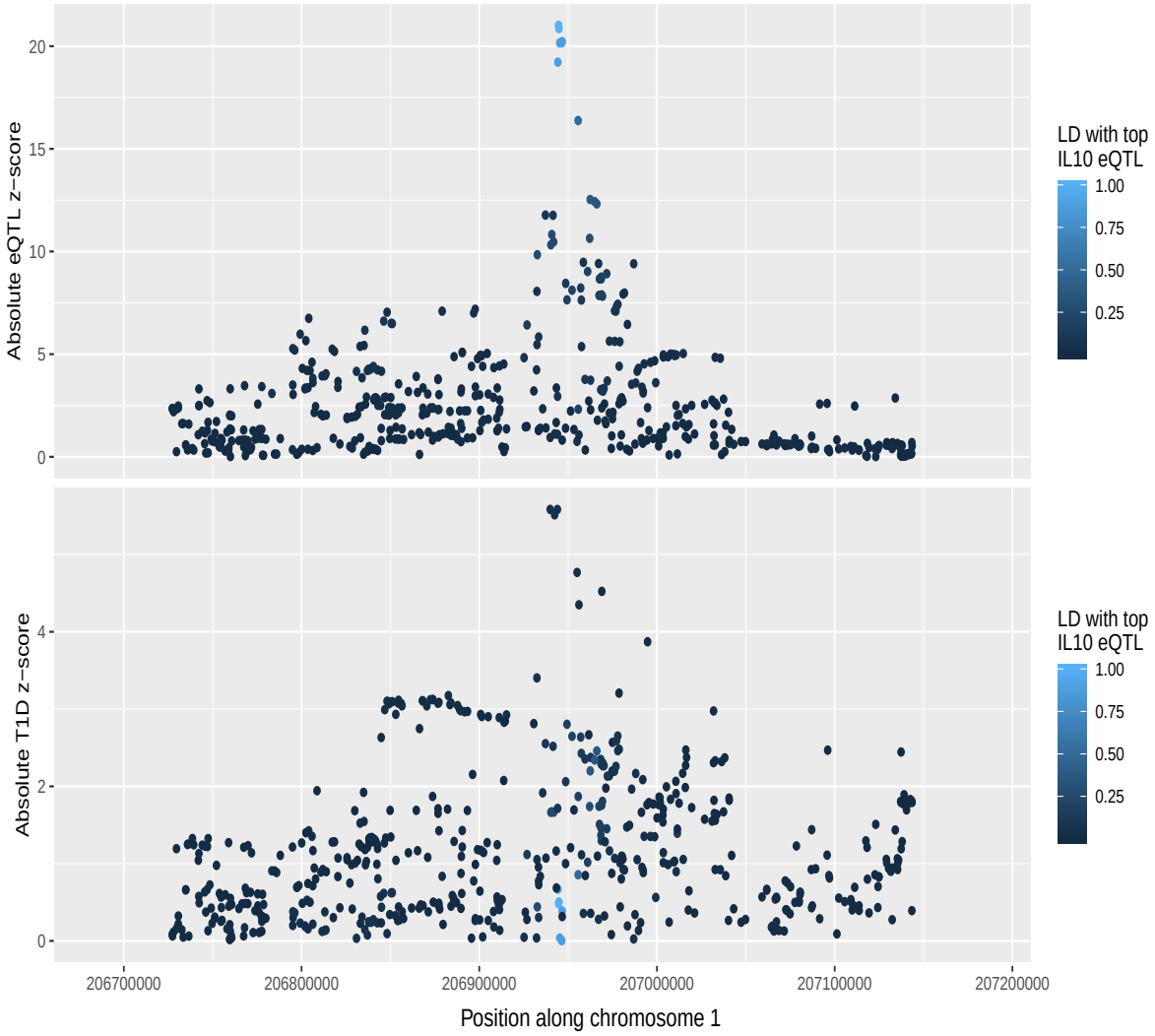


Figure 3.16: Absolute z-scores from *IL10* whole-blood eQTL (top panel) and T1D risk diagnosed at < 7 years, coloured by LD (r^2) to the index eQTL (bottom panel).



3.3.6 Heritability analyses

The heritability estimates decreased with increasing AAD group, whether including the HLA region in the estimation (Table 3.4) or not (Table 3.5). As expected, the heritability estimates were lower when the HLA was not included in the calculation. The < 7 group had an estimated heritability approximately 10% higher than the ≥ 13 group. This trend towards higher heritability in younger AAD groups remained consistent regardless of the assumed prevalences of disease in these groups.

Table 3.4: Heritability estimates from including all variants on the ImmunoChip from mixed model including as fixed effects sex and the ten largest genetic PCs. Rows are different assumed prevalences of T1D to give an indication of the likely true range of heritabilities.

Disease prevalence (%) (< 7,$7-13$,≥ 13)	< 7	$7-13$	≥ 13
0.4,0.4,0.4	0.366 (0.346, 0.386)	0.301 (0.284, 0.319)	0.233 (0.206, 0.26)
0.5,0.3,0.3	0.383 (0.362, 0.404)	0.285 (0.268, 0.301)	0.22 (0.195, 0.245)
0.5,0.2,0.2	0.383 (0.362, 0.404)	0.264 (0.249, 0.28)	0.204 (0.18, 0.228)

Table 3.5: Heritability estimates from including all variants on the ImmunoChip outside the HLA region from mixed model including as fixed effects sex and the ten largest genetic PCs. Rows are different assumed prevalences of T1D to give an indication of the likely true range of heritabilities.

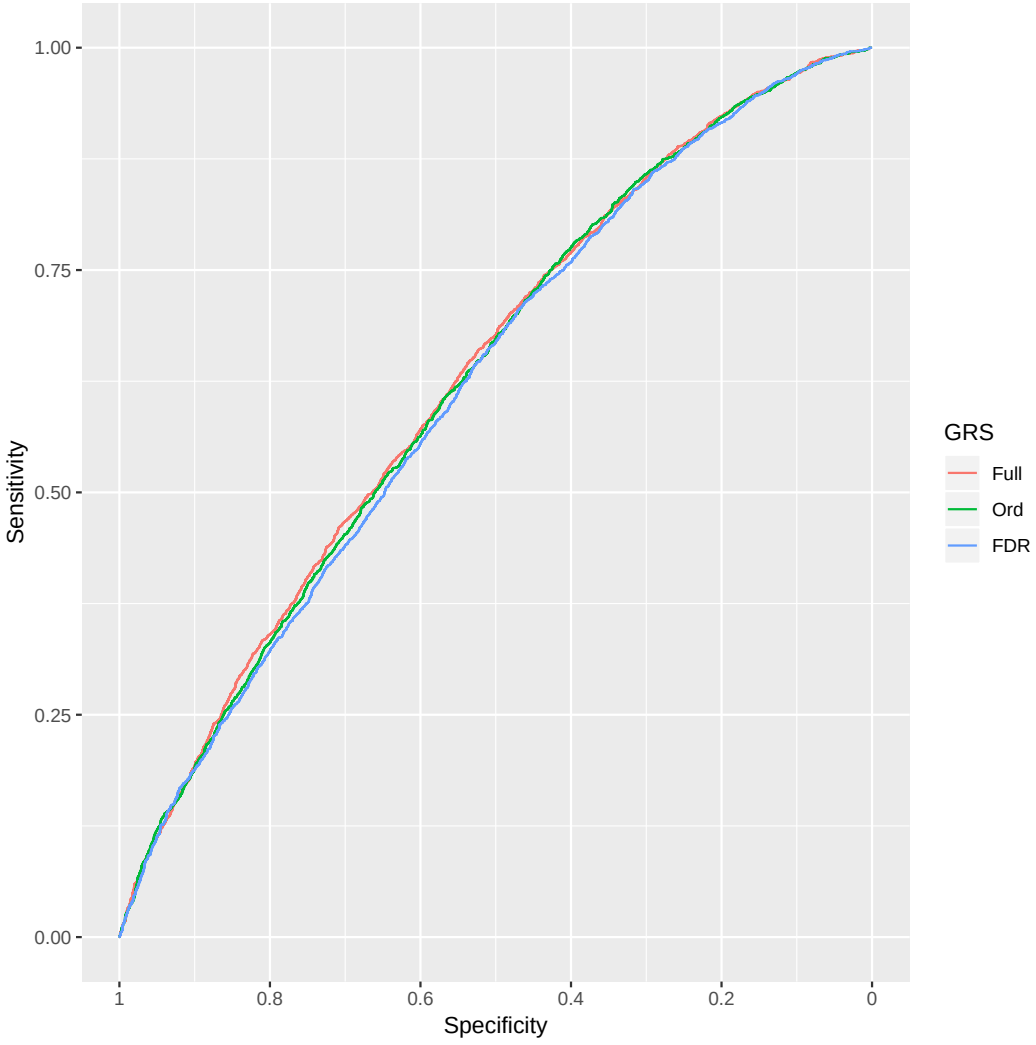
Disease prevalence (%) (< 7,$7-13$,≥ 13)	< 7	$7-13$	≥ 13
0.4,0.4,0.4	0.209 (0.188, 0.229)	0.178 (0.16, 0.195)	0.09 (0.069, 0.11)
0.5,0.3,0.3	0.218 (0.197, 0.239)	0.168 (0.152, 0.184)	0.085 (0.065, 0.104)
0.5,0.2,0.2	0.218 (0.197, 0.239)	0.156 (0.141, 0.171)	0.079 (0.061, 0.097)

3.3.7 Prediction

Including either GRS_{full}, GRS_{ord} or GRS_{fdr} as a covariates in logistic regressions predicting T1D diagnosed at < 7 compared to ≥ 7 had similar predictive ability, though the more variants included in the GRS, the marginally better the prediction ($\text{AUC}_{full} = 0.6295$, $\text{AUC}_{ord} = 0.6261$, $\text{AUC}_{fdr} = 0.6199$). The similarity in ROC curves can be observed in

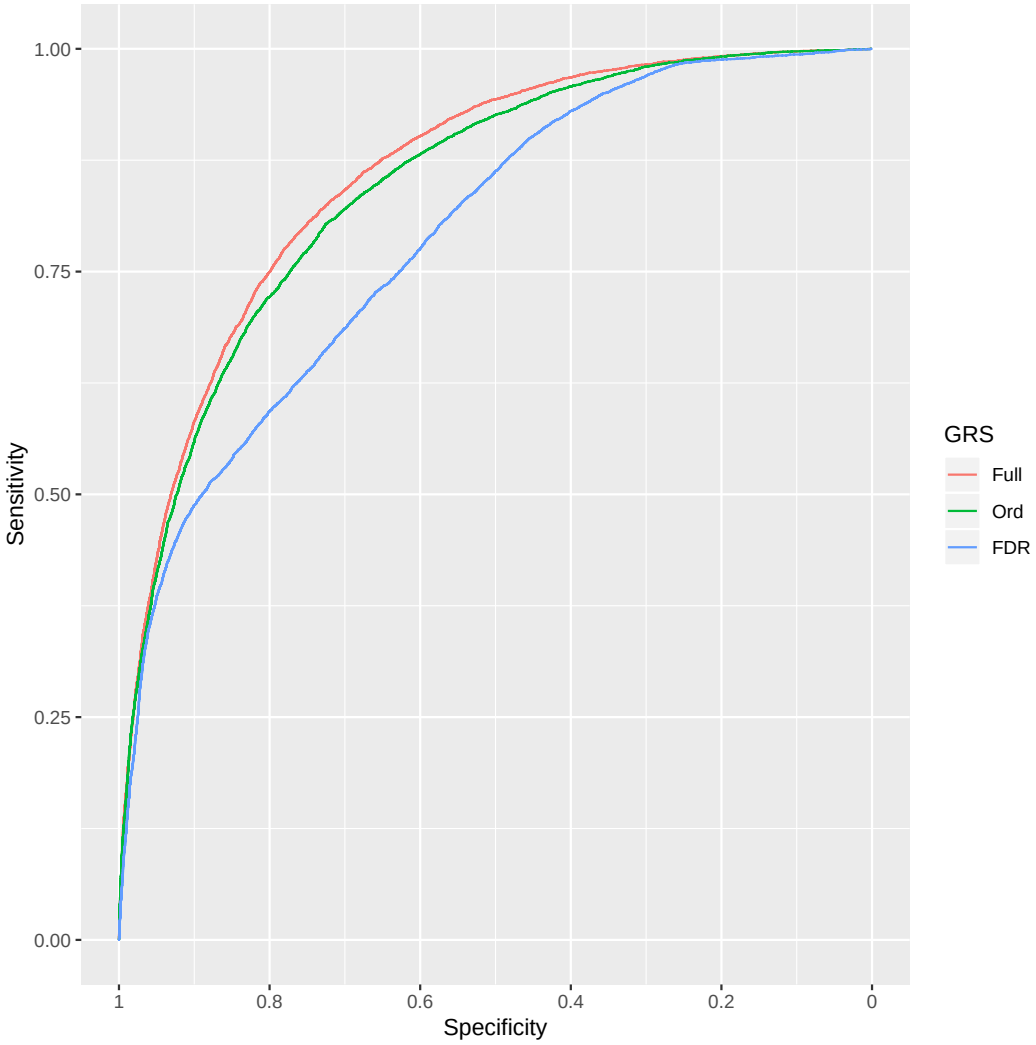
Figure 3.17.

Figure 3.17: Receiver operating characteristic (ROC) curve from logistic regression model including as covariates sex, the ten largest genetic PCs and a GRS including all T1D-associated variants (red), variants with largest effect size in the < 7 group, intermediate in the $7 - 13$ group and smallest in the ≥ 13 group (green) and variants with FDR < 0.1 differential-association between the < 7 and ≥ 13 group (blue). Including T1D cases only and predicting which individuals are diagnosed at < 7 years compared to ≥ 7.



A similar but clearer trend can be observed when examining logistic regression which include cases and controls and include the same three GRS's as covariates in the models to predict T1D status. The GRS_{full} covariate has a much higher AUC than using only the FDR < 0.1 differentially-associated variants in the GRS ($\text{AUC}_{full} = 0.8581$, $\text{AUC}_{fdr} = 0.7892$), suggesting that those loci that are not differentially-associated between AAD groups still improve prediction of T1D status (Figure 3.18).

Figure 3.18: Receiver operating characteristic (ROC) curve from logistic regression model including as covariates sex, the ten largest genetic PCs and a GRS including all T1D-associated variants (red), variants with largest effect size in the < 7 group, intermediate in the $7 - 13$ group and smallest in the ≥ 13 group (green) and variants with FDR < 0.1 differential-association between the < 7 and ≥ 13 group (blue). Including T1D cases and controls and predicting which individuals are cases compared to controls.



3.4 Discussion

In chapter 2, I assessed each variant on the ImmunoChip for its association with T1D AAD, examining cases only. This approach provided evidence that there are AAD associated variants, but had the potential limitation of having to overcome multiple testing burdens and also examined AAD across a broad range of ages in an unspecific way. In this Chapter, I explored a separate question, focussing instead on whether known T1D association regions have a particularly strong effect (or a particularly high prevalence) in T1D diagnosed at a young age (< 7 years) compared to T1D diagnosed at ≥ 13 years. This alternative approach has yielded a number of associations that may have previously lacked power to detect owing to the multiple testing burden encountered from analysing ImmunoChip-wide. The differential-association between the < 7 and ≥ 13 groups was expected for the DR3-DQ2/DR4-DQ8 diplotype as well as the class I A*24:02 and B*39:06, since these effects have been observed previously [84, 195, 83, 135], however, the decreased prevalence of the protective DRB1*15:01-DQB1*06:02 and DRB1*07:01-DQB1*03:03 amongst T1D individuals diagnosed at < 7 compared to ≥ 13 years is a novel finding. This work has been published in a peer-reviewed journal [88].

Outside the HLA region, six T1D-associated regions showed robust evidence of having stronger effect sizes in those diagnosed at < 7 years compared to ≥ 13 years, including the region containing *THEMIS*, which was first identified in chapter 2. Fine mapping and colocalisation analyses of these regions using only controls and cases diagnosed at < 7 years showed that three of the loci, *CTSH*, *IKZF3* and *THEMIS* colocalise with whole-blood eQTLs. The *GLIS3* and *IL10* regions did not show evidence of colocalisation with any whole-blood eQTLs, whilst the *IL2RA* region had evidence of more than one disease causal variant in the region, violating the assumptions of the colocalisation analysis method. Whilst methods do exist to examine colocalisation when multiple signals are present in the region [79], the *IL2RA* region is extremely complicated and would take a long time to evaluate colocalisation with so many disease signals without guaranteeing a satisfactory answer, especially given the mixture of cell types in whole blood. Therefore colocalisation analyses in the *IL2RA* region are beyond the scope of this analysis and to

be considered in a cell-type specific manor in future work.

The *GLIS3* region is likely acting in pancreatic beta cells rather than immune cells, given that *GLIS3* is highly expressed in the pancreas (`https://gtexportal.org/home/`) [47], that there is colocalisation between the T1D and T2D signals in the region [11] and there is no association in the region with any other autoimmune disease (`https://genetics.opentargets.org`). It has been shown previously that the T1D and T2D risk allele in the *GLIS3* region increases beta-cell susceptibility to apoptosis, implying that beta-cell fragility could be an underlying risk factor for T1D and T2D [55, 142].

The remaining five AAD differentially-associated loci (*CTSH*, *IKZF3*, *IL2RA*, *THEMIS* and *IL10*) have roles in T- and/or B- cell biology. It has been suggested that the T1D risk allele near *CTSH* alters disease risk by altering beta-cell sensitivity to apoptosis. At the time this study was carried out, rs3825932 (C>T) was the T1D index variant, likely due to lower statistical power in earlier GWAS analyses for T1D. This variant is in LD ($r^2 = 0.26$) with one of the credible SNPs in the present analysis, rs2289702 (C>T). However, the C allele at the variant studied in the previous analysis is counter-intuitively associated with protection from beta-cell apoptosis, but increased T1D risk. It therefore seems unlikely that sensitivity to beta-cell apopotosis is the primary reason for altered T1D risk in this region. The C allele in one of the credible variants in the present analysis, rs2289702, is associated with increased CTSH expression and increased T1D risk. One hypothesis to explain increased T1D risk in this region is that the increased *CTSH* expression with the T1D risk allele increases cleavage of the N-terminus of the Toll-like receptor 3 (*TLR3*) protein, increasing its functionality [151]. *TLR3* recognises double-stranded RNA, a marker of viral infection, and triggers an immune response against the infection [198]. By increasing the functionality of *TLR3*, the increased expression of *CTSH* could lead to heightened response to viral infections and increased upregulation of the type 1 interferon response. Type 1 interferon potentially upregulates HLA class I expression on beta cells, which would increase their sensitivity to killing by $CD8^+$ cytotoxic T cells. This in turn could lead to an increased risk of T1D, especially since *TLR3* is expressed in the pancreas [156]. Since viral infections are more frequent during childhood, this could help explain

the increased risk of childhood-diagnosed T1D.

The locus near *IKZF3* has been studied in depth due, in part, to the association in the region with asthma and paediatric asthma [167, 28]. The set of variants prioritised by GUESSFM is a large haplotype of 103 variants and it is this haplotype that is associated with asthma and also colocalised with whole-blood eQTLs for *IKZF3*, *ORMDL3* and *GSDMB*, suggesting one or more of these genes could be causal for T1D and/or asthma. However, the effect direction is opposite between asthma and autoimmue diseases, for example the C allele at rs921649 (C>T), a variant in the haplotype, increases susceptibility to autoimmune diseases rheumatoid arthritis [143], irritable bowel disease [117], primary biliary sclerosis [118] and T1D [144], but is protective for asthma. The three candidate genes are all expressed in lymphocytes and *IKZF3* is upregulated following activation with anti-CD3/CD28 beads, whilst *GSDMB* and *ORMDL3* are downregulated following activation (`https://dice-database.org/`). I can therefore hypothesise how these genes might be causally associated with T1D. For example, *IKZF3*, also known as Aiolos, is important for B-cell activation and both B- and T- cell differentiation [204, 153]. Increased *IKZF3* expression in whole blood due to carrying the T1D risk allele could lead to increased baseline levels of B-cell activation. On the other hand, *ORMDL3* has been shown to negatively regulate lymphocyte activation and cytokine production [167, 33]. The decreased expression of this gene in carriers of the T1D risk allele could lead to less negative regulation of cytokine production and thus a more active immune system under resting conditions and therefore an increased T1D risk. Finally, *GSDMB* can act as a pryoptotic protein [145], so the direction of T1D risk would not be so obvious to link to this gene but it is possible that reduced pryoptosis could lead to increased T1D risk by autoreactive cells surviving.

The locus near *PTPRK* and *THEMIS* was first identified as T1D-associated in chapter 2. The whole-blood eQTL colocalisation between T1D risk at < 7 years and *THEMIS* expression, and lack of colocalisation with *PTPRK* expression implies that *THEMIS* is the most likely causal gene in the region. *THEMIS* is expressed in the thymus and alters the threshold for selection of single positive (either $CD4^+$ or $CD8^+$) T cells from

double positive ($CD4^+8^+$) cells. Deleting *THEMIS* in mice reduces the selection of double positive thymocytes [93, 68]. The T1D risk allele, for example the G allele at rs13204742 (G>T) is associated with increased *THEMIS* expression. It is possible therefore that this increased *THEMIS* expression leads to increased transition of double positive to single positive thymocytes and this may lead to an increased probability of an autoreactive T cell entering circulation.

In chapter 2, when fine mapping AAD as a continuous variable, two signals in the *THEMIS* region were identified. However, only one is identified here. This is likely due to an increase in statistical power in the analysis in chapter 2, where a mixed model was used and all cases were included in the analysis including T1D affected sibling pairs and no controls, as opposed to the analysis here where independent cases and controls only were used. The lead signal, however, is consistent between this analysis and the analysis in chapter 2, which is encouraging since the analyses were approached in different ways and the same conclusion was reached.

The six FDR differentially-associated regions are the primary focus of this chapter. However, the ≥ 13 group contained just 1,708 individuals and therefore the effect size estimate for this group would be more precise if the sample size increased and the standard error around the estimate would decrease. This would likely lead to a number of the regions where the ordering is such that the < 7 group has the strongest effect, the $7 - 13$ group has the intermediate effect and the ≥ 13 group has the weakest effect reaching an association of FDR < 0.1. Given the large number of loci with log-odds ratio in this order (21) compared to the number expected by chance (8), it suggests many T1D-associated loci are more prevalent in early-diagnosed individuals than later-diagnosed individuals, larger sample sizes would be required to determine how many of these 21 loci are appearing in this order by chance and how many are truly AAD-associated. A recent analysis of AAD of T1D using an older cohort has been performed and would be a useful cohort to examine as a follow-up analysis to determine consistency of the effect estimates in the ≥ 13 group between this analysis and theirs [130].

The heritability estimates in this chapter show an increased heritability in the < 7 group

compared to the $7-13$ and ≥ 13 groups, regardless of inclusion or exclusion of the HLA region and also under various disease prevalence assumptions. This result mirrors the results in [159, 158], where the probability of developing T1D if a sibling is diagnosed with T1D at a young age is higher than the probability of being diagnosed with T1D if a sibling is diagnosed with T1D at an older age, as well as the increased heritability estimates when examining childhood-onset T1D [112] compared to the estimates examining T1D diagnosed at all ages [108, 180]. However, the estimates obtained here must be interpreted with caution since the genotype platform is the ImmunoChip. This platform was designed to genotype regions previous associated with autoimmune diseases based on underpowered studies and also based on T1D individuals diagnosed at < 16 years. It is therefore possible that these heritability estimates simply reflect the ImmunoChip design, since the ImmunoChip is likely enriched for early-diagnosed T1D regions. However, there is still a difference in heritability estimates between AAD groups that are bounded within the < 16 years cutoff and thus there is likely genuinely higher heritability with earlier diagnosis, consistent with the trend in the family heritability estimates. The lower heritability estimates in the ≥ 13 group is consistent with increasing age accompanying greater risk and length of exposure to environmental factors associated with T1D. This finding is the opposite to what is seen for IQ [27], where the heritability increases in later life since the effect of parenting diminishes.

In conclusion, early-diagnosed T1D is driven primarily by particular HLA alleles, the most important factor being carriage of the DR3-DQ2/DR4-DQ8 diplotype, but also class I classical alleles A*24:02 and B*39:06. In addition, many non-HLA regions appear to have larger effects in the early-diagnosed T1D compared to later-diagnosed, in particular the loci near *CTSH*, *IKZF3*, *GLIS3*, *IL2RA*, *THEMIS* and *IL10*, suggesting malfunctions in the infiltrating immune cells and pancreatic beta cells themselves could all combine with HLA susceptibility to drive down the age at which T1D is diagnosed.

Chapter 4

ImmunoChip analysis of type 1 diabetes with over 60,000 individuals

4.1 Introduction

In the previous two chapters, I used existing data and attempted to identify variants associated with younger T1D AAD, a potentially important question, which pointed towards particular chromosome regions most associated with early-diagnosed T1D. However, since the last T1D association study in 2015 [144], which contained approximately $30,000$ individuals, genetic association studies examining other autoimmune diseases have expanded in sample size in a way that T1D genetic analyses have not. In rheumatoid arthritis (RA), for example, a genetic analysis of over $100,000$ individuals was performed [143] and a study of approximately $95,000$ individuals was carried out examining inflammatory bowel disease (IBD) [117]. These studies identified novel associations with respective diseases due to an increase in statistical power to detect either variants with lower MAF or variants with smaller effect on disease risk. The importance of increasing sample sizes to uncover novel disease insights is perhaps best illustrated by type 2 diabetes (T2D), in which the first genetic analysis contained $1,924$ cases and identified three disease-associated regions [48], to the largest GWAS which contained $74,124$ cases and identified 243 conditionally independent associations with disease [123].

It is likely that with increased sample size, novel insights into T1D aetiology could be gained in a similar way to other diseases. With that in mind, colleagues at the University of Virginia (UVA) collected DNA samples from $61,427$ individuals and genotyped the samples using the ImmunoChip. The association of each variant with T1D was tested with the aim of increasing our understanding of the genetics underlying T1D risk.

4.2 Methods

4.2.1 Genotyping

DNA samples from 61,427 individuals were genotyped on the ImmunoChip at the University of Virginia Genome Sciences Laboratory. I will refer to the entire dataset collectively as the MEGAT1D dataset.

4.2.2 Relatedness and ancestry calculation

In order to classify individuals into ancestry homogeneous populations, the ImmunoChip data including all individuals were pruned to an $r^2 < 0.2$, variants with a low MAF (< 0.005) were removed, and three regions with particularly large LD blocks were removed, as defined in [149]. Genotype data from the 1000 Genomes Project individuals were then downloaded [176], and the intersect of variants used on this platform and those in the pruned ImmunoChip dataset was taken, leaving a total of 8297 variants in both the MEGAT1D dataset and the 1000 Genomes Project dataset. Principal components were calculated for the 1000 Genomes Project individuals as in chapter 2, since this dataset represents a population of individuals without T1D from multiple ancestries.

The loadings from these PCs, $\mathbf{W_{1000G}} = \mathbf{Z^T_{1000G}Q_{1000G}}$, where $\mathbf{Z_{1000G}}$ is the mean-centred genotype matrix and $\mathbf{Q_{1000G}}$ is the eigenvectors from the 1000 Genomes Project dataset, were used to project PCs onto the MEGAT1D individuals using the same formula as in chapter 2, to obtain $\mathbf{Q_{MEGAT1D}}$, the projected PCs for the MEGAT1D dataset:

$$\mathbf{Q_{MEGAT1D}} = \frac{1}{M}\mathbf{Z_{MEGAT1D}W_{1000G}}\mathbf{\Lambda}^{-1}_{1000G}. \tag{4.1}$$

The ten largest PCs were then used in k-means clustering to define five clusters of ancestrally-similar individuals, namely European (EUR), African-American (AFR), Eastern (EAS), Finnish (FIN) and Admixed (AMR), labeled according to their closest 1000 Genomes Project ancestry label. The k-means clustering was carried out by Catherine Robertson at UVA.

Case-control ImmunoChip-wide association analyses were performed separately for each

ancestry cluster. Within each cluster, affected trios were excluded and analysed separately, and a set of unrelated individuals (more distant than 2nd degree) were selected from the remaining subjects using the KING software [125], as in chapter 2 and removing (randomly) one individual in a pair with a kinship coefficient estimate of greater than 0.0884.

For independent cases and controls within each ancestry group, PCs were calculated by performing PCA on the pruned genetic matrix of unrelated controls only and using these loadings to project them onto the cases, as in equation 4.1. This was done in an attempt to not regress out any true T1D association signal that might be detected in the PCs if including them in the PCA.

4.2.3 Defining ImmunoChip regions

There are 184 previously defined ImmunoChip regions which were densely genotyped due to their association with immune-related disorders. Using the 'humarray' R package, I mapped these regions from genome build GRCh36 to GRCh38. For each region, the variants were mapped to the latest genome build and then the region was defined as any variants lying between the lowest and highest positions of the variants as they mapped on the GRCh38 coordinates.

4.2.4 Imputation and post-imputation quality control

Genotypes were imputed using the Trans-Omics for Precision Medicine (TOPMed) [186] reference panel (freeze 5), which is a panel of over $50,000$ whole genome sequences with diverse ancestry groups included. The Michigan imputation server was used for imputation, which first phases the genotypes into probable haplotypes using the Eagle software (version 2.4) [119] and then uses the Minimac4 software to impute the variants not present on the genotyping platform [53]. Minimac4 is similar to IMPUTE2 in that it uses Hidden Markov Models (HMMs) to infer the probability the genotypes are part of a particular hidden state (reference haplotype) but differs in that the recombination rate is estimated from the data in Minimac4 rather than being pre-supplied, as in IMPUTE2.

Post-imputation quality control was performed in an ancestry-specific manor, excluding all variants with an imputation $r^2 < 0.8$ and a MAF< 0.005. In addition, I removed variants where the imputation information score differed between cases and controls by > 0.05, since this can artificially drive an association signal. Finally, I removed any variants imputed outside of the 184 densely genotyped regions on the ImmunoChip, since imputing variants without many genotyped variants in the vicinity will lead to convergence towards reference haplotype frequencies for the imputed variants, as there is not enough information to guide the imputation for the non-genotyped variants other than the reference haplotypes. Therefore in order to avoid either reporting false negative or false positive results, imputed variants outside of the ImmunoChip regions were removed, although those variants that were genotyped outside the ImmunoChip regions were included in the analysis.

For family trio analyses, the Mendelian error rates were calculated for each variant, which is the rate at which the offspring from two parents inherit a genotype not consistent with the genotypes of the parents, for example if both mother and father are homozygous AA and the offspring is heterozygous AT or homozygous TT, this represents either a *de novo* mutation or a Mendelian error. This was particularly important for family collections, since they were analysed using the transmission disequilibrium test (TDT), which takes the 'best guess' genotype from the imputation posterior probabilities. This can lead to high Mendelian error rates, particularly for variants with posterior probabilities of each genotype greater than 0, since the best guess could then not be correct for parent or child and thus a Mendelian error could be recorded. Using the TDT, false positive associations are more likely to be identified at variants with high Mendelian error rates [187], therefore in family collections, variants with a Mendelian error rate of > 0.01 were removed from the analysis.

4.2.5 Independent case-control analyses

Independent case-control analyses were carried out in four ancestry groups, excluding the EAS collection due to small sample size leading to effect estimates with very large

standard errors (38 cases and 106 controls). I adjusted for the five largest ancestry-specific genetic PCs in each model. In addition, the uncertainty around the genotype value which result from imputation was accounted for by including the genotype dose in the model rather than the best guess genotype for each variant. This means taking $(2 \times P_{AA}) + P_{AB} + (2 \times P_{BB})$ as the genotype value at that variant for each individual, where P_{XX} is the posterior probability of having genotype XX. The SNPTEST software was used to carry out these analyses [127]. I combined results from each ancestry group in an inverse-variance weighted fixed-effects meta-analysis, implemented using the METAL software [208]. The results from this analysis are referred to as the Phase I results. Forward stepwise logistic regression was performed at each associated region to identify loci with more than one independent association with T1D, noting all conditionally independent associated variants with $p < 5 \times 10^{-8}$.

Finally, case-control analyses were performed under recessive and dominant models of inheritance. To evaluate the relative fit of the three models (additive, recessive and dominant), I compared the Akaike information criterion (AIC) and identified the inheritance model providing the lowest AIC, which indicates the best fit. The AIC is a function of the number of parameters in a model and the likelihood of the model and is a way of comparing two non-nested models. The formal definition of the AIC is:

$$AIC = 2k - ln(L), \tag{4.2}$$

where k is the number of covariates included in the model and L is the likelihood of the model.

4.2.6 Transmission disequilibrium test: family-based analysis

The affected trios were also analysed by ancestry group and the TDT was used to test for association between each variant and T1D [181].

The TDT works by examining, for each family where both parents are heterozygous at a variant, how many transmissions of both alleles at that variant are observed in the child, who is known to be a T1D case. If, across a population of T1D case offspring,

the offspring inherit more of one allele over the other, this indicates the over-inherited variant may be associated with disease. The test statistic is distributed as χ^2 with 1 degree of freedom under the null hypothesis of no over-inheritance of either allele over that expected by chance. The statistic is defined as:

$$tdt = (b - c)^2/(b + c), \tag{4.3}$$

where b is the number of transmissions of allele B and c is the number of transmissions of allele C, it is evident from the numerator in Equation 4.3 that an equal transmission of allele B and C from parents to offspring will lead to a test statistic of 0 and therefore no evidence against the null hypothesis, whereas an imbalance in transmissions will increase the numerator and increase evidence against the null.

The PLINK software was used to carry out these analyses [150], and this was performed by Catherine Robertson at UVA.

The results from the TDT family analyses are referred to as the Phase II results.

4.2.7 Phase I and II Meta-analysis

From TDT summary statistics, effect sizes and standard errors were derived for each variant to enable meta-analysis with the Phase I additive case-control results, as suggested in [100]. Specifically, the log-odds ratio for the TDT can be given as:

$$log(OR_{TDT}) = log(\frac{b}{c}), \tag{4.4}$$

the log of the ratio of transmitted B alleles to transmitted C alleles, with variance:

$$Var(log(OR_{TDT})) = \frac{1}{b} + \frac{1}{c}. \tag{4.5}$$

These can be used along with the results from the Phase I analysis to calculate the Phase I and II meta-analysis results, considered the primary results of this project. However, the Phase II results were also used as a replication dataset so any novel association from the Phase I results that also reached nominal significance ($p < 0.05$) in the Phase II

results provided some evidence of replication in independent samples and hence more confidence the association was genuine.

4.2.8 Exploration of genomic inflation factor statistic

Since the ImmunoChip intentionally covers regions of the genome previously associated with immune-mediated disease, it is expected that the genomic inflation statistic, (λ), will be greater than one, since over half of the variants on the ImmunoChip lie in autoimmune-associated regions and therefore comparing the median test statistic to the null distribution is unlikely to indicate artificially inflated statistics, for example caused by population stratification, and instead simply reflect the ImmunoChip design. However, in order to assess whether the observed inflation factor was due to chip design or artificial inflation of test statistics, the inflation factor from the family-based analysis, which is robust to population structure [58, 139], was compared to the inflation factor from a subset of cases and controls from the case-control analysis with equivalent statistical power, taken as one case and one control per family trio [129]. Similarity between these inflation statistics would indicate the inflation observed is due to ImmunoChip design rather than population structure driving artificial associations. Five sets of cases and controls with equivalent power to the family-based collection were generated and the inflation factors $\lambda_i,\ i = 1, ..., 5$ were compared to the family based inflation factor, λ_{tdt}.

The sampling of cases and controls to generate these datasets was performed by Catherine Robertson.

4.2.9 Significance thresholds

The standard genome-wide significance threshold ($p < 5 \times 10^{-8}$) was used to declare significantly associated variants. However, I also examined all variants with FDR< 0.01, as the FDR procedure provides the opportunity to identify additionally associated regions, since this is a less stringent multiple comparison correction whilst still declaring a small (and pre-specified) number of false-positive associations. Since our test statistics were not all independent due to LD between variants, the FDR procedure I employed

was the Benjamini-Yekutieli FDR [20] (BY-FDR). As an additional check, I also examined regions associated to a Bonferroni-corrected signifiance threshold of 5%, which is $0.05/410234 = 1.24 \times 10^{-7}$, since association statistics were available for $410,234$ variants.

4.2.10 Fine mapping, colocalisation and haplotype analyses

GUESSFM

As has been described in chapter 2 and chapter 3, GUESSFM was used to fine map each region that was associated with T1D and also in a densely genotyped ImmunoChip region. Only European individuals were included in the analysis, since this represents a relatively ancestry homogeneous population and is therefore unlikely to prioritise variants spuriously due to differences in ancestry.

Variants were pruned to an r^2 of 0.99 prior to carrying out the stochastic search. The top 10,000 most visited models were saved for downstream GUESSFM analysis. The prior for the number of variants associated with the phenotype in each region was a binomial prior centred at 3 causal variants.

Trans-ethnic PAINTOR

In order to examine the possibility of further refining the variants prioritised from fine mapping, I included more than one ancestry collection in the fine-mapping analyses. Populations of divergent ancestry may contain recombination hotspots not present in the other ancestry group. For this reason, if a large haplotype is prioritised using GUESSFM in Europeans, it might be possible to substantially reduce the number of likely causal variants within the haplotype, since the haplotype could be broken up in a different ancestry group, with only one half of the haplotype showing disease association. However, including results from other ancestry populations that themselves have no association with T1D ($p > 5 \times 10^{-4}$) can prioritise variants in a haplotype based on a very weak association signal from an ancestry group, which might simply be prioritising statistical fluctuation. Therefore I did not include results from other ancestry populations in this analysis if the minimum association p value in the region was $> 5 \times 10^{-4}$.

A method called trans-ethnic Probabilistic Annotation INtegraTOR (PAINTOR) has been developed to harness the differences in LD structures between the ancestry divergent groups to reduce the number of potentially causally associated variants in the region [103]. The method uses z-scores from genetic association studies, where the following is assumed for each z-score of association, for variant j:

$$z_j = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)}, \tag{4.6}$$

$$\hat{\beta}_j = \frac{Cov(\mathbf{G_j}, \mathbf{Y})}{Var(\mathbf{G_j})}, \tag{4.7}$$

where $\mathbf{G_j}$ is the genotype vector at variant j and $\mathbf{Y}$ is the outcome. In addition,

$$Var(\hat{\beta}_j) = \frac{\sigma^2}{(N-1) \times Var(\mathbf{G_j})} \approx \frac{\sigma^2}{N \times Var(\mathbf{G_j})}, \tag{4.8}$$

in large sample sizes (with σ^2 the residual variance), then:

$$z_j \approx \frac{Cov(\mathbf{G_j}, \mathbf{Y})}{Var(\mathbf{G_j})} \times \frac{\sqrt{N \times Var(\mathbf{G_j})}}{\sqrt{\sigma^2}} = \frac{\hat{\beta}_j \sqrt{Var(\mathbf{G_j})}}{\sigma} \sqrt{N}, \tag{4.9}$$

which is normally distributed with variance 1 under the null hypothesis of no association between variant and disease. If there is LD between a non-causal variant k and the causal variant(s) j, the expected z-score at k will be non-zero due to the correlation between the alleles. The distribution of z-scores is therefore, in a similar way to GUESSFM, made to depend on a binary latent vector $\mathbf{C}$, where $c_j = 1$ if variant j is causal and 0 if not. The objective is to identify the marginal posterior distribution of $\mathbf{C}$. This is done by, for each combination of the vector $\mathbf{C}$, first calculating non centrally parameters of the expected distribution of z-scores, for SNPs with known LD with the assumed causal variant(s) under consideration. If a parameter λ_j is equal to the z-score at variant j if the $c_j = 1$ at that variant and 0 otherwise, then the expected non-centrality parameter of another variant under consideration in the analysis, k, is $\lambda_j \times r_{j,k}$, which is the z-score for variant j multiplied by the correlation between the assumed non-causal variant k and the assumed

causal variant j, $r_{j,k}$. More generally, and summing across more than one causal variant if $\mathbf{C}$ contains more than one non-zero element, the non-centrality parameter, $\mathbf{\Lambda}$ is defined as, for variant j, $\Lambda_j = \sum_c r_{j,c}\lambda_c$.

For each combination of $\mathbf{C}$ therefore $\mathbf{\Lambda}$ will vary depending on $\mathbf{C}$ but it can be quickly enumerated by:

$$\mathbf{\Lambda} = \mathbf{\Sigma}(\boldsymbol{\lambda} \circ \mathbf{C}), \tag{4.10}$$

where $\mathbf{\Sigma}$ is the correlation, r, between each variant under consideration and $\circ$ is the element-wise multiplication between the two vectors.

Crucially, the LD matrix and association z-scores in each ancestry population, p, are likely different, therefore the expected non-centrality parameter for each population is defined as:

$$\boldsymbol{\Lambda_p} = \boldsymbol{\Sigma_p}(\boldsymbol{\lambda_p} \circ \mathbf{C}), \tag{4.11}$$

where the 'causal vector', $\mathbf{C}$, is assumed identical across populations. The distribution of z-scores across all populations can be expressed as a multivariate normal distribution with a non-centrality parameter equal to $\boldsymbol{\Lambda_p}$:

$$\mathbf{Z_p}|\boldsymbol{\lambda_p}, \mathbf{C_p} \sim MVN(\boldsymbol{\Lambda_p}, \boldsymbol{\Sigma_p}). \tag{4.12}$$

The likelihood function multiplies the probability of each ancestry group together to obtain the overall likelihood. In an analysis of P populations, this is:

$$L(\mathbf{Z_1}, ..., \mathbf{Z_P}, \boldsymbol{\lambda_1}, ..., \boldsymbol{\lambda_P}) = \sum_{\mathbf{C} \in \mathbf{C}_{all}} \prod_p P(\mathbf{Z_p}|\boldsymbol{\lambda_p}, \mathbf{C}) P(\mathbf{C}), \tag{4.13}$$

with $P(\mathbf{C})$ the prior probability of the causal set under consideration, $\mathbf{C}$, which can be made to depend on functional data, but I chose not to do this since my focus is on the benefits of trans-ethnic fine mapping using genetic data alone rather than the integration of other data sources. Instead, I chose an uninformative binomial prior so each variant

has equal probability of causality *a priori*.

The marginal posterior probability of each causal set is then determined by:

$$P(\mathbf{C}|\mathbf{Z_1},...,\mathbf{Z_P},\boldsymbol{\lambda_1},...,\boldsymbol{\lambda_P}) = \frac{\prod_P P(\mathbf{Z_p}|\boldsymbol{\lambda_p},\mathbf{C})P(\mathbf{C})}{\sum_{\mathbf{C}\in\mathbf{C}_{all}}\prod_P P(\mathbf{Z_p}|\boldsymbol{\lambda_p},\mathbf{C})P(\mathbf{C})}, \tag{4.14}$$

This likelihood can be calculated explicitly for each causal set of variants (up to a maximum number of causal variants in the region), or alternatively the likelihood can be sampled in an MCMC scheme that explores different combinations of causal variants.

Haplotype analyses

Haplotype analyses were performed as in chapter 2 for each region so regions with more than one T1D-associated haplotype could be visualised. Any index variant with GUESSFM group posterior probability > 0.5 was included in the analysis, as well as the index variant in univariable analyses if that variant had group posterior probability < 0.5 in the fine-mapping analysis, since this could indicate the index variant tags two or more disease-associated haplotypes. Haplotype analyses show the effect of the index variant in isolation without the disease-associated haplotypes it is in LD with, allowing visualisation of the tag effect.

4.2.11 ATAC-seq data processing

Publicly available ATAC-seq data were downloaded for 24 immune cell types [32] as well as islets [196] and a control cell type, fibroblasts [94], a connective tissue cell type that has no obvious link to T1D.

I also used in-house generated ATAC-seq data for various cell types, both resting and under various stimuli. Table 4.1 lists all cell types examined in this analysis.

Table 4.1: ATAC-seq data used in analysis.

Cell type	N	Source	Stimulus	Called
$CD4^+$ T cells unstim 240 minutes resting	7	In house		In_house-CD4
$CD4^+$ T cells stim 240 minutes	6	In house	anti CD3/CD28 beads	In_house-CD4
Naive T effector cells unstim	4	Calderon		Calderon-Naive_Teffs
Naive T effector cells stim 24 hours	5	Calderon	anti CD3/CD28 beads, human IL-2	Calderon-Naive_Teffs
Effector $CD4^+$ unstim	4	Calderon		Calderon-Effector_CD4pos_T
Effector $CD4^+$ stim 24 hours	3	Calderon	anti CD3/CD28 beads, human IL-2	Calderon-Effector_CD4pos_T
Memory T effector cells unstim	4	Calderon		Calderon-Memory_Teffs
Memory T effector cells stim 24 hours	4	Calderon	anti CD3/CD28 beads, human IL-2	Calderon-Memory_Teffs
Regulatory T cells unstim	4	Calderon		Calderon-Regulatory_T
Regulatory T cells stim 24 hours	4	Calderon	anti CD3/CD28 beads, human IL-2	Calderon-Regulatory_T
Naive regulatory T cells unstim	2	Calderon		Calderon-Naive_Tregs
Naive regulatory T cells stim 24 hours	2	Calderon	anti CD3/CD28 beads, human IL-2	Calderon-Naive_Tregs
Memory regulatory T cells unstim	4	Calderon		Calderon-Memory_Tregs
Memory regulatory T cells stim 24 hours	4	Calderon	anti CD3/CD28 beads, human IL-2	Calderon-Memory_Tregs
T Follicular helper cells unstim	5	Calderon		Calderon-Follicular_T_Helper
T Follicular helper cells stim 24 hours	4	Calderon	anti CD3/CD28 beads, human IL-2	Calderon-Follicular_T_Helper

Central memory $CD8^+$ T cells unstim	4	Calderon		Calderon-Central_memory_CD8pos_T
Central memory $CD8^+$ T cells stim 24 hours	4	Calderon	anti CD3/CD28 beads, human IL-2	Calderon-Central_memory_CD8pos_T
$CD8^+$ T cells unstim	4	Calderon		Calderon-CD8pos_T
$CD8^+$ T cells stim 24 hours	3	Calderon	anti CD3/CD28 beads, human IL-2	Calderon-CD8pos_T
Naive $CD8^+$ T cells unstim	4	Calderon		Calderon-Naive_CD8_T
Naive $CD8^+$ T cells stim 24 hours	4	Calderon	anti CD3/CD28 beads, human IL-2	Calderon-Naive_CD8_T
Effector memory $CD8^+$ T cells unstim	4	Calderon		Calderon-Effector_memory_CD8pos_T
Effector memory $CD8^+$ T cells stim 24 hours	4	Calderon	anti CD3/CD28 beads, human IL-2	Calderon-Effector_memory_CD8pos_T
Th1 precursors unstim	4	Calderon		Calderon-Th1_precursors
Th1 precursors stim 24 hours	4	Calderon	anti CD3/CD28 beads, human IL-2	Calderon-Th1_precursors
Th2 precursors unstim	4	Calderon		Calderon-Th2_precursors
Th2 precursors stim 24 hours	4	Calderon	anti CD3/CD28 beads, human IL-2	Calderon-Th2_precursors
Th17 precursors unstim	3	Calderon		Calderon-Th17_precursors
Th17 precursors stim 24 hours	4	Calderon	anti CD3/CD28 beads, human IL-2	Calderon-Th17_precursors
Gamma delta T cells unstim	4	Calderon		Calderon-Gamma_delta_T
Gamma delta T cells stim 24 hours	3	Calderon	anti CD3/CD28 beads, human IL-2	Calderon-Gamma_delta_T
Plasmablasts unstim	3	Calderon		Calderon-Plasmablasts
Bulk B cells unstim	4	Calderon		Calderon-Bulk_B
Bulk B cells stim 24 hours	3	Calderon	F(ab)‘2 anti IgG/IgM, human IL-4	Calderon-Bulk_B

Naive B cells unstim	4	Calderon		Calderon-Naive_B
Naive B cells stim 24 hours	3	Calderon	F(ab)‘2 anti IgG/IgM, human IL-4	Calderon-Naive_B
Memory B cells unstim	4	Calderon		Calderon-Mem_B
Memory B cells stim 24 hours	4	Calderon	F(ab)‘2 anti IgG/IgM, human IL-4	Calderon-Mem_B
B cells unstim	4	In house		In_house-B
B cells stim 4 hours	4	In house	anti-Ig, CD40L mega IL-21, IL-4	In_house-B
Immature NK cells	5	Calderon		Calderon-Immature_NK
Mature NK cells unstim	4	Calderon		Calderon-Mature_NK
Mature NK cells stim 48 hours	6	Calderon	human IL-2	Calderon-Mature_NK
Memory NK cells	6	Calderon		Calderon-Memory_NK
Monocytes unstim	3	Calderon		Calderon-Monocytes
Monocytes stim 6 hours	6	Calderon	LPS	Calderon-Monocytes
Dendritic Cells unstim	3	Calderon		Calderon-pDCs
Myeloid Dendritic Cells unstim	3	Calderon		Calderon-Myeloid_DCs
Islets	2	Varshney		Varshney-Islets
Fetal fibroblasts	1	Jonsson		Jonsson-Cardiac_Fibroblasts-f
Adult fibroblasts	1	Jonsson		Jonsson-Cardiac_Fibroblasts-a

Reads were mapped to genome genome build 38 (GRCh38.p12) using minimap2 [110] and peaks were called for each cell type using macs2 [70].

A consensus list of peaks, which were used in differential-expression analyses, were created from all cell types in the Calderon publication [32]. To do this, for each cell type, the number of reads were counted from the donor contributing the fewest number of reads for that cell type, then divided this count by two. The reads were then randomly pooled by that number of reads from all the samples for that cell type, and thus creating a representative alignment file for that cell type. This was done two times in order to have two pseudo-replicates. Peaks were called from these two pseudo replicates using macs2 [70]. A measure of consistency of each peak across the two pseudo replicates was calculated, called the irreducible discovery rate (IDR) [111]. Any peak sufficiently consistent between pseudo replicates (IDR ≥ 0.05) was included in the consensus list of peaks. This list was then used as a feature reference and reads were counted in every cell type using featureCounts [114].

In addition to these datasets, where typically a small number of individuals are included in the analyses, ATAC-seq data in total $CD4^+$ T cells were available for a subset of 123 individuals without T1D from the MEGAT1D dataset.

David Flores implemented the ATAC-seq processing pipeline. In-house data were generated by Dr. Anthony Cutler.

4.2.12 Enrichment of prioritised T1D variants with ATAC-seq peaks

In order to determine whether or not the variants prioritised by GUESSFM had regulatory function in a particular cell type, the ATAC-seq data were used and if a variant overlapped an ATAC-seq peak, it was inferred that this variant is more likely functional in the cell type under consideration. If the number of variants prioritised by GUESSFM overlapped open chromatin more often than expected by chance, it would imply that the cell type under consideration might be important in the aetiology of T1D.

To examine enrichment of the prioritised variants in each cell type, I chose to compare

the overlap of GUESSFM prioritised variants with a null distribution of overlaps across the genome. However, the variants within groups from GUESSFM are in high LD with each other and I wanted to mimic this LD structure in the variants I was comparing to, in order to ensure there was no bias in the enrichment analysis due to differences in LD structure between GUESSFM prioritised variants and a randomly sampled selection of variants across the genome.

I therefore used 1000 Genomes Project data, analysing European ancestry individuals only, and calculated Pearson correlations between all variants using a software called LDstore, which allows these correlations to be calculated in a highly parallel way (`http://www.christianbenner.com/`). I kept all correlations with absolute Pearson $r > 0.8$ which I used as a proxy for a haplotype and the number of variants with $r > 0.8$ with an index variant I assumed was the haplotype size.

I then took each group with posterior probability sum > 0.8 from GUESSFM and assigned a bin size to these groups, with bin sizes 1-9, 10-19, 20-49, 50-74, 75-99, 100-149 and 150-250. For each group, I matched the variants to a haplotype from the 1000 Genomes Project data of the same bin size, so the total number of variants sampled from the 1000 Genomes Project data were similar in size to the prioritised variants from GUESSFM and also mimicked the approximate LD structure. In addition, ATAC-seq peaks are often present in the proximity of promoters and enhancers, which are usually physically close to genes. Therefore, if the randomly sampled regions of the genome contained less genes than the regions containing the GUESSFM prioritised variants, ATAC-seq enrichment could be observed due to differences in the number of genes in the observed set and the randomly sampled set of variants. To minimise the probability of this occurring, once the randomly sampled regions of similar LD structure were identified, I filtered them so the regions overlapped the same number of genes as the observed GUESSFM prioritised variant set overlapped (or a similar number of overlaps for GUESSFM prioritised regions overlapping many genes). For each GUESSFM prioritised group, the comparator randomly sampled group from across the genome was then sampled from this filtered set of regions with similar gene density to the GUESSFM prioritised group. I exam-

ined how many of these randomly sampled variants from the 1000 Genomes Project data overlapped open chromatin in each cell type.

I did this 100 times, so the results did not suffer from a random bias in the sampling procedure. For each of the 100 randomly sampled datasets, I performed a Fisher's exact test between the number of variants in open chromatin in the GUESSFM prioritised variants and the number of variants in open chromatin from the randomly sampled dataset. The mean z-score across all 100 tests was then compared to the normal distribution to obtain a two sided enrichment p value.

This procedure was carried out for each cell type and this allowed inference about cell types likely involved in T1D aetiology. To test the enrichment analysis procedure was performing appropriately, I included fibroblasts as a cell type in the enrichment analyses in order to have a negative control where I expected no enrichment of T1D-associated variants.

4.2.13 Differential openness between activated and non-activated cells overlapping potentially causal T1D variants

Since chromatin conformation can vary depending on the activation state of the cell, I wanted to examine whether or not there were differences in the ATAC-seq counts between resting and stimulated cells in those peaks that overlap prioritised variants from GUESSFM. This was motivated by the observation that if there is an ATAC-seq peak in the stimulated cells, there is likely an ATAC-seq peak in the unstimulated cells, therefore the enrichment by examining overlapping peaks with T1D prioritised variants would be similar between conditions, but the number of reads observed between states was often different and this might be important to examine. An increase in read count upon activation might point towards this variant acting in activated cells and a decrease in read count upon activation may suggest the opposite, that the variant is perturbing disease risk in non-activated cells.

I therefore examined differential-openness of every peak in the set of consensus list of peaks across immune cells to highlight those peaks where the read count was signifi-

cantly different between stimulation conditions. I then took every peak that was considered differentially-open and examined enrichment of GUESSFM prioritised T1D variants in these stimulation-specific peaks, using the same enrichment analysis as the analysis examining all peaks regardless of read count.

DESeq2

A method called DESeq2 [120], was used to assess differential-openness of peaks across conditions. In a setup with J individuals, j, and I ATAC-seq peaks, i, the counts of each peak are modelled as a negative binomial distribution with mean μ_{ij} and dispersion parameter α_i. The reason the counts are modelled as a negative-binomial distribution over a Poisson, which might be the natural choice for count data, is that the Poisson distribution assumes the mean is equal to the variance, whereas it is expected that the variance will be larger than the mean, since this is often the case for RNA-seq data [162] and therefore type 1 error rates would be increased from assuming a Poisson distribution. To account for the fact that each individual, or sample, may have been sequenced to different depths, first the counts are normalised to account for this by being divided by a individual/sample-specific normalising factor, s_j. This normalising factor is taken to be the median value across all peaks of the count at peak i, K_i divided by the expected count at peak i across all samples:

$$s_j = median \frac{K_{ij}}{K_i^R}, \tag{4.15}$$

where $K_i^R = (\prod_{j=1}^{J} K_{ij})^{1/J}$. The median ratio is taken since if the sum of counts was divided by the total expected sum of counts, this could be influenced heavily by peaks with lots of reads.

The normalised count, $q_{ij} = \frac{K_{ij}}{s_j}$ is then used as the outcome in a generalised linear model with log_2 link function, so:

$$log_2 q_{ij} = \sum_r x_{jr} \beta_{ir}, \tag{4.16}$$

where r is the number of covariates included in the model, x_{jr} is the value of these covariates for individual j and β_{ir} is the parameter estimate effect size for each covariate r on the counts in peak i.

An empirical Bayes method is used to fit Equation 4.16, meaning the data is used to estimate the prior distribution for the parameters of interest. Two sets of parameters are estimated, the effect sizes of the stimulation coefficient for each peak, β_i and the dispersion parameter for each peak, α_i. The dispersion parameter relates to the variance of the negative binomial distribution, $Var(q_i) = \mu_i + \alpha_i \mu_i^2$, with μ_i the mean normalised count, $\bar{q}_i$ across samples, j. Therefore the β_i estimates depend on the dispersion parameter, which can be unreliably estimated in small sample sizes. Accurate estimation of this parameter is crucial to accurate estimation of β_i, therefore information is shared across all peaks and a prior distribution for the dispersion parameter is estimated as a function of the mean count at the peak.

The prior distribution for α_i is assumed to be log-normal, but the mean of the distribution is made to depend on two hyper-parameters, α_0 and a_1, which are estimated across all peaks from the data itself. The way these are estimated is by first obtaining peak-wise dispersion parameter estimates, α_i^{PW}, by maximising the likelihood function from fitting model 4.16 with respect to α_i, conditional on a set of fitted values for μ_i^0 which were fitted from the same generalised linear model but using an initial estimate for α_i^0, based on within group variances and means.

Once a set of peak-wise dispersion parameters are estimated, α_i^{PW} they are regressed in a generalised linear model with gamma link function against mean normalised count $\bar{\mu}_i$ as:

$$g(\alpha_i^{PW}) = \alpha_0 + a_1 \bar{\mu}_i, \tag{4.17}$$

with link function $g(\alpha_i^{PW}) = -\frac{1}{\alpha_i^{PW}}$. Through fitting the model, the hyper-parameters α_0 and a_1 are estimated, which describe the expected relationship across all peaks between the dispersion parameter and mean count, which influences the prior distribution of α_i:

$$log\alpha_i \sim N(log(\frac{a_1}{\bar{\mu}_i} + \alpha_0), \sigma_d^2). \tag{4.18}$$

Here, σ_d is taken as the difference between the variance of the log dispersion estimates across all peaks and the expected sampling variance of the log dispersion estimates. When combined with the likelihood, this prior has the effect of shrinking the dispersion parameter estimates, α_i, towards the expected value for a peak with that count across all peaks, meaning if a dispersion parameter is very high given the mean count at that peak, it will remain high but be reduced towards the prior value given this estimate could be unstable due to low counts at that peak.

Similarly, the parameter estimates for β_i are calculated using empirical Bayes, with the normal prior assumed for β_i with a prior variance parameter estimated from the data, σ_r^2, $\beta \sim N(0, \sigma_r^2)$. This is done to shrink the effect size estimates towards zero, and the level of shrinkage adjusts to the dispersion parameter (high dispersion leads to increased shrinkage) and mean count at the peak (smaller count leads to increased shrinkage due to instability of effect estimate at low counts). To estimate σ_r^2, first the maximum likelihood estimates for β_i are calculated, conditional on the maximum *a posteriori* (MAP) estimate for the α_i parameters. The σ_r value is then set so that the 0.95 quantile value of the distribution of β_i estimates, $Q_{|\beta_{MLE}|}$, matches the 0.975 theoretical quantile from $N(0, \sigma_r^2)$, Q_N, so:

$$\sigma_r = \frac{Q_{|\beta_{MLE}|} \times 0.95}{Q_N \times 0.975}. \tag{4.19}$$

This ensures the prior variance around β_i is slightly larger than that estimated from the data itself. This prior for β_i is then combined with likelihood to obtain MAP values for each β_i. Association tests are then performed using a Wald test, dividing the MAP β_i estimate by the standard error around it, and comparing this to the standard normal distribution.

Colocalisation analyses

Colocalisation analyses were carried out as in chapter 3, but with the T1D association summary statistics combined with chromatin accessibility QTL (caQTL) summary statistics. These caQTL summary statistics were generated by first normalising the counts for sequencing depth and then log transforming the values so they were approximately normally distributed. Then, for each peak, a linear model was fitted with the normalised peak as the outcome and batch as a covariate as well as the genetic variant under consideration. The association of each peak with every variant was performed so that each peak had a corresponding set of variant association summary statistics. Colocalisation between the caQTLs and the T1D summary statistics from $CD4^+$ T cells would indicate potential mechanism behind the association with disease. In addition, once the colocalisation between caQTL and T1D summary statistics were examined, I looked at colocalisation between T1D summary statistics and *cis* eQTLs from a whole-blood eQTL study of over $30,000$ individuals [199].

Catharine Robertson carried out these analyses at UVA.

4.3 Results

4.3.1 Study population

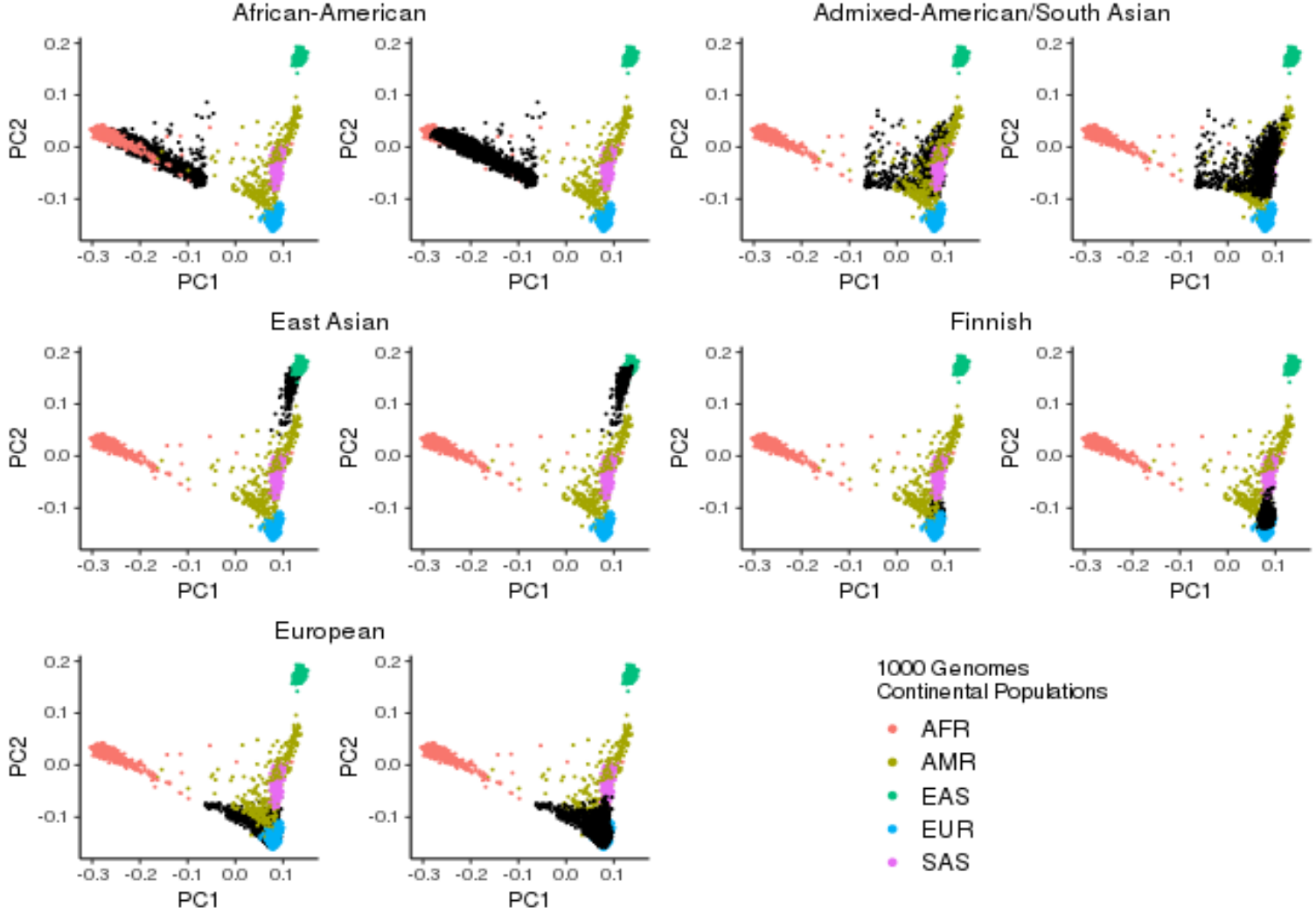
Table 4.2 shows the number of individuals in each ancestry cluster and Figure 4.1 shows PCA plots from each ancestry group in the analysis projected onto the 1000 Genomes Project PCs, where the genetically-determined ancestry groups from the MEGAT1D datasets cluster with the 1000 Genomes Project self-reported ancestry groups. Figure 4.1 and Table 4.2 were produced by Catherine Robertson at UVA.

Table 4.2: Individuals included in the MEGAT1D analysis.

Cohort	Country	Total	Cases	Controls	Missing	AFR	AMR	EAS	EUR	FIN
BC58	United Kingdom	6553	0	6548	5	0	0	0	6549	4
BDA	United Kingdom	1481	784	697	0	0	0	0	1481	0
Cambridge	United Kingdom	2941	0	2941	0	13	27	5	2892	4
GRID	United Kingdom	6679	6673	6	0	5	10	1	6661	2
UKBS	United Kingdom	2904	0	2904	0	3	6	4	2891	0
T1DGC-UK	United Kingdom	738	398	337	3	8	9	2	719	0
BRI	United States	1299	572	727	0	27	61	80	1125	6
CLEAR	United States	379	0	379	0	377	1	0	1	0
EDIC	United States	1415	465	950	0	9	17	9	1373	7
GoKinD	United States	2816	1779	1037	0	52	44	7	2694	19
NYCP	United States	1541	0	1541	0	792	10	0	737	2
NIMH	United States	1656	0	1656	0	0	2	0	1652	2
SEARCH	United States	855	855	0	0	233	339	13	268	2
TrialNet	United States	1725	1030	695	0	41	128	4	1525	27
T1DGC-NA	United States	6808	3284	3508	16	1496	907	10	4374	21

T1DGC	United States	6	2	4	0	6	0	0	0	0
Univ. of Alabama	United States	1432	0	1432	0	622	8	0	801	1
Univ. of Colorado	United States	99	67	32	0	86	11	0	2	0
UCSF	United States	481	0	481	0	464	9	0	8	0
IDDMGEN	Finland	3435	1061	2374	0	3	0	0	50	3382
T1DGEN	Finland	246	246	0	0	0	0	0	4	242
MILWAUKEE	Finland	2915	937	1276	702	0	0	2	37	2876
T1DGC-NI	Northern Ireland	502	501	1	0	0	0	0	502	0
T1DGC-YH	Northern Ireland	477	0	477	0	0	1	1	474	1
DAN-SDC	Denmark	3648	1693	1955	0	12	39	5	3566	26
IC-Cases-Steno	Denmark	1048	1048	0	0	0	0	0	1047	1
IC-Cases-HSG	Denmark	202	202	0	0	1	5	4	187	5
T1DGC-EUR	Other Europe	4931	2615	2294	22	40	5	0	4526	360
HapMap	Other Europe	6	0	0	6	0	0	0	6	0
T1DGC-AP	Asia Pacific	2209	981	1224	4	0	600	441	1167	1
Total		61427	25193	35476	758	4290	2239	588	47319	6991

Figure 4.1: The largest versus second-largest genetic PCs projected onto the 1000 Genomes Project PCs. MEGAT1D data coloured black, data from 1000 Genomes Project coloured by ancestry group. Two plots shown for each ancestry group, one with the 1000 Genomes Project data at the front, the other with the data in this study at the front.



4.3.2 Variant Quality control

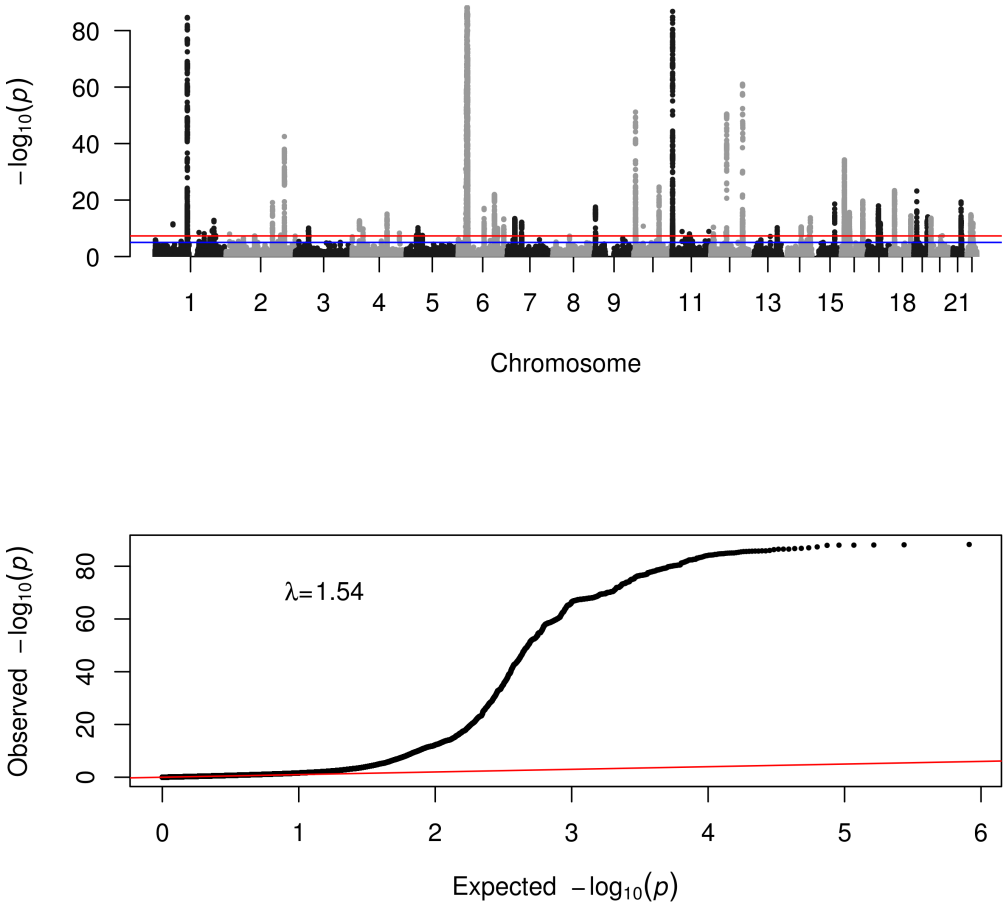
In total, 59,974 individuals, including 41,545 unrelated individuals (16,159 cases and 25,386 controls) and 6143 family trios with an affected offspring were included in the analysis. This is different to the number of individuals included in Table 4.2, since this table includes individuals from one of the two following scenarios: i) if a family has two children, one with T1D and the other without, the child without T1D is included in Table 4.2 but not the association testing and ii) if a family has two children and both are T1D cases, the parents are included twice in the association tests and the transmissions to the two children are counted as two separate family trios in the association TDT analysis.
After post-imputation variant QC, 645,918 variants were examined in the European population, 1,141,842 in the African population, 684,512 in the Finnish population and 588,195 variants in the admixed population.

4.3.3 ImmunoChip-wide meta-analysis

Phase I

In an initial Phase I analysis, including 41,545 unrelated individuals, 16,159 cases and 25,386 controls, 60 regions outside the HLA region were associated with T1D to genome-wide significance ($p < 5 \times 10^{-8}$), including 11 novel associated regions. The Manhattan and QQ plots for this analysis can be seen in Figure 4.2.

Figure 4.2: Manhattan and Quantile-Quantile plot from Phase I results.



When examining forward stepwise conditional regression, there were an additional 16 conditionally independent genome-wide associated variants, taking the total number of conditionally independent signals to 76. Namely, eight regions with evidence for two independent signals (near the candidate causal genes *IL10*, *COBL/IKZF1*, *INS*, *IKZF4*, *ZFP36L1*, *DEXI*, *PTPN2* and *TYK2*), two regions with three independent signals (near the candidate causal genes *IFIH1* and *CTLA4*), and one region with five independent signals (near the candidate causal gene *IL2RA*).

After fitting each variant in models under the inheritance assumption of recessive and dominance, 11 variants showed a better model fit than the additive model as measured by a lower AIC. One of these showed a better fit under the recessive model, near the candidate causal gene *GSDMB*, the other ten variants showed a better fit under the

dominance model (Table 4.3). This comparison was in Europeans only, since the AIC cannot be obtained from the meta-analysis of all ancestry groups. For these 11 variants only, to assess whether or not they truly had dominance effects, I fitted a general genetic model, with a parameter for each genotype, and compared the AIC of these variants under the general model to the AIC under a dominance or recessive mode of inheritance. For every variant except for rs8074437, near *GSDMB*, the AIC was lowest for the dominance model. This is not proof that these variants act in a dominant manner, but of the models examined, they are most consistent with a dominant mode of inheritance.

Table 4.3: Variants where models assuming dominant or recessive modes of inheritance have a lower AIC than the additive model.

ID	chromosome	position	candidate gene	A1	A2	DOM/REC	AIC: DOM/REC	AIC: ADD
rs3024493	1	206770623	*IL10*	C	A	Dominant	45155.2	45156.6
rs35667974	2	162268127	*IFIH1*	T	C	Dominant	45136.0	45138.4
rs10933559	2	241468331	*FARP2*	A	G	Dominant	45192.2	45194.0
rs1018942	7	50960703	*COBL*	T	C	Dominant	45159.6	45160.8
rs722988	10	33137219	*NRP1*	T	C	Dominant	45176.6	45179.6
rs11170466	12	53192075	*ITGB7*	C	T	Dominant	45182.6	45183.2
rs2057456	13	99424191	*GPR183*	C	T	Dominant	45184.2	45185.2
rs151234	16	28494339	*IL27*	G	C	Dominant	45150.2	45155.6
rs425105	19	46705224	*PRKD2*	T	C	Dominant	45166.6	45169.4
rs281377	19	48703346	*FUT2*	C	T	Dominant	45170.8	45173.2
rs8074437	17	39919884	*GSDMB*	T	G	Recessive	45162.2	45163.8

Phase II

In Phase II, the family-based analysis, 12 regions outside the HLA region were associated to genome-wide significance, all of which were genome-wide associated in the Phase I results. The Manhattan and QQ plot for the Phase II results can be found in Figure 4.3. Of the 11 novel associations identified in the Phase I results, four showed some association in the Phase II results ($p < 0.05$), albeit with lower statistical power (median (IQR) = $0.88(0.86, 0.90)$), calculated assuming the same effect size as from the Phase I results, which may be over-inflated due to the winner's curse. Additionally, case-control (Phase I) and family-based (Phase II) association statistics were concordant in magnitude and direction of effect across all but three loci, rs10815223 (C>T) near the Plasminogen Receptor, C-Terminal Lysine Transmembrane (PLGRKT) gene, rs34536443 ($G > C$) near the Tyrosine Kinase 2 (TYK2) gene and rs6072275 (G>A) near the Lipin 3 (LPIN3) gene (Figure 4.4). The variants near *PLGRKT* and *LPIN3* showed little association in the Phase II analyses, with a reasonable number of observed transmissions from heterozygous parents at these alleles ($N > 1000$), indicating these variants might not be associated with T1D and the Phase I results are false positives. However, the variant near *TYK2* only passed QC in the African ancestry group in the Phase II analysis, and a total of three transmissions were observed from heterozygous parents at this allele in the African ancestry group, so the Phase II effect estimate in this group is extremely unstable and potentially misleading. Therefore this variant is likely truly associated with T1D but the effect estimate in the Phase II analysis is likely not accurate due to small sample size, this is the reason why there appears to be an outlier in the top left hand corner of Figure 4.4.

Figure 4.3: Manhattan and Quantile-Quantile plot from Phase II results.

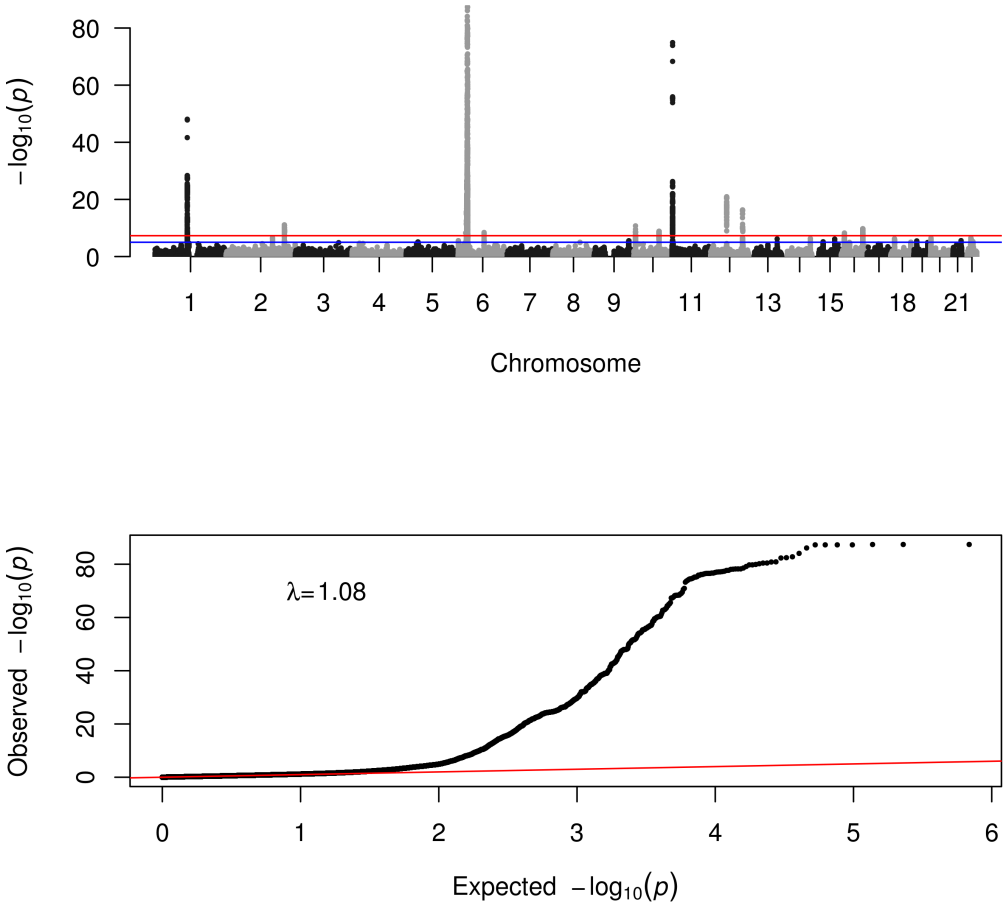
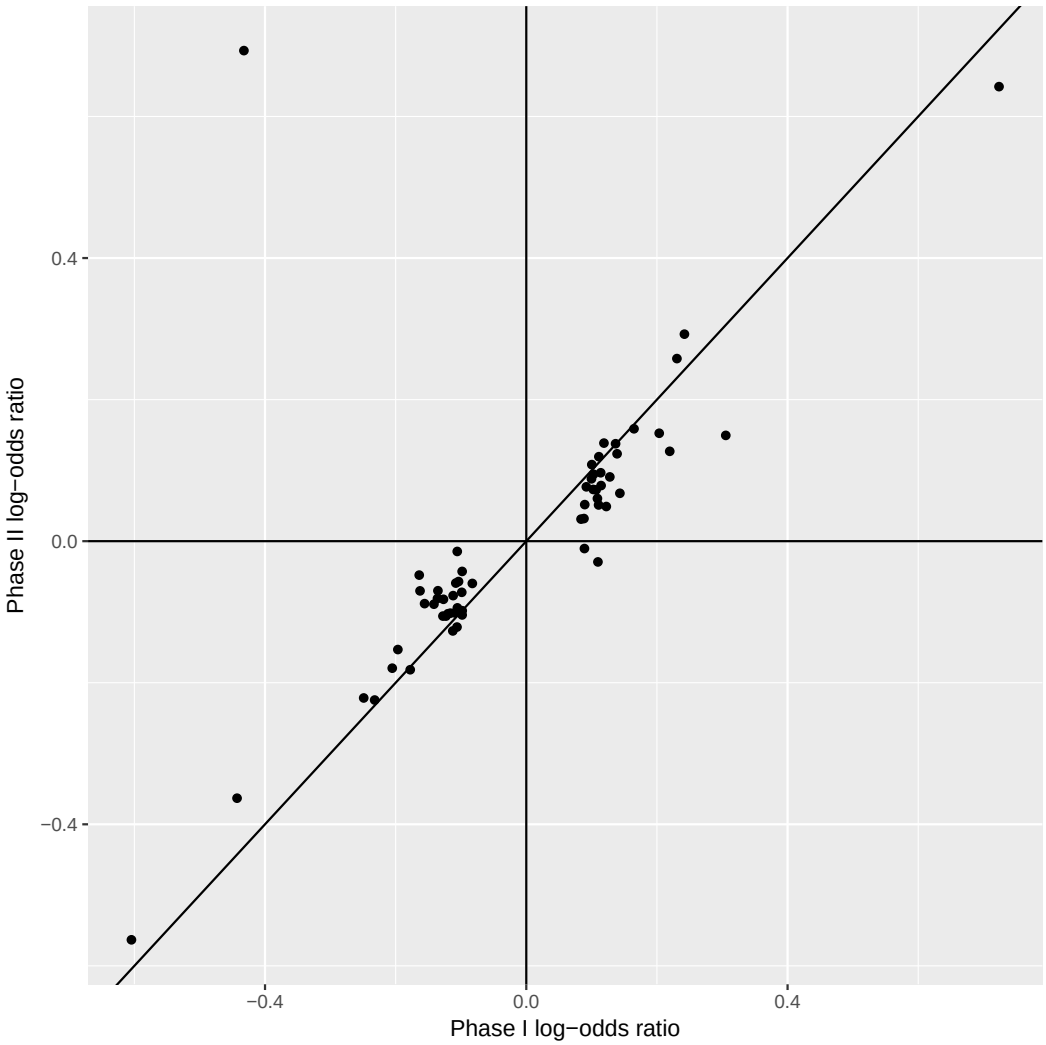


Figure 4.4: Phase I versus Phase II log-odds ratio for genome-wide significant variants from Phase I.



Phase I and II meta-analysis

When Phase I and Phase II were combined in an inverse-variance weighted fixed-effect meta-analysis, 74 regions were associated with T1D ($p < 5 \times 10^{-8}$), including 22 novel disease-associated regions (Figure 4.5). Of these 22 novel regions, eight resided in densely genotyped ImmunoChip regions. In addition, eight of these variants are in LD ($r^2 > 0.95$ according to the 1000 Genomes Project European population) with variants associated with other autoimmune diseases (Table 4.4).

Figure 4.5: Manhattan and Quantile-Quantile plot from Phase I and Phase II results combined.

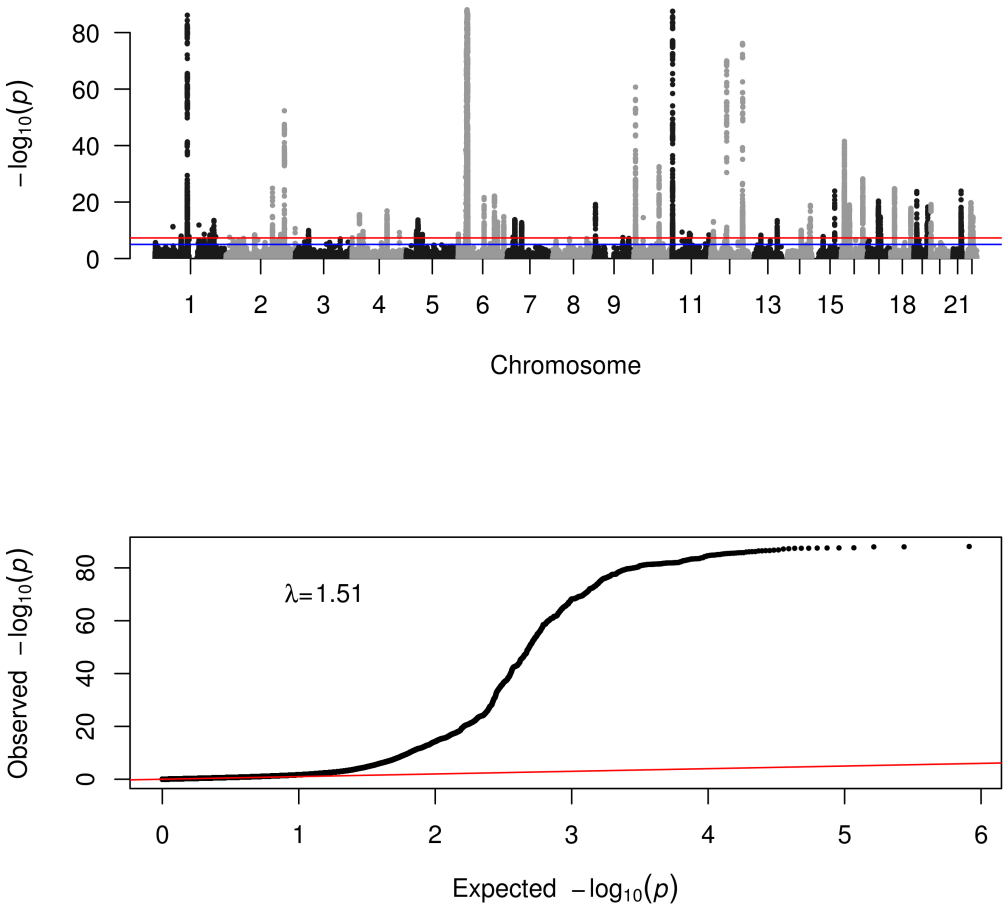


Table 4.4: Novel chromosome regions associated with T1D. AF=allele frequency of A2 allele. Beta estimates are for the additive addition of A2. iChip=in densely genotyped ImmunoChip region. SLE=systemic lupus erythematosus. CD=coeliac disease. T2D= type 2 diabetes. PBC=primary biliary cholangitis. RA=rheumatoid arthritis. HTD=hypothyroidism.

Chr	Position	A1	A2	ID	Candidate gene	AF	iChip	Beta_{META}	P_{META}	Other associations
1	92358141	G	C	rs34090353	*RPAP2*	0.361	Yes	0.075	1.1×10^{-8}	
1	119895261	A	G	rs2641348	*NOTCH2*	0.107	No	0.107	1.61×10^{-8}	Crohn's, T2D.
1	154465420	T	C	rs2229238	*IL6R*	0.813	No	-0.110	1.38×10^{-12}	
1	192570207	G	A	rs2816313	*RGS1*	0.719	Yes	0.086	4.57×10^{-9}	
1	212796238	G	T	rs11120029	*TATDN3*	0.147	No	0.098	1.82×10^{-8}	
2	12512805	C	T	rs10169963	*TRIB2*	0.58	No	0.071	2.78×10^{-8}	HTD.
2	191105394	C	G	rs7582694	*STAT4*	0.773	Yes	-0.088	2.83×10^{-9}	HTD, SLE, PBC, RA, CD.
2	241468331	A	G	rs10933559	*FARP2*	0.208	No	0.104	2.39×10^{-11}	
4	960459	G	T	rs3733345	*TMEM175*	0.53	No	0.070	4.39×10^{-8}	SLE, Parkinson's, blood protein levels.
4	38602849	G	A	rs337637	*KLF3*	0.364	No	-0.085	2.57×10^{-10}	
5	40521603	G	T	rs1876142	*PTGER4*	0.658	Yes	-0.100	2.18×10^{-14}	
6	424915	C	A	rs9405661	*IRF4*	0.514	Yes	0.077	2.26×10^{-9}	HTD.
7	20557306	A	G	rs17143056	*ABCB5*	0.183	No	-0.095	2.44×10^{-8}	
7	28102567	G	T	rs10245867	*JAZF1*	0.331	Yes	-0.074	3.15×10^{-8}	Hayfever, SLE, eczema.
9	99823263	T	C	rs1405209	*NR4A3*	0.375	No	0.072	3.45×10^{-8}	
11	35285601	T	C	rs3763888	*SLC1A2*	0.392	No	0.080	4.41×10^{-10}	Vitiligo.
11	60961822	G	T	rs79538630	*CD5/CD6*	0.035	Yes	0.193	1.14×10^{-9}	
11	61828092	C	T	rs968567	*FADS2*	0.177	No	-0.102	8.42×10^{-9}	RA, blood metabolite levels, asthma.
11	128734337	G	T	rs605093	*FLI1*	0.47	No	0.075	4.25×10^{-9}	
12	8942630	T	C	rs1805731	*M6PR*	0.389	No	0.070	4.15×10^{-8}	
13	42343795	C	T	rs74537115	*AKAP11*	0.141	Yes	0.103	5.41×10^{-9}	
17	47956725	A	G	rs2597169	*PRR15*L	0.348	No	0.078	3.35×10^{-9}	

False discovery rates

If the significance threshold is relaxed from the conservative genome-wide threshold of 5×10^{-8} and instead variants are considered associated if they have BY-FDR < 0.01, 146 regions showed association with T1D. The absolute log-odds ratios for variants reaching the genome-wide threshold were, as expected, larger than those satisfying BY-FDR< 0.01 with $p > 5 \times 10^{-8}$ but with similar MAFs (median (IQR) OR = 1.13 (1.10, 1.14) versus 1.07 (1.06, 1.08); median (IQR) MAF= 0.314 (0.179, 0.382), verses 0.308 (0.187, 0.398)), indicating that there are likely many more regions associated with T1D with increasingly smaller effect sizes. Figure 4.6 shows each of the 146 independent associations to BY-FDR < 0.01 and it can be observed that those BY-FDR < 0.01 variants with $p > 5 \times 10^{-8}$ have smaller effect sizes, with the exception of rare variants, which require large effect sizes to have sufficient statistical power in order to show significant association with disease. There are two such variants with odds ratios > 1.25, rs141672851 (European allele frequency= 0.021) and rs150558343 (European allele frequency= 0.013), near the Linker For Activation Of T Cells (LAT) and F-Box And Leucine Rich Repeat Protein 20 (FBXL20) genes, respectively. Neither of these variants show association to other currently measured and examined phenotypes (`https://genetics.opentargets.org`). The sample size required for the rarest variant with an FDR association < 0.01 with T1D, rs150558343, to have 80% power to detect association to genome-wide significance ($< 5 \times 10^{-8}$), assuming the same effect size observed in this analysis, would be $39,681$ cases and controls. Therefore in order to robustly identify rare variants associated with T1D, increases in sample sizes are required.

When examining associations to a Bonferroni-corrected significance threshold of 1.24×10^{-7}, 81 regions were associated with disease. However, the variance explained by each additional associated variant being included in the model decreased with decreasing association strength, as expected. Figure 4.7 shows the variance explained by including each variant of association in the model, measured by McFadden r^2, defined as $1 - \frac{ln(L_M)}{ln(L_0)}$, with $ln(L_M)$ and $ln(L_0)$ the natural logarithm of the likelihood under the model examined and the null model including only as intercept, respectively. Cumulatively, less than 20% of

the trait variance is explained by these non-HLA variants.

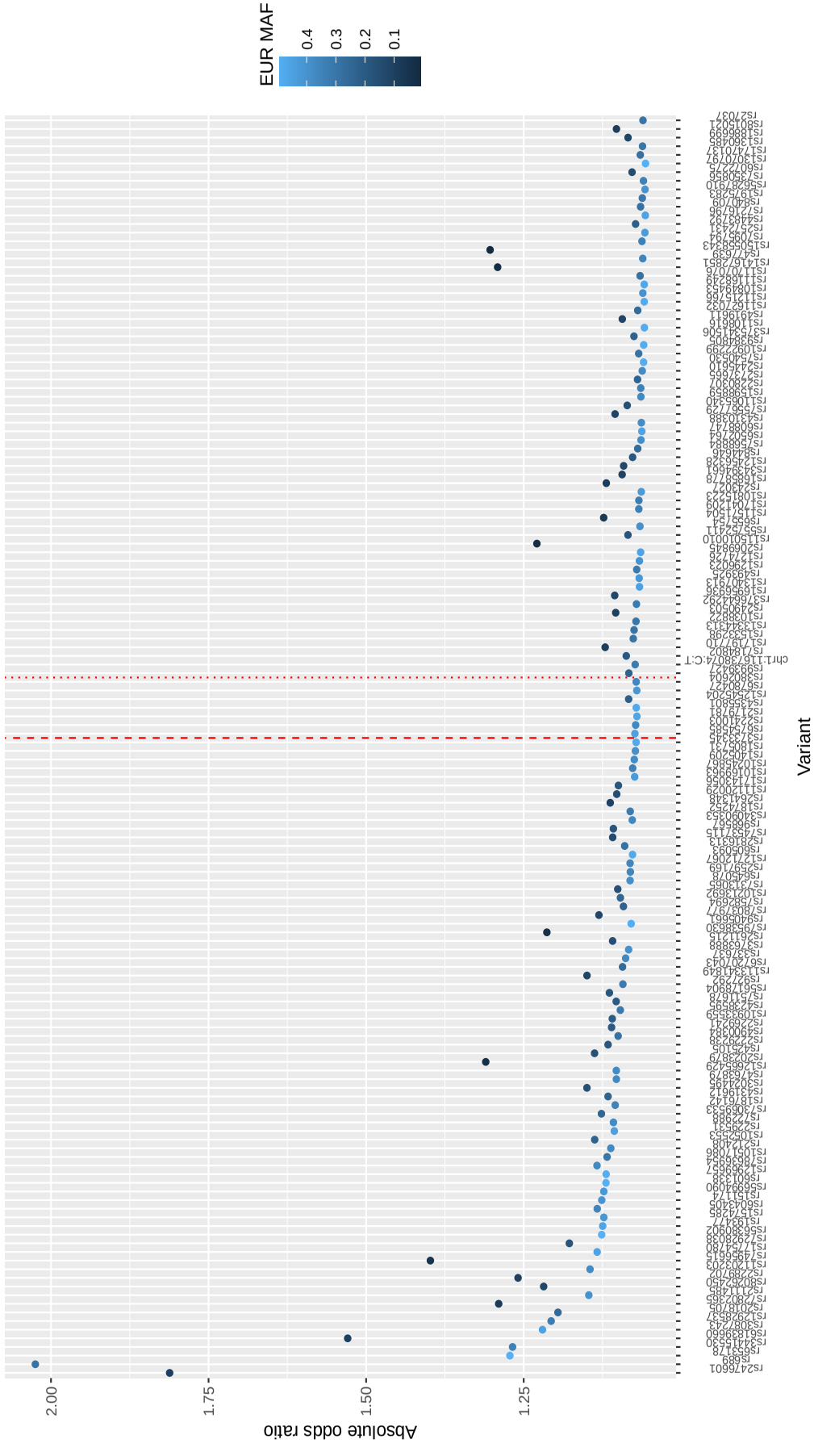


Figure 4.6: Each variant associated to a False Discovery Rate < 0.01, ordered by decreasing strength of association with T1D and coloured by the minor allele frequency of the variant in the European population. Variants to left of red dashed line are genome-wide significant, variants to the left of the red dotted line are Bonferroni-corrected significant at 5%.

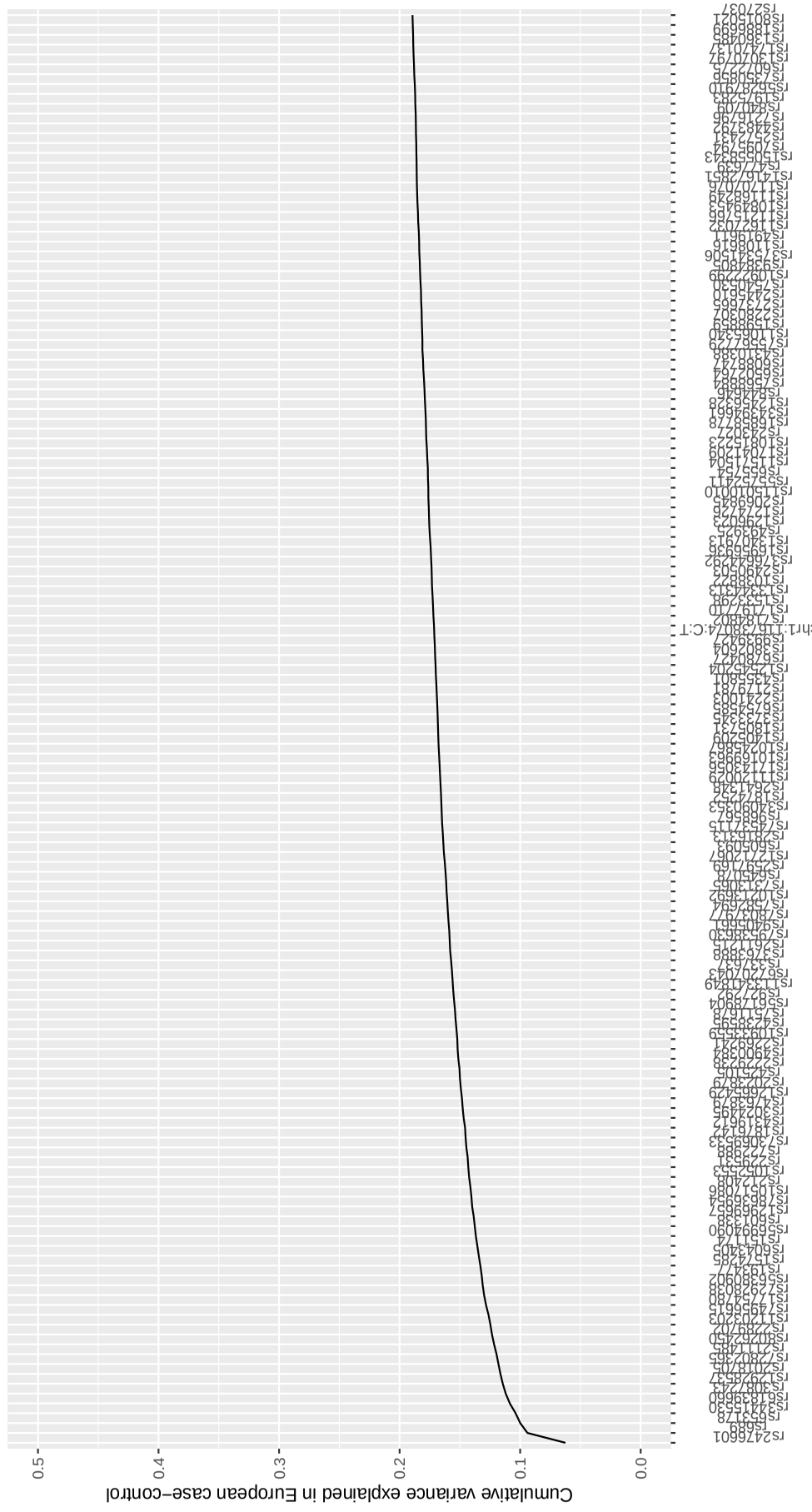


Figure 4.7: Each variant associated to a False Discovery Rate < 0.01, ordered by decreasing strength of association with T1D with the McFadden r^2 displayed from a model including each variant to the left of the current dot in a logistic regression model.

Stronger association in individuals diagnosed at < 7 compared to $\geq$13?

In a similar analysis to in chapter 3, I examined each variant that passed variant QC in the analysis in chapter 3, using the same $27,071$ individuals used in that analysis. I fitted a multinomial logistic regression examining the log-odds ratio of the variant for those diagnosed at < 7, $7-13$ and ≥ 13. I then counted the number of times the ordering had the largest effect size in the < 7 group, the intermediate effect in the $7-13$ group and the smallest effect in the ≥ 13 group, and compared this to the number of times expected by chance using a binomial test. There were 17 of the 22 novel variants associated with T1D that passed variant QC in the chapter 3 analysis. Of these 17 variants, eight had the log odds ratio with the largest effect in the < 7 group, intermediate in the $7-13$ group and largest in the ≥ 13 group, compared to three expected by chance ($p = 0.003$), indicating that these novel associations are also preferentially associated in those individuals with early-diagnosed T1D.

Genomic inflation factor assessment

The observed inflation factor in the Phase I and II meta-analysis was 1.51. This is higher than the expected value of 1 for an unbiased genotyping platform and no artificial inflation of test statistics. The inflation factor from the European family-based analysis was 1.42, and in the five European case-control analyses performed with equivalent statistical power to the family-based analyses, the inflation factor statistics observed were 1.29, 1.32, 1.30, 1.30 and 1.26 (Figure 4.8), indicating that the inflation factor of 1.51 is as a result of genuine association as opposed to spurious associations, since the family-based inflation factor, which is robust to population structure, is higher than the inflation factor of case-control analyses of similar statistical power.

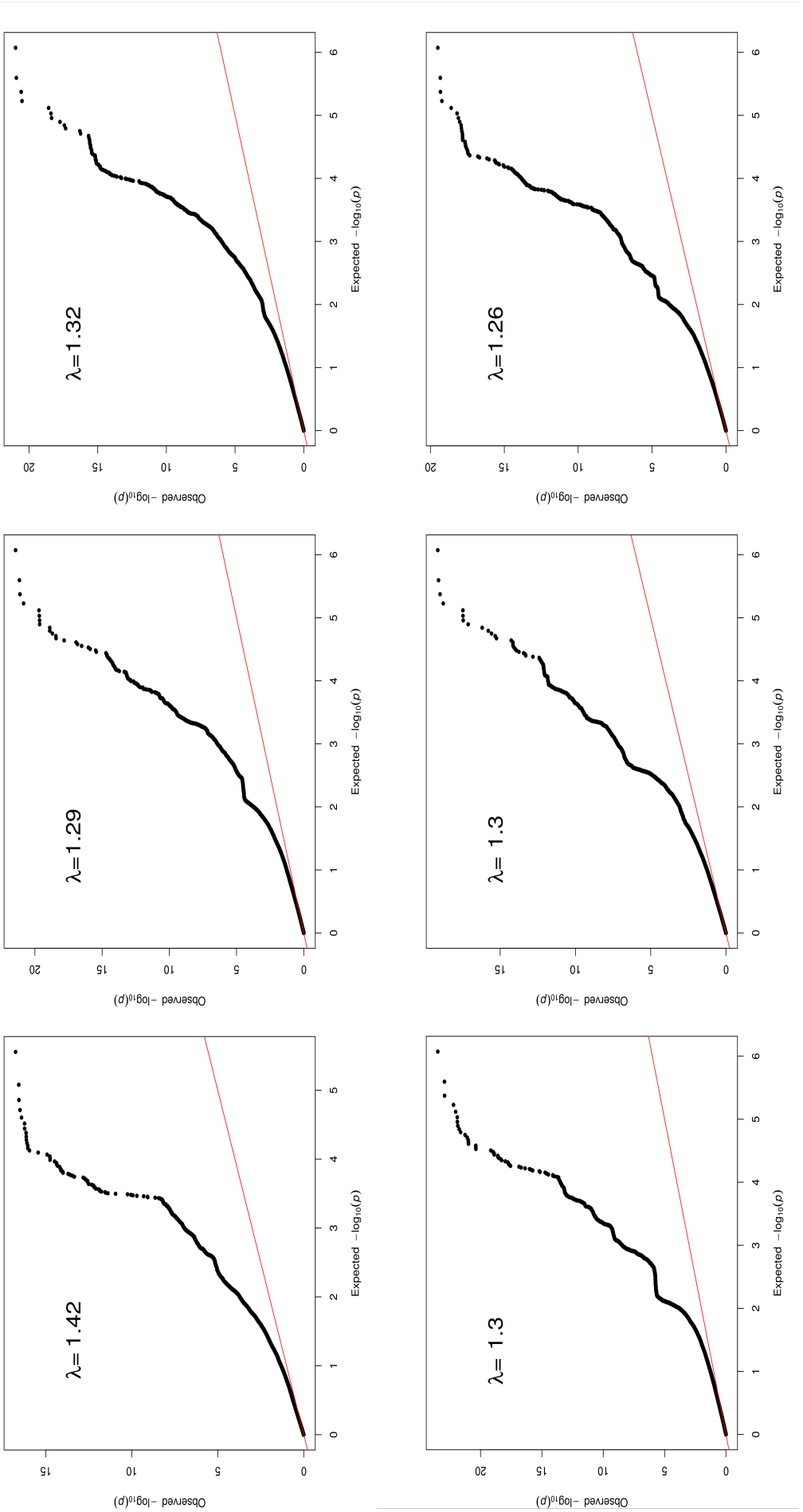


Figure 4.8: Quantile-quantile plots for European family-based collections (top left), and the five randomly sampled case-control populations (top centre, right, bottom left, centre, right) with equivalent statistical power to the family-based analyses.

The random sampling of cases and controls to generate the five European case-control datasets was carried out by Catharine Robertson.

4.3.4 Fine mapping - European only using GUESSFM

Global overview

GUESSFM analyses using only individuals clustering into the European ancestry group was performed on 52 regions associated with disease in the Phase I and II meta-analysis. Results were broadly comparable to the previous iteration of fine mapping T1D associated regions using GUESSFM, which is expected given many of the samples used in the present analysis were also used in that previous iteration [31]. Of the 59 index variants from groups with posterior probability > 0.8 in the previous analysis, just five were not prioritised by GUESSFM in the current analysis, four of which are located near the INS gene. At this locus, it is likely that the previous iteration of fine-mapping was incorrect, since rs689 (T>A) was not prioritised, which tags a variable tandem repeat that is very likely to be causal [194]. The complex LD in the region means removing the rs689 variant and those in high LD with it from the analysis leads to many association signals (data not shown), implying this is likely what happened in the previous fine-mapping analysis [31]. For the remaining group not prioritised by GUESSFM in the current analysis but prioritised in the previous analysis, near the RAS Guanyl Releasing Protein 1 (RASGRP1) gene on chromosome 15q14, the index variant in the region was different to the previous analysis (rs1874252 (G>T) in this analysis, rs55728265 (C>T) in the previous, $r^2 = 0.02$) and therefore it is unsurprising that GUESSFM prioritises different signals at this locus.

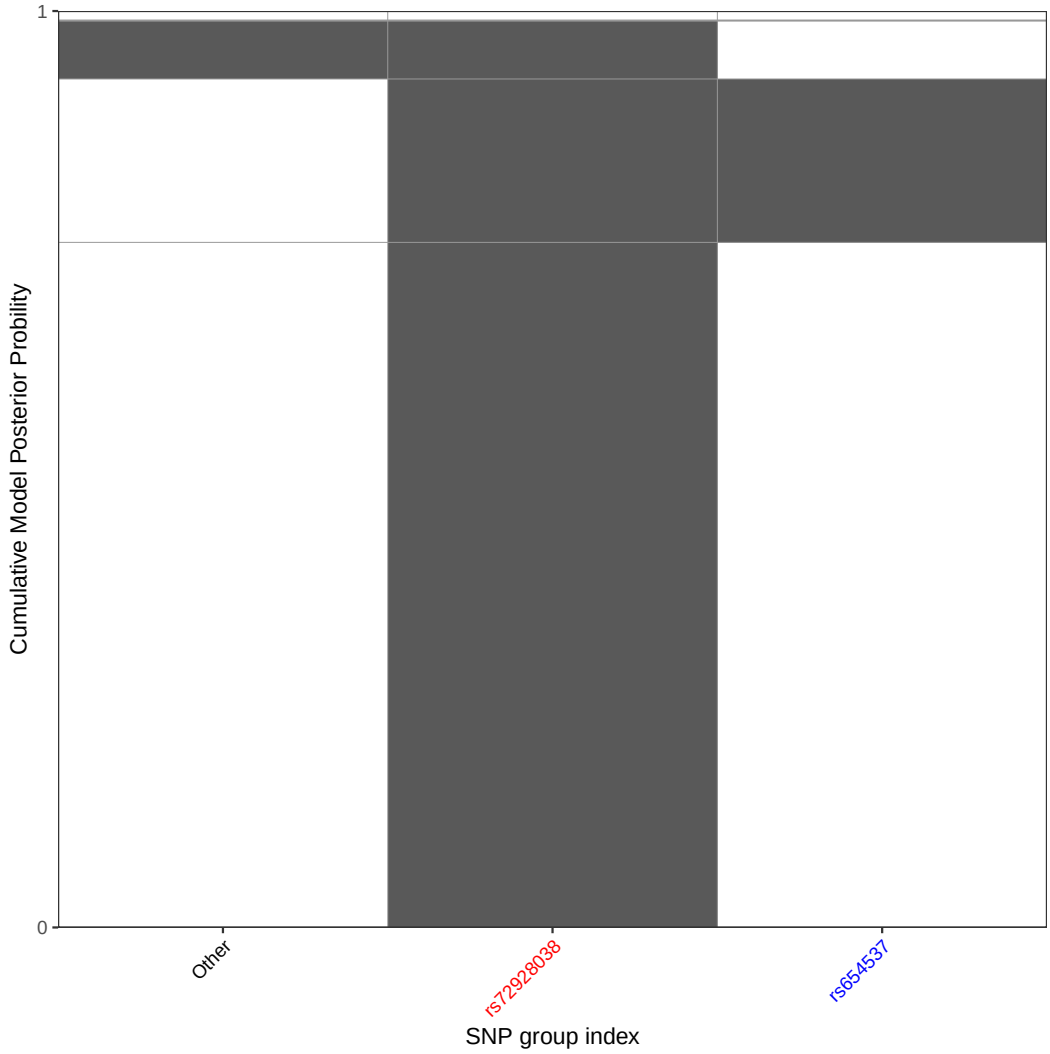
Example region 1 - *BACH2*

An example of a region where the fine mapping resolution has increased due to increase in sample size is near the BTB Domain And CNC Homolog 2 (BACH2) gene on chromosome 6q15. In the previous fine mapping analysis, this region had four groups prioritised by GUESSFM, but two of them, one indexed by rs72928038 (G>A, group

posterior probability= 0.98) and the other indexed by rs56258221 (T>C, group posterior probability= 0.02), are in LD (r^2 = 0.4) and their combined posterior sums to 1, indicating that they are likely identifying the same signal. Therefore when considering functional follow up studies, researchers would be required to consider the two variants in the group indexed by rs72928038 and the seven variants indexed by rs56258221.

In the current fine mapping analysis, the seven variants indexed by rs56258221 are not prioritised and therefore researchers can consider just two variants to test in functional follow-up studies. There is some evidence for a secondary signal in the region but the majority of the posterior probability (> 0.7) points towards a single causal variant in the region. Figure 4.9 shows the cumulative probability plot for models visited during the stochastic search. It shows that in almost every model searched, a variant from the group indexed by rs72928038 is included and in a subset of models, a variant from a group indexed by rs654537 (A>G) is included as well as a variable in the group indexed by rs72928038. Since this group consists of just two variants, identification of the causal variant in functional follow up studies should be simpler than when having to consider nine variants. However, it is of course possible that a nearby untyped variant, which failed imputation QC and is in very strong LD with the analysed variants, is the causal variant.

Figure 4.9: Cumulative probability plot showing the models visited during the stochastic search in the *BACH2* region on chromosome 6q15. Shaded black area is the proportion of times the model includes a variant from the group that the column indexes in the stochastic search. Shaded black in more than one column indicates the proportion of times the stochastic search visits a model with one variant from each group.

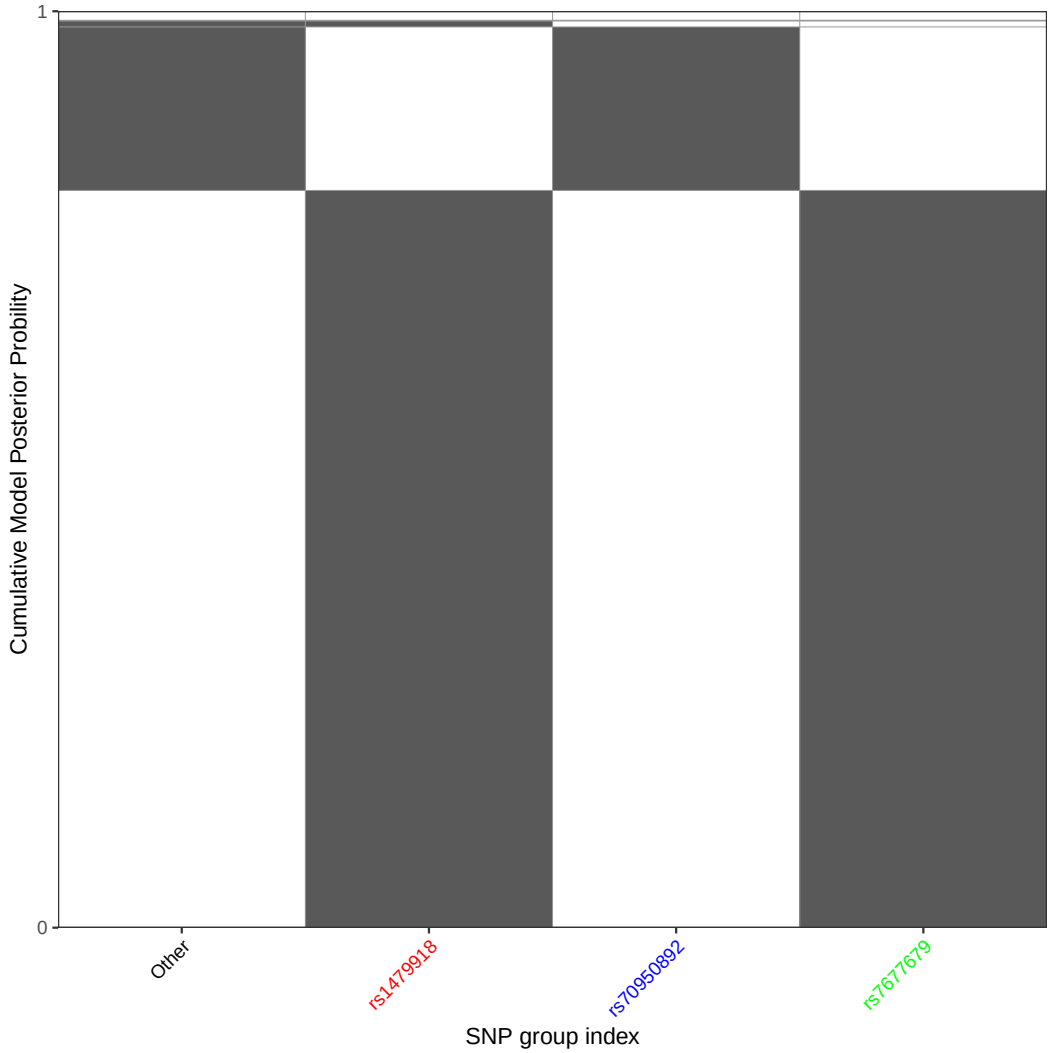


Example region 2 - *IL2/IL21*

A clear example of a region where a variant tags two disease-associated haplotypes can be seen on chromosome 4q27, near the interleukin-2 (IL-2) and interleukin-21 (IL-21) genes. The index variant in the univariable analysis is rs70950892 (ATC>A), yet this variant has a relatively low posterior probability in the GUESSFM analysis (posterior probability of the group containing rs70950892= 0.179). The GUESSFM analysis suggests there are two causal variants in the region (posterior probability for two causal variants in the region > 0.9) and the two variants that index the groups with highest posterior proba-

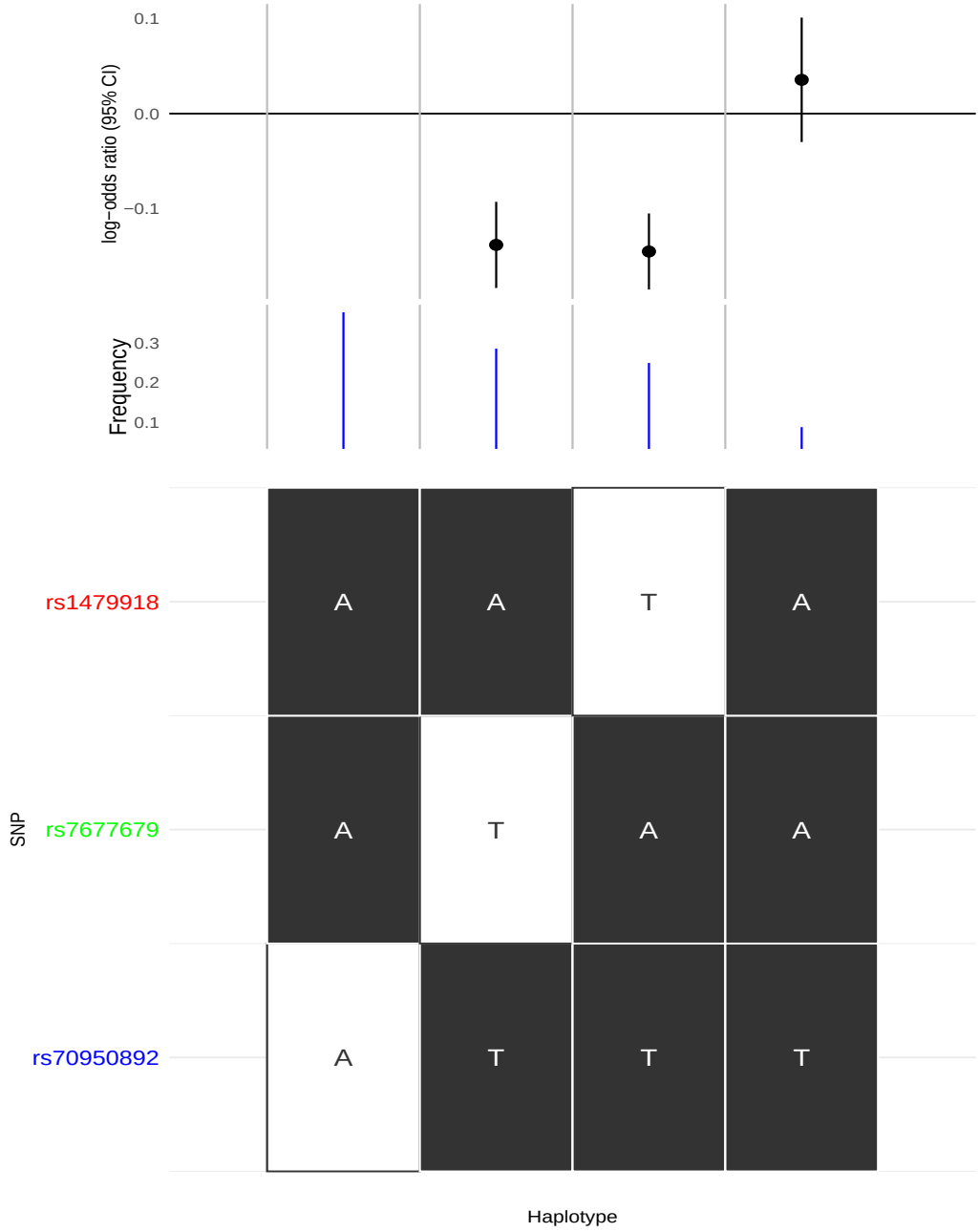
bility, rs1479918 (A>T, group posterior probability= 0.82) and rs7677679 (A>T, group posterior probability= 0.81) are both in LD with the index variant from the univariable analysis (r^2 = 0.21 and 0.23 for rs1479918 and rs7677679, respectively). The stochastic search visits the model including only the univariable index variants approximately 20% of the time (Figure 4.10) yet the other 80% of the time prioritises the model including a variant from both of the groups indexed by rs1479918 and rs7677679. This indicates that the phenotype can be explained by the tag variant, rs70950892, and in the simplest univariable analysis, this variant shows the strongest association, but when considered in combination, both rs1479918 and rs7677679 together better explain the phenotype.

Figure 4.10: Cumulative probability plot showing the models visited during the stochastic search in the *IL2/IL21* region on chromosome 4q27. Shaded black area is the proportion of times the model includes a variant from the group that the column indexes in the stochastic search. Shaded black in more than one column indicates the proportion of times the stochastic search visits a model with one variant from each group.



This result is clearly evident in the haplotype analysis of the region. In the univariable analysis, the insertion at the tag variant rs70950892 is associated with protection from T1D. However, in the haplotype analysis in Figure 4.11 the ATC allele (represented simply as "T" in the Figure 4.11) is not associated with disease risk when it appears on its own without the minor allele from rs1479918 or rs7677679, both of which have a protective effect on disease risk.

Figure 4.11: Haplotype analysis in the *IL2/IL21* region, using three variants prioritised by GUESSFM, which index three groups of variants potentially associated with T1D. Variant rs70950892 shows no association with disease risk when appearing without the minor alleles at either rs1479918 or rs7677679.



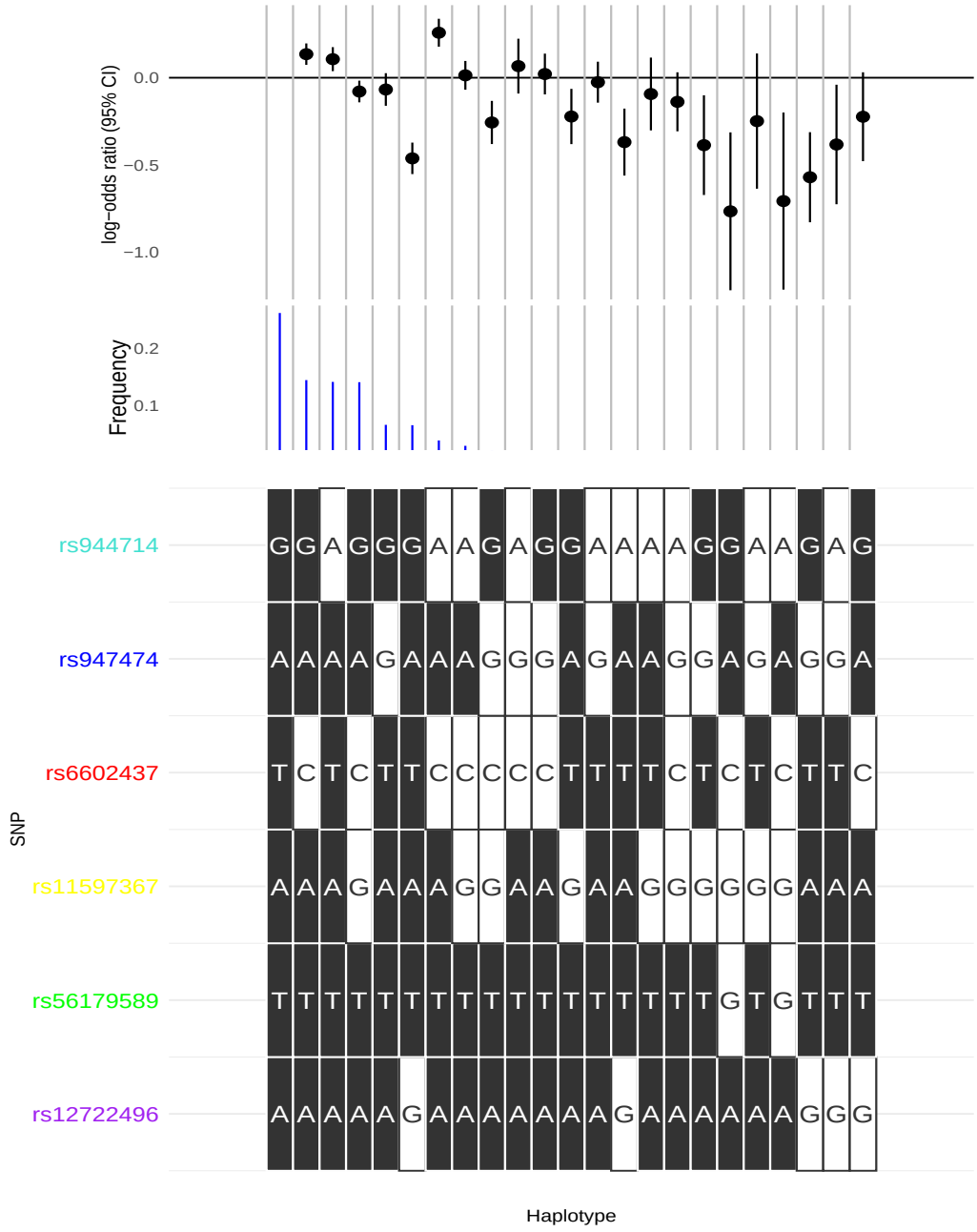
Example region 3 - *IL2RA*

The region on chromosome 10p15.1 containing the IL-2Ra and RNA Binding Motif Protein 17 (RBM17) genes has been examined in depth and four disease-associated haplotypes have been identified, referred to as groups A (indexed by rs12722496 A>G), C (indexed by rs11597367 A>G), E (indexed by rs6602437 T>C) and F (indexed by rs56179589 T>G), where the index variants named here are corresponding to the present analysis for consistency within this thesis rather than between this thesis and the original publication [202], but each group identified in the original publication is prioritised in this analysis (at least one variant from groups A, C, E and F are prioritised here with group posterior probability> 0.9).

However, in addition to the four previously identified disease associations in the region, the increase in statistical power of the present analysis allowed identification of two further signals with group posterior probability > 0.8, indexed by rs947474 (A>G) and rs944714 (G>A). Figure 4.12 shows a haplotype analysis including a variant from each of the six groups with posterior probability > 0.8. It is clear that rs12722496 (group A) and rs56179589 (group F) are the rarest variants and have the largest effect size, whilst rs11597367 (group C), rs6602437 (group E) and the two additional signals, rs947474 and rs944714, are more common and with more subtle effects.

This is the most complex region with the largest number of associations with T1D and it is likely that with further increased sample size, additional signals could be detected.

Figure 4.12: Haplotype analysis in the *IL2RA/RBM17* region, using variants prioritised by GUESSFM, which index six groups with posterior probability > 0.8.



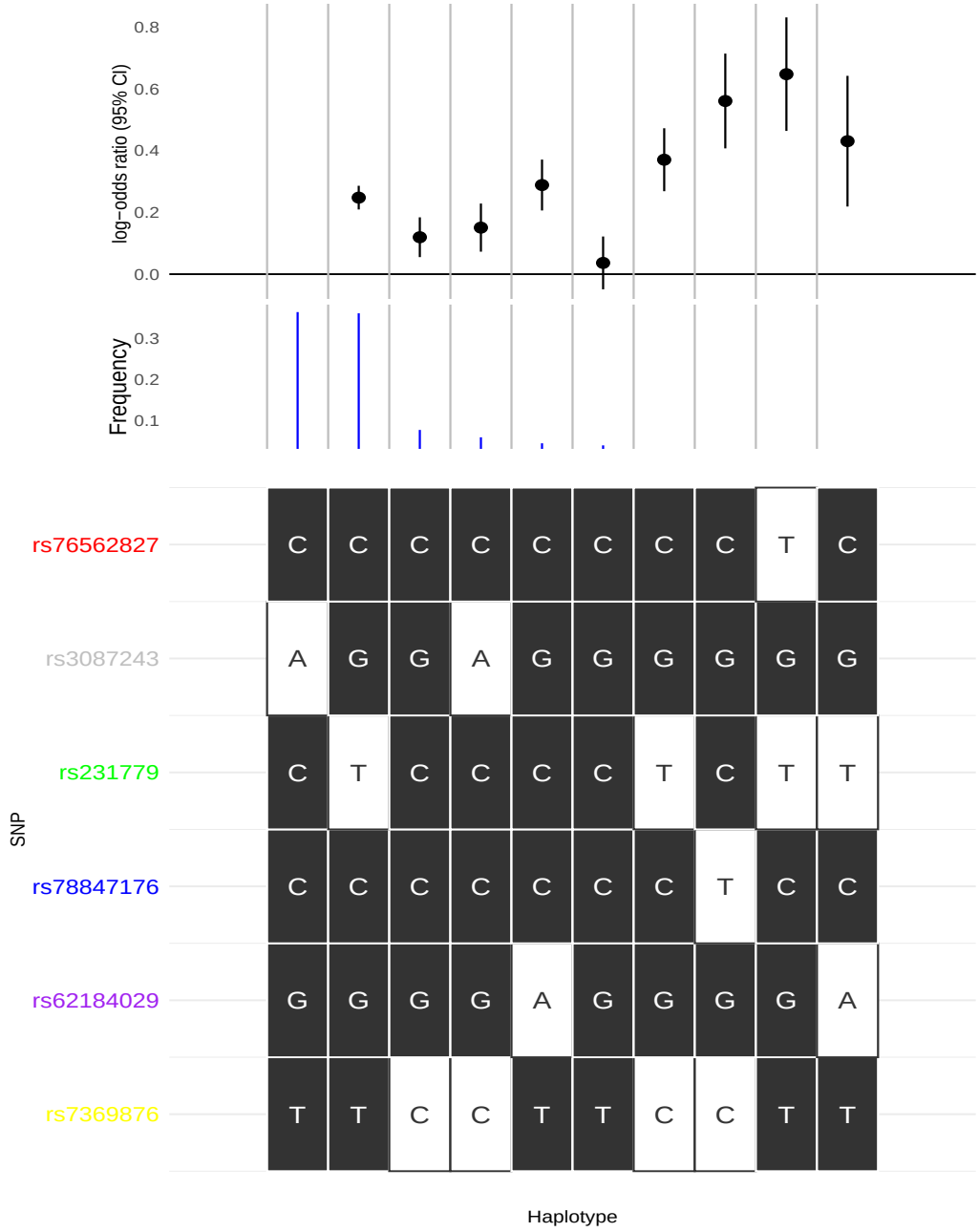
Example region 4 - *CTLA4*

Another complex region associated with T1D is a region on chromosome 2q33.2, containing the CTLA4 and Cluster of Differentiation 28 (CD28) genes. The index variant from the univariate analysis was rs3087243 (G>A), which is also known as CT60. This variant has consistently been the most strongly associated variant in the region since the first T1D GWAS [48, 15, 30, 144, 190, 207] and is located in the 3' UTR of *CTLA4*;

it has been studied in depth for association with many autoimmune diseases [138, 62, 191, 96]. However, this variant is not prioritised by GUESSFM, instead five independent signals are prioritised, each with some LD with rs3087243, as can be visualised in Figure 4.13. Of these prioritised variants, three are quite rare (MAF= 0.011, 0.013 and 0.053 for rs76562827 (C>T), rs78847176 (C>T) and rs62184029 (G>A), respectively) and therefore there may have not been statistical power to detect these variants previously. The two groups containing more common variants, rs231779 (C>T) and rs7369876 (T>C) were prioritised as a pair over rs3087243 in a recent fine-mapping analysis, which incorporated multiple immune-related diseases [8], providing reassurance the results observed here are genuine.

However, one of the rare groups prioritised by GUESSFM, indexed by rs78847176, contained no variants that were genotyped on the ImmunoChip and therefore I will caution against considering this association as robust until there has been a replication where this variant is directly genotyped, rather than imputed.

Figure 4.13: Haplotype analysis in the *CTLA4/CD28* region, using variants prioritised by GUESSFM, which index five groups of variants with posterior probability > 0.8, as well as the index variant from the univariable analysis, rs3087243.



PAINTOR trans-ethnic fine mapping

Of the 52 regions examined using European individuals, I considered only those with highest posterior probability of one causal variant in the region to take forward to PAINTOR fine mapping. This is because I wanted some clear and interpretable examples and think the trans-ethnic aspect is useful to reduce causal variant group size when there is a single causal variant in the region. When examining regions with more than one association

with disease, PAINTOR relies on population LD structure to identify the most likely combination of variables associated with disease; using these approximations along with association summary statistics will not be as accurate as using the individual level data, as GUESSFM does. Therefore I considered 30 regions with posterior probability > 0.5 of one causal variant in the region. Of these 30 regions, three ($p > 5 \times 10^{-4}$), showed association in European and African ancestry groups and five regions showed association in European and Finnish ancestry groups, none showed association in all three ancestry groups. None of the five regions associated in European and Finnish ancestry groups had any gain in fine-mapping resolution as a result of incorporating two ancestry groups, either due to the already small groups of prioritised variants using GUESSFM or due to relatively weak disease associations in the European ancestry group, leading to uncertain estimates in both GUESSFM and PAINTOR. However, the three regions showing association in European and African populations showed benefit of trans-ethnic fine mapping, by producing a smaller number of variants with posterior probability > 0.01 after using the summary statistics from European and African populations.

The first of these regions is on chromosome 4p15.2, near the Recombination Signal Binding Protein For Immunoglobulin Kappa J Region (RBPJ) gene, where there are 24 variants in the group prioritised by GUESSFM but just 11 variants with posterior probability > 0.01 in the PAINTOR analysis. Figure 4.14 shows the Manhattan plots from the European ancestry group, African ancestry group and the PAINTOR posterior probability of association for each variant, highlighting the smaller number of variants in the African ancestry group that show strong association relative to the European ancestry group. Figure 4.15 shows the location of the GUESSFM prioritised variants and the ATAC-seq read count from various cell types. The three variants highlighted in blue show the three variants prioritised by PAINTOR with posterior probability > 0.1. None of these variants intersect a large ATAC-seq peak that is present in almost all cell types (not all cell types shown), but one of the variants does overlap an ATAC-seq peak that is present in a subset of cell types, rs35944082 (A>G). The GUESSFM, PAINTOR and ATAC-seq datasets combined would therefore point towards this variant being the most likely causal variant

in the region if the assumption is made that the causal variant will be in open chromatin in any of the cell types shown in Figure 4.15. Without the use of PAINTOR, however, the variant overlapping the ATAC-seq peak present in all cell types except immature NK cells in Figure 4.15 might be prioritised, which highlights the potential advantage of incorporating many analysis techniques and/or datasets when attempting to identify the most likely causal variants.

Figure 4.14: Manhattan plots from the European ancestry group analysis and the African ancestry group analysis in the *RBPJ* region, as well as the PAINTOR posterior probability of each variant, coloured by the r^2 to the lead variant from the PAINTOR analysis.

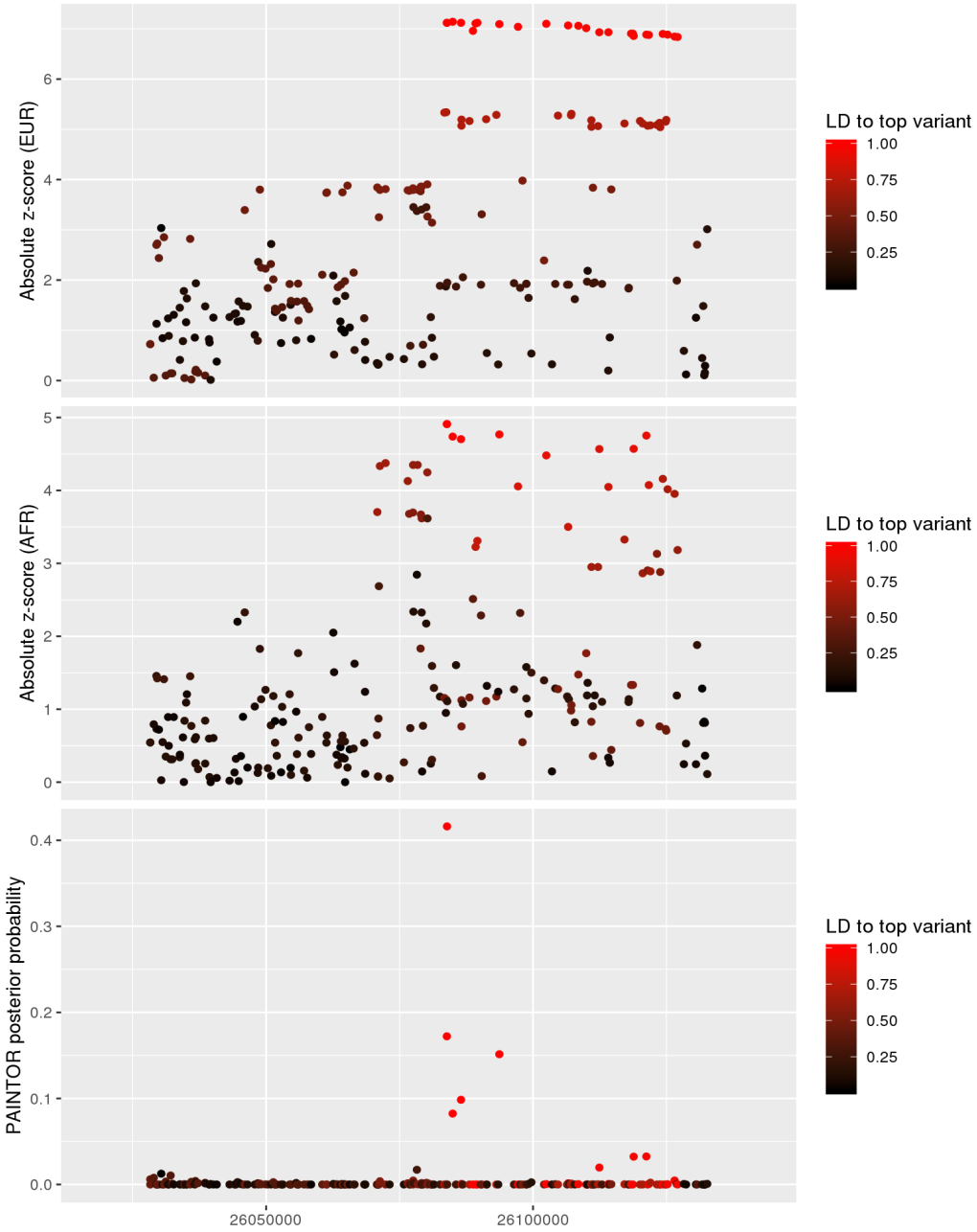
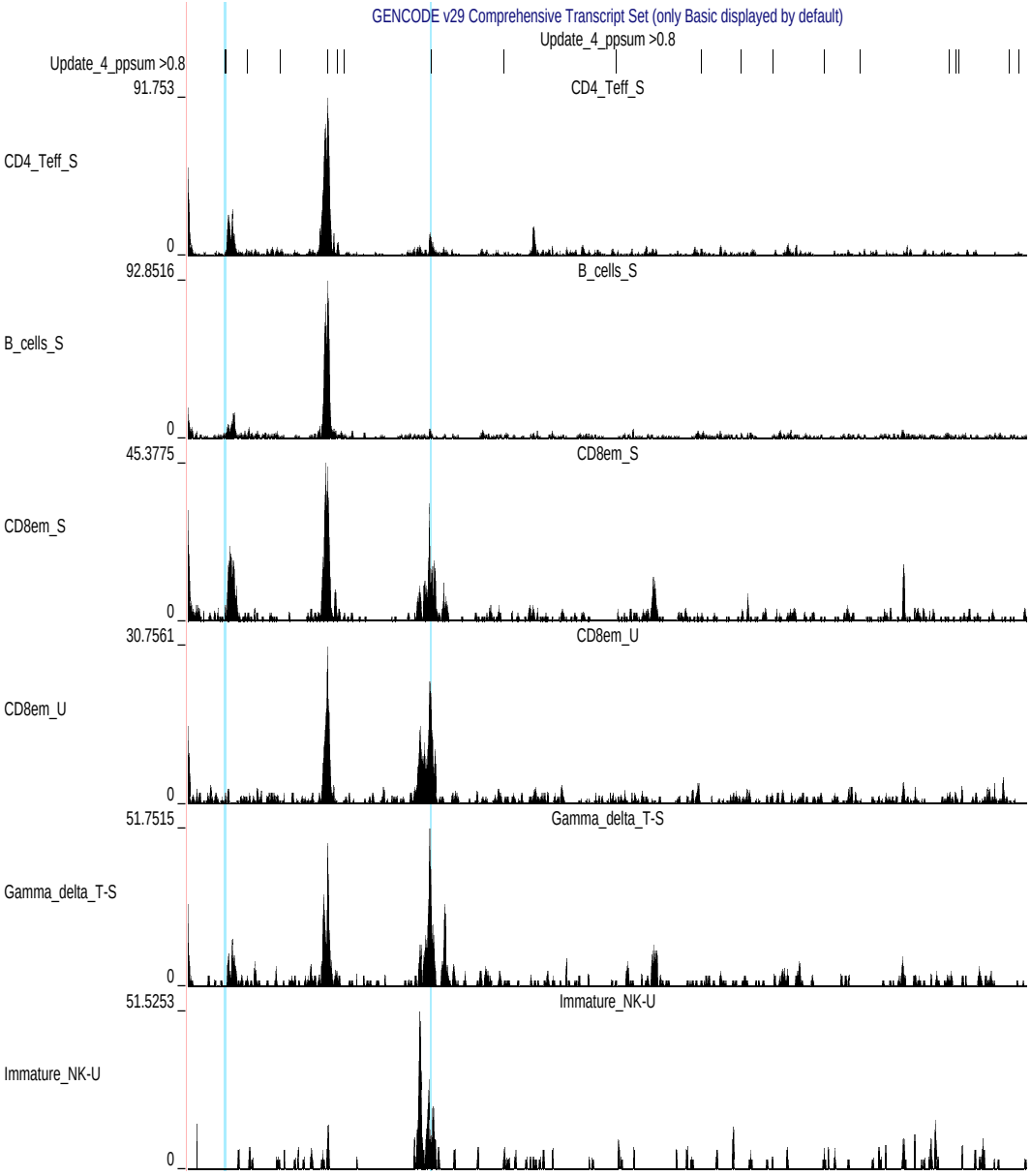


Figure 4.15: GUESSFM prioritised variants in the *RBPJ* region, with variants prioritised by PAINTOR (posterior probability > 0.1) highlighted in blue. ATAC-seq read count shown for stimulated effector $CD4^+$ T cells, stimulated B cells, stimulated effector memory $CD8^+$ T cells, unstimulated effector memory $CD8^+$ T cells, Gamma-delta T cells and unstimulated immature NK cells. Cell types shown are a subset of cell types that show an ATAC-seq peak overlapping variant rs35944082, except B cells, which is shown to highlight not all immune cell types and states have a peak in this location.



A similar association pattern can be observed in Figure 4.16 showing a regions on chromosome 6q22.32, near the Centromere Protein W (CENPW) gene, where the number of variants prioritised in GUESSFM was 86, compared to 47 with posterior probability > 0.01 using PAINTOR. In addition, the three variants prioritised by PAINTOR with posterior probability > 0.1 are rs67126256 (TG>T, probability= 0.164), rs2045258

(A>G, probability=0.126) and rs9388486 (T>C, probability= 0.112). Figure 4.17 shows that of the 86 variants prioritised by GUESSFM in the region, there is one variant located in a clear ATAC-seq peak in multiple cell types, and this is rs9388486, which is prioritised by PAINTOR. In addition, rs9388486 has a low regulomeDB score (`http://www.regulomedb.org/`), indicating that the variant is likely to affect transcription factor binding, including affecting binding of transcription factors Transcription Factor 7 Like 2 (*TCF7L2*) and SIN3 Transcription Regulator Family Member A (*SIN3A*) in a human pancreatic cancer cell-line, PANC-1. These are three orthogonal pieces of evidence suggesting that rs9388486 might be causal due to association in more than one ancestry group and it's location in an open chromatin region in many immune cell types, increasing its candidacy as the causal variant.

Figure 4.16: Manhattan plots from the European ancestry group analysis and the African ancestry group analysis in the *CENPW* region, as well as the PAINTOR posterior probability of each variant, coloured by the r^2 to the lead variant from the PAINTOR analysis.

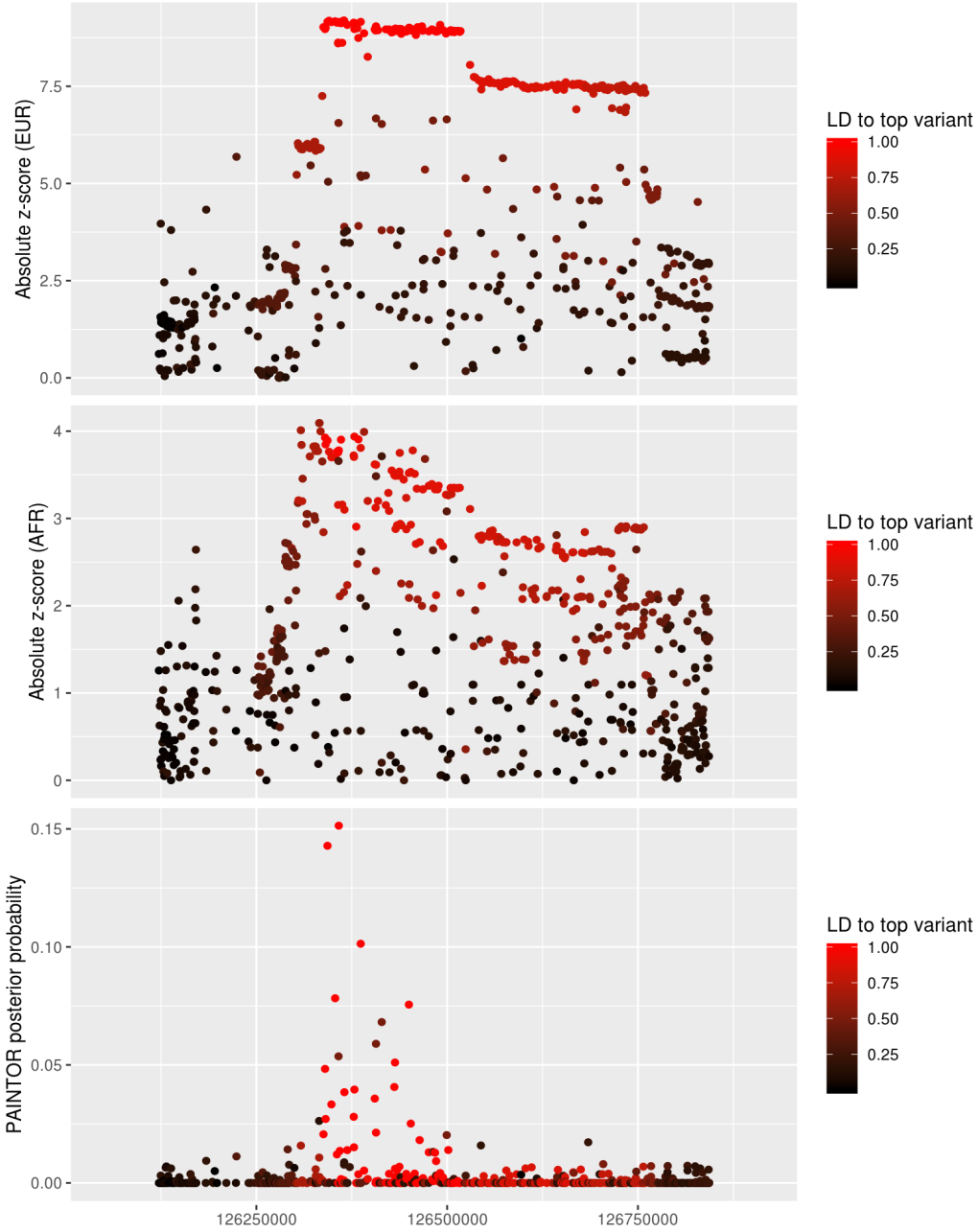
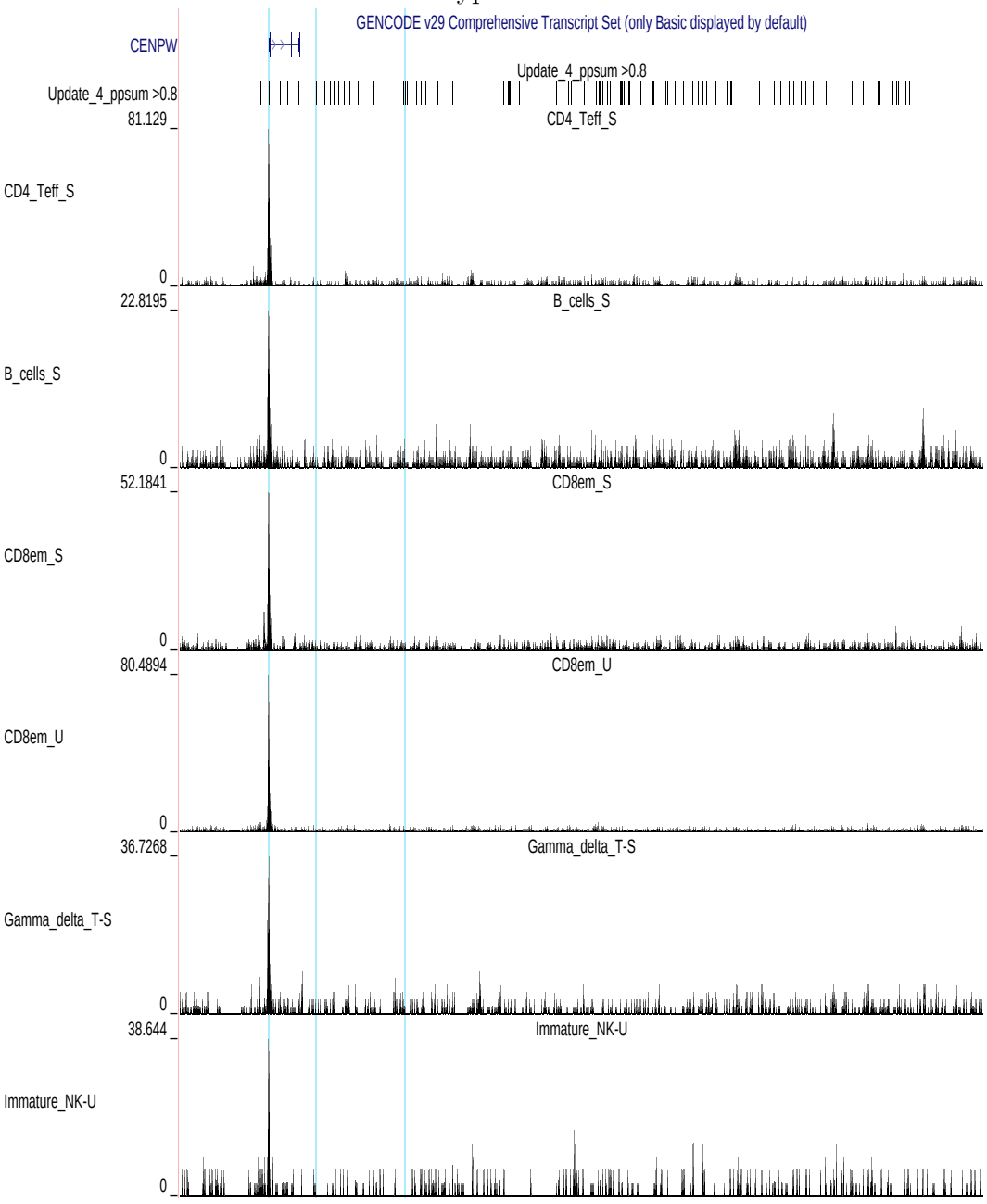


Figure 4.17: GUESSFM prioritised variants in the *CENPW* region, with variants prioritised by PAINTOR (posterior probability > 0.1) highlighted in blue. ATAC-seq read count shown for stimulated effector $CD4^+$ T cells, stimulated B cells, stimulated effector memory $CD8^+$ T cells, unstimulated effector memory $CD8^+$ T cells, Gamma-delta T cells and unstimulated immature NK cells. Variant rs9388486 overlaps ATAC-seq peaks in multiple cell types, increasing candidacy as the disease causal variant. Cell types shown are a subset of immune cell types with data available.



Finally, a region on chromosome 18q22.2, near the Cluster of Differentiation 226 (CD226) gene, had 38 variants prioritised by GUESSFM compared with 21 after using PAINTOR (Figure 4.18). However, the three variants with posterior probability > 0.1 from PAINTOR in this region did not have such obvious candidacy for causality. None of the variants

prioritised by PAINTOR (probability > 0.1) highlighted in blue in Figure 4.19 lie in ATAC-seq peaks in any immune cell type or islets (not all cell types shown). Despite integration of multiple ancestry groups and functional data, there is no variant in the region that could yet be considered most likely the causal variant.

Figure 4.18: Manhattan plots from the European ancestry group analysis and the African ancestry group analysis in the *CD226* region, as well as the PAINTOR posterior probability of each variant, coloured by the r^2 to the lead variant from the PAINTOR analysis.

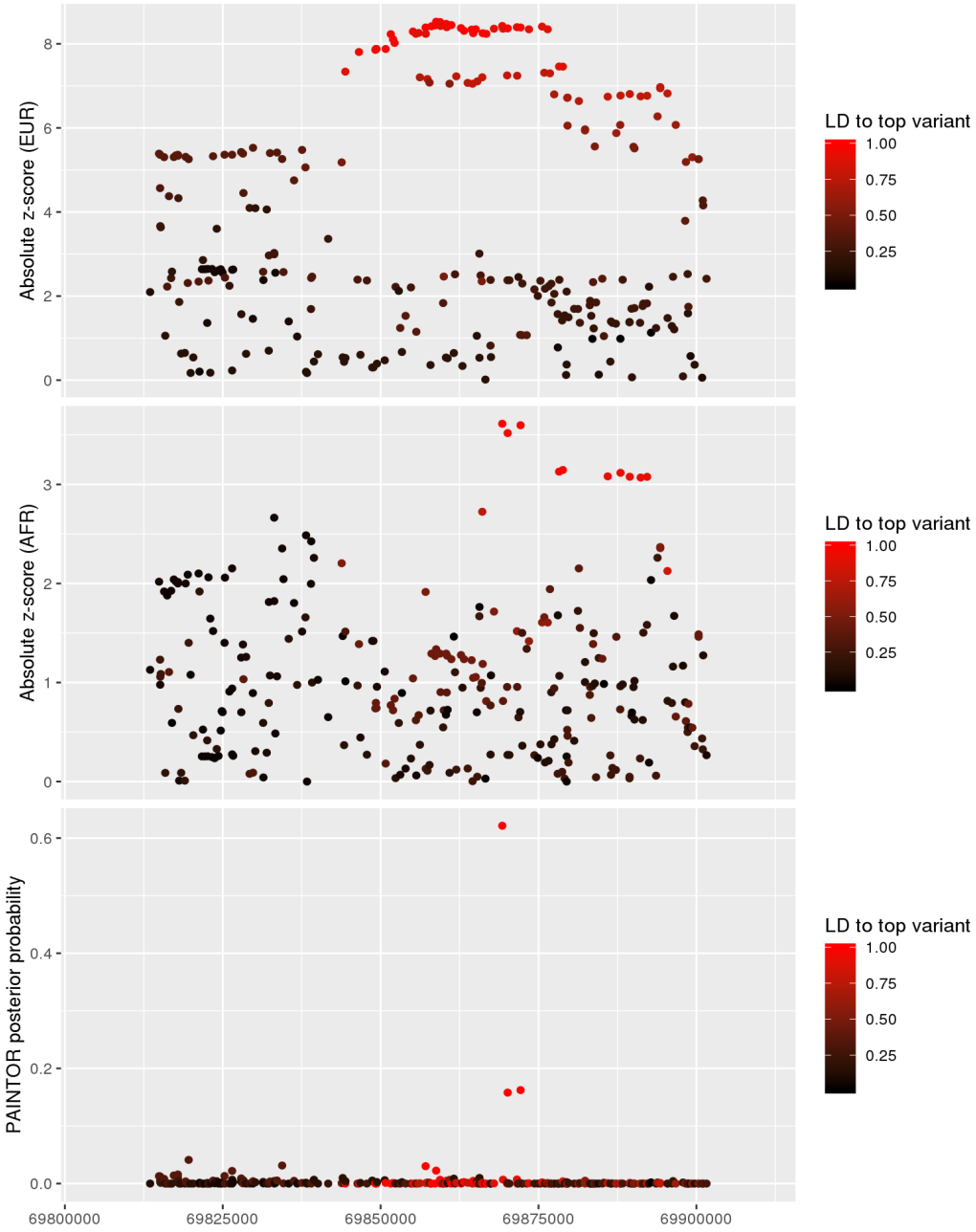
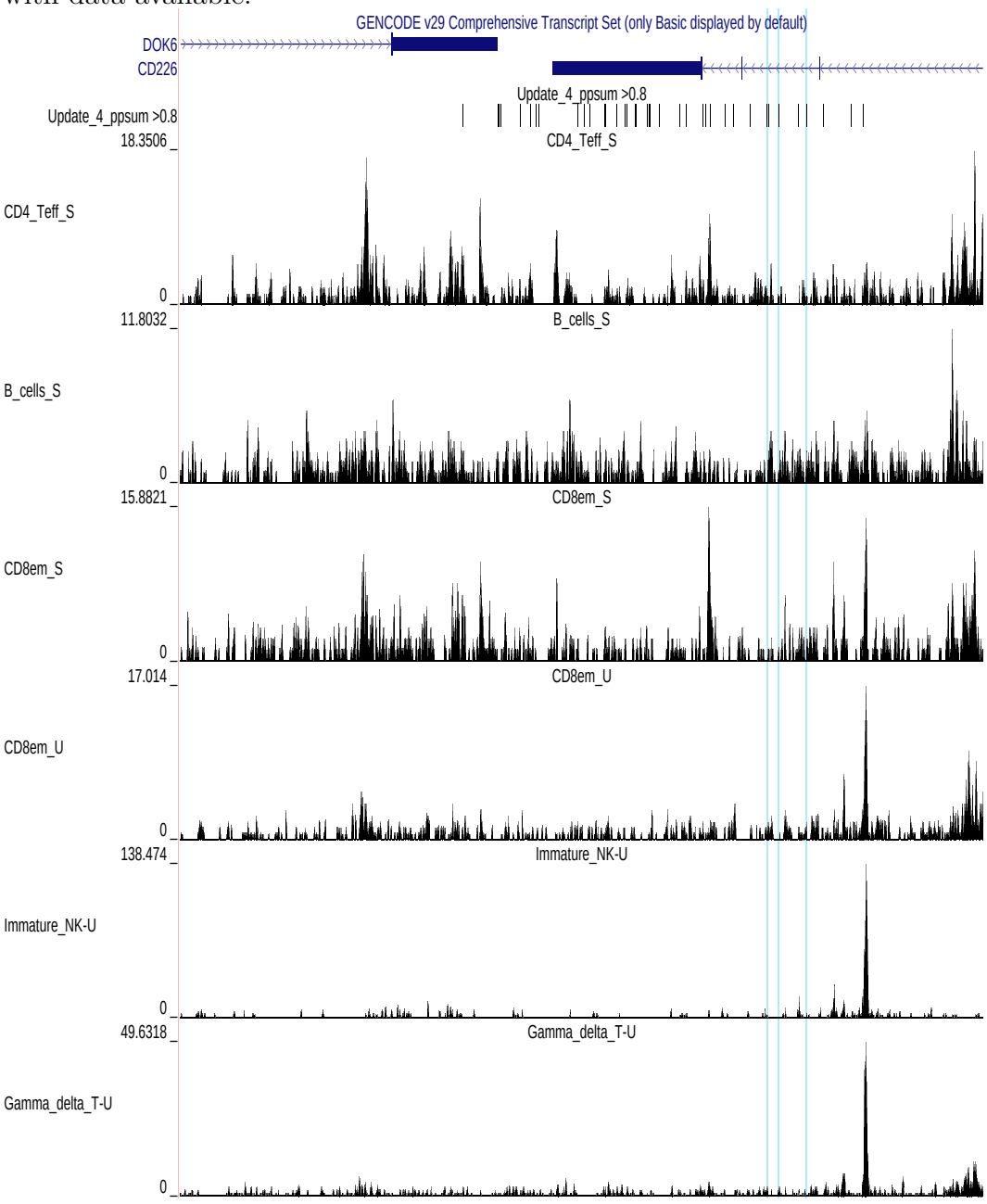


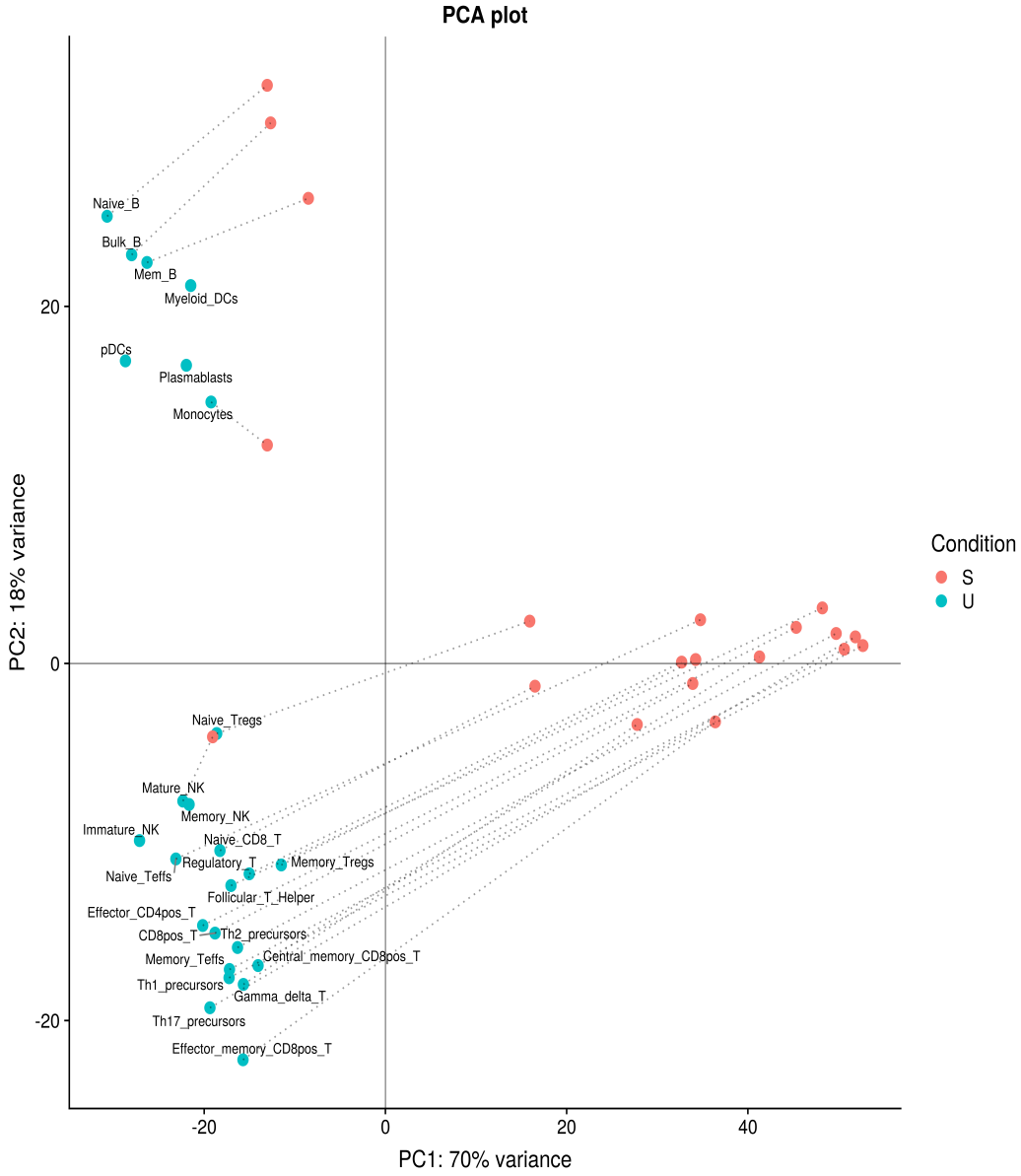
Figure 4.19: GUESSFM prioritised variants in the *CD226* region, with variants prioritised by PAINTOR (posterior probability > 0.1) highlighted in blue. ATAC-seq read count shown for stimulated effector $CD4^+$ T cells, stimulated B cells, stimulated effector memory $CD8^+$ T cells, unstimulated effector memory $CD8^+$ T cells, Gamma-delta T cells and unstimulated immature NK cells. Cell types shown are a subset of immune cell types with data available.



4.3.5 T1D variant enrichment ATAC-seq enrichment in ATAC-seq peaks

In order to confirm the ATAC-seq data were sufficiently different from one another, PCA was run using the counts from the consensus list of peaks, using the Stanford data only. Both the stimulation condition and the cell types separated in the first two PCs (Figure 4.20), indicating the data were suitable to use in enrichment analyses.

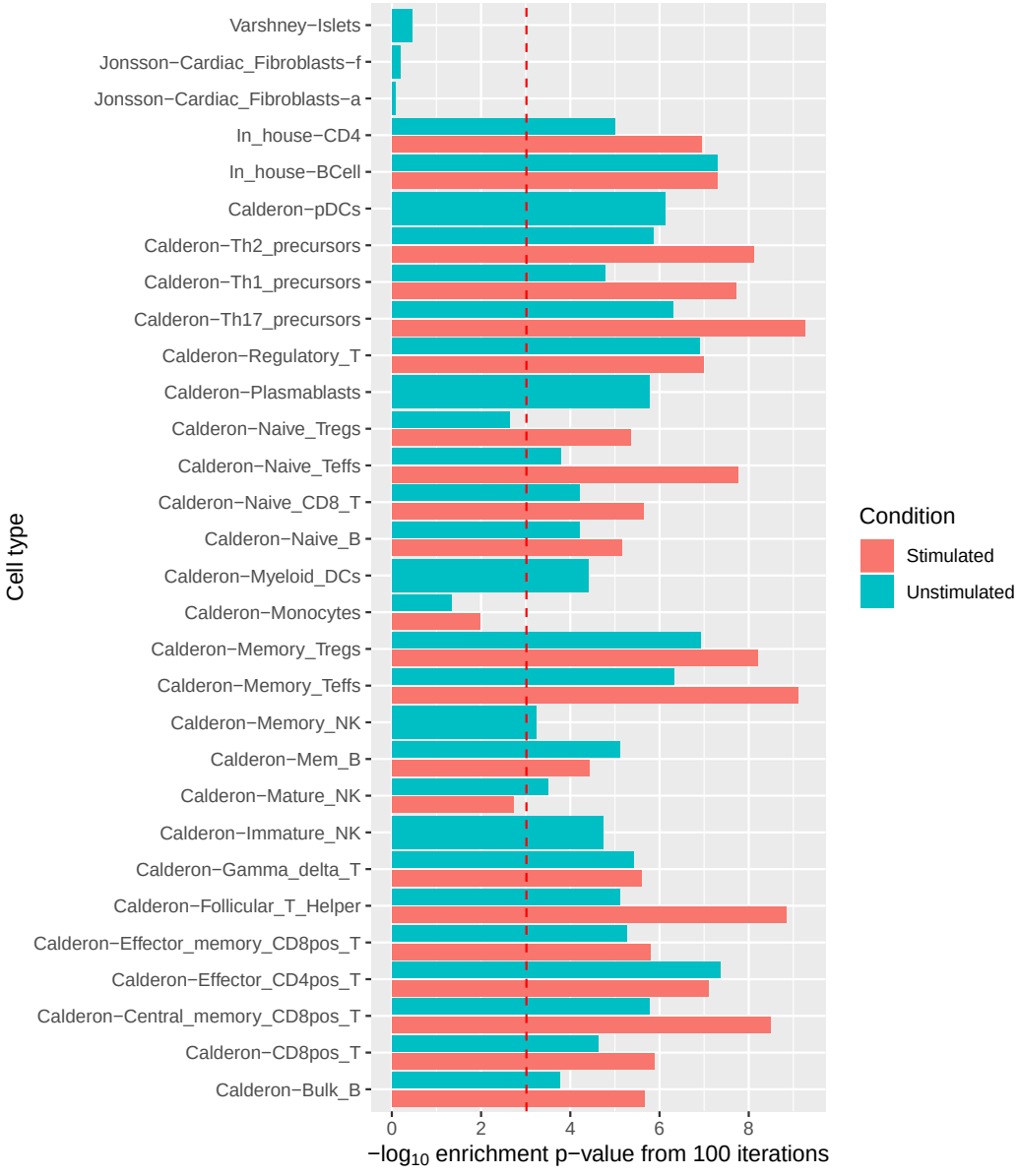
Figure 4.20: First vs. second PC from Stanford ATAC-seq data, using counts from the consensus list of peaks.



ATAC-seq peaks in immune cell type, both in unstimulated and stimulated conditions,

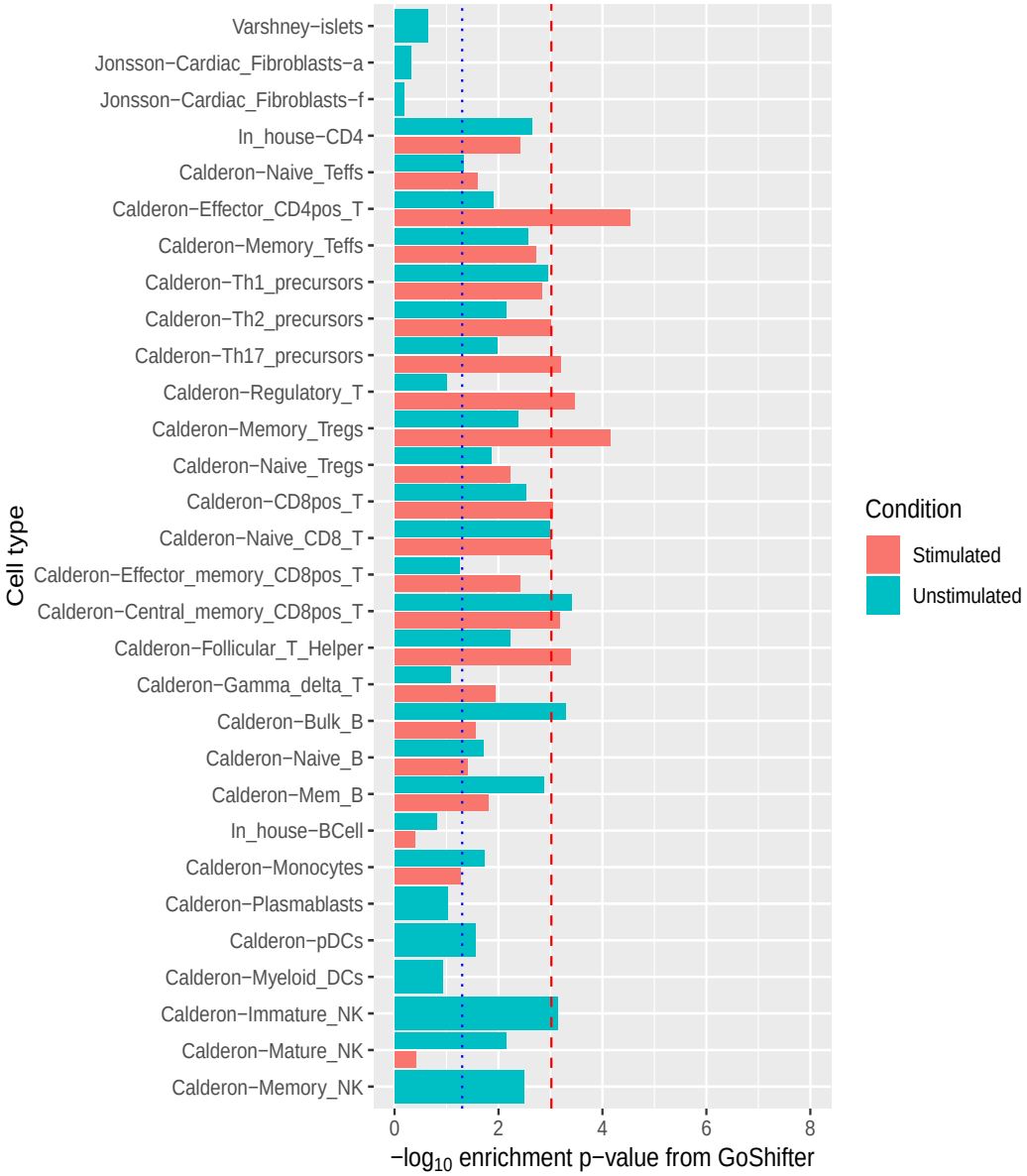
were enriched for T1D variants prioritised by GUESSFM ($p < 0.05/52 = 9.6 \times 10^{-4}$), with the exception of monocytes, unstimulated naive regulatory T cells and stimulated mature NK cells. There was no enrichment of GUESSFM prioritised variants in open chromatin in islets, nor in the control cell types, adult and foetal fibroblasts (Figure 4.21). These data are complimentary to the previously identified observation that T1D associations are enriched in functional regions of immune cells [144, 63]. In addition, for most cell types with stimulated and unstimulated conditions available, the enrichment in the stimulated state was stronger than the enrichment in the resting state.

Figure 4.21: Enrichment of T1D variants prioritised by GUESSFM in ATAC-seq peaks of various cell types, under resting and stimulated conditions when available. Red dashed line, Bonferroni significance at 5%.



I compared these enrichment results to results from GoShifter [192], a method that examines each region and compares the overlap of variants in open chromatin in each region to those where the ATAC-seq peaks are randomly shuffled within the region. The enrichment methods are similar but GoShifter ensures that the number of ATAC-seq peaks in the observed and the null distributions are matched, whereas the primary enrichment method used here matches for LD structure within the GUESSFM prioritised variants and the number of genes within that group of variants, allowing the number of peaks in the null distribution to vary. The GoShifter results were similar but displayed less strong enrichments in most cell types (Figure 4.22), though the results pointed to enrichment of T1D variants in ATAC-seq peaks in most immune cell types but not islets or the control cell types.

Figure 4.22: Enrichment of T1D variants prioritised by GUESSFM in ATAC-seq peaks of various cell types, under resting and stimulated conditions when available, using GoShifter, blue dotted line, nominal significance (p=0.05), red dashed line, Bonferroni significance at 5%.



Of the 2279 GUESSFM prioritised variants, 264 overlapped open chromatin in at least one cell type and state examined. Figure 4.23 shows there are some variants that lie in open chromatin in all cell types, including the control cell types and islets. However, some variants appear to be distinct between cell types and cell condition. There are a number of variants that are in open chromatin in stimulated cells but not unstimulated, and the reverse is also sometimes observed. There are only a small number of peaks that are specific to one cell type and condition only, which is perhaps expected given the large

number of similar cell types examined.

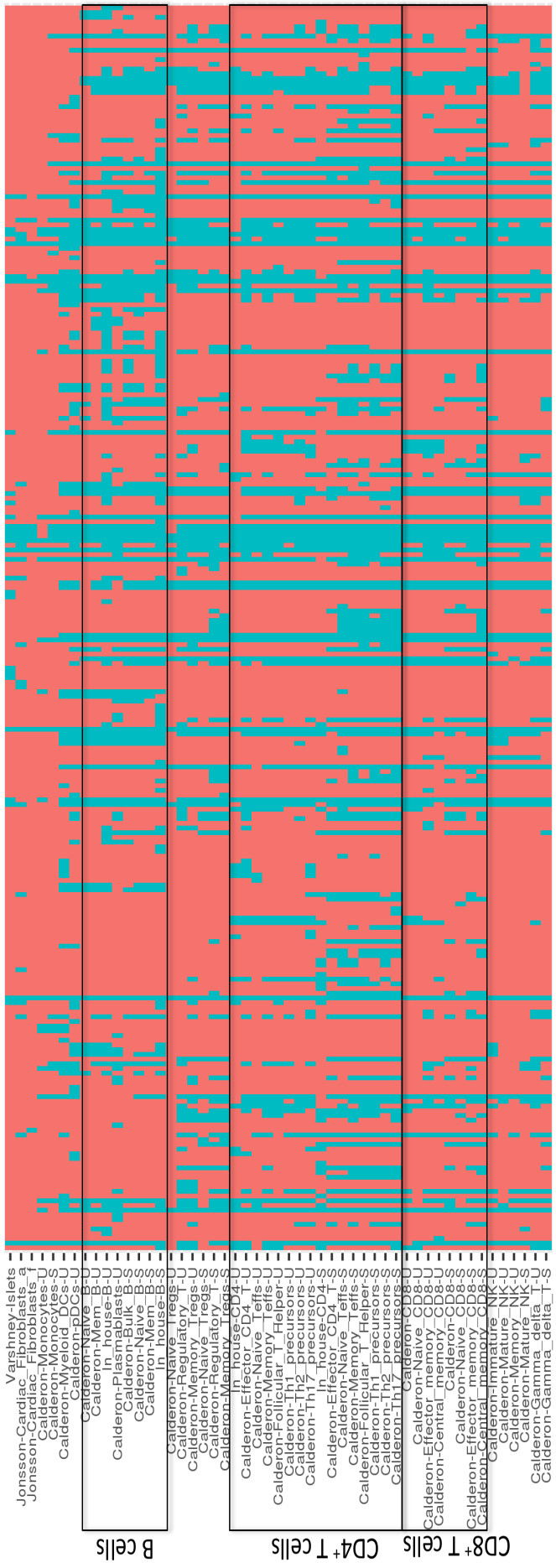


Figure 4.23: For each cell type, showing whether the variant overlaps open chromatin (blue) or not (red), displaying only the 264 GUESSFM prioritised variants that overlapped open chromatin in at least one cell type.

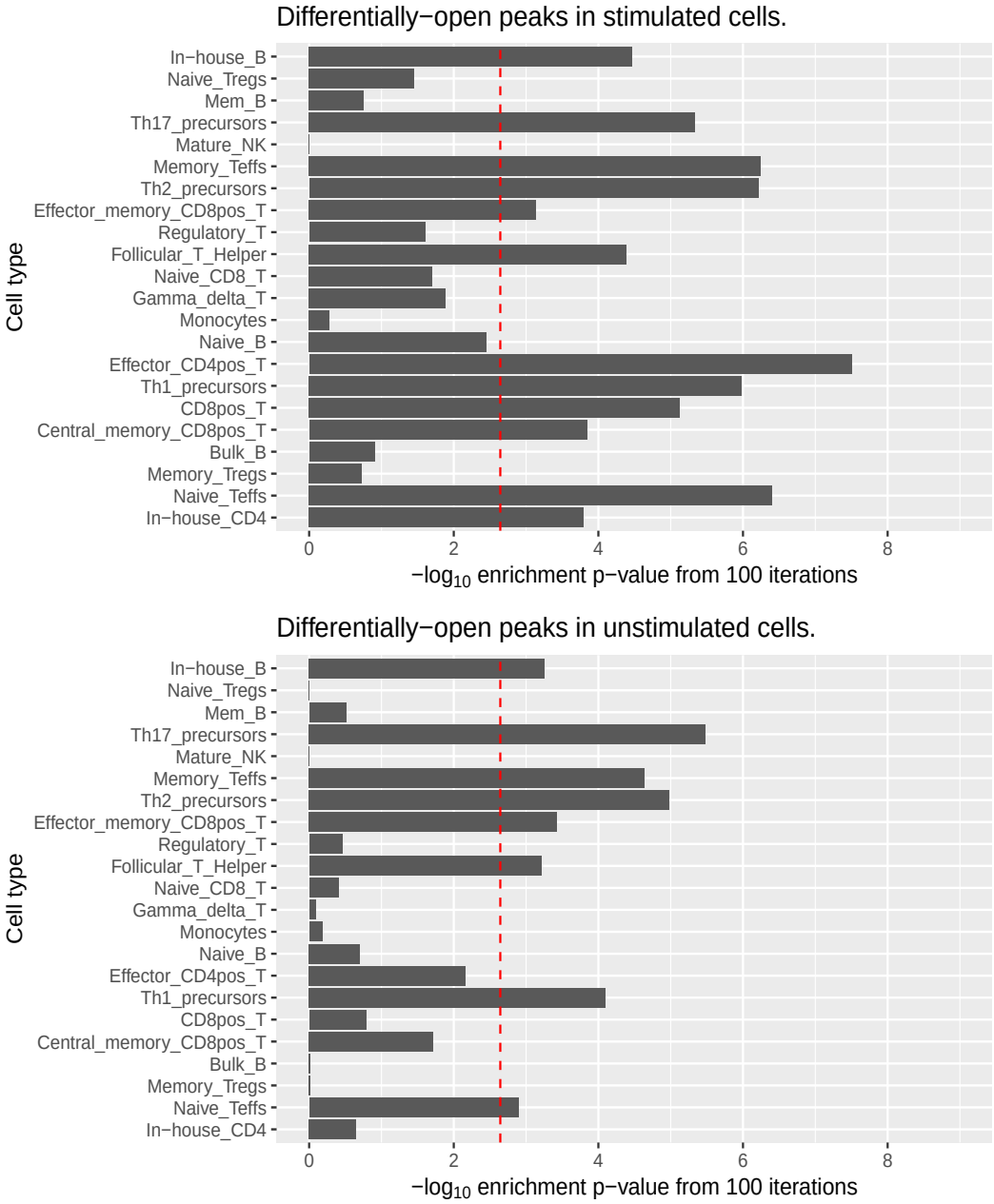
4.3.6 Stimulated vs unstimulated differential openness in ATAC-seq peaks overlapping T1D prioritised variants

The analysis examining whether variants are in ATAC-seq peaks or not without taking into consideration the counts within the ATAC-seq peaks did identify greater enrichments amongst stimulated immune cells but the focus is now on stimulation or non-stimulation specific peaks. Table 4.24 shows the number of stimulation or non-stimulation differential peaks observed in each cell type. Generally, there were more peaks that were differentially-open following stimulation, ranging for 0% to 15% of all peaks observed depending on the cell type. When the differential openness is taken into consideration in the ATAC-seq peaks, the greatest enrichment was observed in effector $CD4^+$ T cells (Figure 4.24), with similar enrichments observed in subsets of these cells, for example the memory T effectors and the Th1, Th2 and Th17 precursors. Results are broadly consistent with [12], which examined cell types after various different stimuli at different time points to those examined here but also identified stimulation-specific disease-associated variant enrichment in active chromatin regions in $CD4^+$ T cell subsets compared to other $CD4^+$ T cell subsets and resting $CD4^+$ T cells. Differentially-open ATAC-seq peaks were enriched for GUESSFM prioritised variants in Th17 precursors, in both stimulation-specific and unstimulation-specific peaks, implying that this is a dynamic cell type that could be affecting T1D risk in various states.

Table 4.5: Stimulation- or unstimulation-specific ATAC-seq peaks identified using DE-seq2.

Cell type	Activation-specific peaks (%)	Non activation-specific peaks (%)
In_house-CD4	1431 (1.03%)	1518 (1.1%)
Calderon-Naive_Teffs	11866 (8.56%)	4008 (2.89%)
Calderon-Memory_Tregs	5287 (3.81%)	390 (0.28%)
Calderon-Bulk_B	2529 (1.82%)	151 (0.11%)
Calderon-Central_memory_CD8pos_T	10544 (7.61%)	2389 (1.72%)
Calderon-CD8pos_T	9974 (7.2%)	2402 (1.73%)
Calderon-Th1_precursors	19452 (14.04%)	6965 (5.03%)
Calderon-Effector_CD4pos T	15823 (11.42%)	5247 (3.79%)
Calderon-Naive_B	4699 (3.39%)	625 (0.45%)
Calderon-Monocytes	539 (0.39%)	761 (0.55%)
Calderon-Gamma_delta_T	3118 (2.25%)	1029 (0.74%)
Calderon-Naive_CD8_T	3829 (2.76%)	598 (0.43%)
Calderon-Follicular_T_Helper	15641 (11.29%)	4001 (2.89%)
Calderon-Regulatory_T	6121 (4.42%)	510 (0.37%)
Calderon-Effector_memory_CD8pos_T	14604 (10.54%)	13637 (9.84%)
Calderon-Th2_precursors	6559 (11.95%)	6407 (4.62%)
Calderon-Memory_Teffs	18107 (13.06%)	6320 (4.56%)
Calderon-Mature_NK	1 (0%)	193 (0.14%)
Calderon-Th17_precursors	21195 (15.29%)	12339 (8.9%)
Calderon-Mem_B	2081 (1.5%)	561 (0.4%)
Calderon-Naive_Tregs	1118 (0.81%)	67 (0.05%)
In_house-B	13354 (9.64%)	8284 (5.98%)

Figure 4.24: Enrichment of T1D variants prioritised by GUESSFM in stimulated-specific (top panel) or unstimulated-specific (bottom panel) ATAC-seq peaks of various cell types.



4.3.7 Colocalisation analyses

There were four ATAC-seq peaks showing colocalisation between T1D risk variants and chromatin accessibility QTL in CD4$^+$ T cells (posterior probability of colocalisation $>$ 0.9). The first was on chromosome 5q11.2, positions $56,147,921 - 56,149,111$ (posterior probability of colocalisation$=$ 0.99). The T1D disease association in this regions also colocalises with whole-blood eQTLs for Ankyrin Repeat Domain 55 (*ANKRD55*) and

Interleukin 6 Signal Transducer (*IL6ST*) (posterior probability of colocalisation > 0.99), indicating that the variant in this region affects the accessibility of chromatin in this peak which possibly affects downstream transcription of these two genes.

Similarly, a chromatin accessibility QTL at a peak on chromosome 6q15 (positions $90,266,766 - 90,267,715$) near *BACH2* colocalised with the T1D association in the region (posterior probability of colocalisation = 0.97) and the same T1D association region colocalised with a whole-blood eQTL for *BACH2* (posterior probability of colocalisation = 1.00).

Another peak on chromosome 6q22.32, near Centromere Protein W (*CENPW*) (positions $126,339,725 - 126,340,580$) colocalised between chromatin accessibility QTL and the T1D disease association (posterior probability of colocalisation = 0.97), with T1D association and whole-blood eQTL colocalisation also observed with *CENPW* (posterior probability of colocalisation = 0.90).

Finally, a chromatin-accessibility QTL was identified that colocalised with the T1D-association signal at a peak on chromosome 14q32.2 (positions $98,018,322 - 98,019,163$, posterior probability of colocalisation = 0.97). There was not strong evidence for colocalisation between the T1D association and any whole-blood eQTL in the region, but there was suggestive evidence with *RP11-61O1.1* (posterior probability of colocalisation = 0.63).

4.4 Discussion

As a result of increased sample size and therefore statistical power, 22 novel regions of association have been identified in this analysis to genome-wide significance. Moreover, relaxing the association threshold has more than doubled the number of regions associated with T1D to date, from 60 to 146. This indicates that there are likely many more associations with T1D, which are either very rare or have increasingly small effect sizes. Two recent publications indicate that much trait heritability can be explained by rare variants [200, 78], and this may also be the case for T1D. The ultimate aim of T1D genetic analyses are to uncover drug targets that can slow, reverse or prevent T1D disease course. The rich source of genes and pathways that the current analysis highlights as associated with T1D provide an opportunity to examine how these targets can affect risk of autoimmunity in animal models and then ultimately in humans. The benefits of performing a genetic analysis of increased sample size has been to increase certainty that previously borderline T1D-associated regions are genuine, to increase fine-mapping resolution for each associated region, and to begin uncovering rare variants associated with disease, which tend to have larger effect sizes. Now a larger genetic analysis has been carried out, it is appropriate to, instead of increasing the sample size again to identify more regions with smaller effect sizes, focus on fully understanding the mechanism behind the most associated regions and how to target this mechanism to alter disease risk. There would still be benefit, however, in performing a large genetic analysis for T1D using an unbiased GWAS chip, where entirely new pathways could be identified, unlike the current ImmunoChip analysis, which is biased towards known immune genes. Indeed, of the 22 novel genome-wide associated chromosome regions, 14 are located outside the 184 densely genotyped ImmunoChip regions, suggesting that the increased sample size may highlight many additional chromosome regions away from the ImmunoChip regions, which were defined based on initial GWAS analyses with lower statistical power.

By identifying the causal genes in these regions, and examining which pathways they are in and which other proteins they are interacting with, novel therapeutic targets to treat T1D could be identified. For example, the novel associated region on chromosome 2q32.2,

which has been associated with T1D previously by combining multiple autoimmune diseases but never by using T1D cases and controls only [128] contains the Signal Transducer And Activator Of Transcription 4 (STAT4) gene. This is a transcription factor known to be phosphorylated by Janus Kinase (JAK) proteins, which induces downstream transcription. Inhibition of these JAK proteins using a drug called Tofacitinib has been shown to be safe [43] and efficacious in reducing the disease activity in rheumatoid arthritis patients [209] and increase remission rates amongst individuals with active ulcerative colitis [164]. There is some evidence from mouse models that this inhibitor could slow disease course in coeliac disease [214], systemic lupus erythematosus [69] and T1D under conditions of increased IL-15 expression [40]. Therefore there may be benefit in identifying drug targets not only from examining the candidate gene of association in the region, but also examining the genes that the candidate gene interacts with. A recent analysis [61] prioritised genes in a similar way to this in order to identify drug targets in a number of immune-related diseases, including T1D, but using summary statistics from a smaller sample size than reported here. Of the 32 target genes identified in that analysis, just one modulator of the target gene has been tested in a phase I trial to treat autoimmunity, *IL2RA*. It is possible that with this increased sample size and number of associated regions, more genes might be identified that already have drugs developed to target and could be repurposed as a T1D therapy. This analysis is to be considered in future work. Whilst the region on chromosome 11q12.2, near the CD6 gene has been implicated in other autoimmune diseases [117, 57], the association signal for T1D appears distinct from the other autoimmune disease associations, with the index variant being rs11230563 (C>T) for IBD and rs79538630 (G>T) for T1D ($r^2 = 0.02$). Although the genetic signals are distinct, the CD5 and CD6 genes are now implicated in T1D, which are suggested therapeutic targets to suppress autoimmunity [113, 3]. Similarly, the associated region on chromosome 5p13.1, near Prostaglandin E Receptor 4 (*PTGER4*) has shown association with Crohn's disease, but the T1D and Crohn's disease variants appear distinct from one another (r^2 between T1D index variant rs1876142 (T>G) and Crohn's index variant rs7711427 (C>A) = 0.04). It has been shown that inhibiting PTGER4 reduces IL-1

signalling and reduces lymphocyte infiltration in pre-diabetic mice [154], suggesting this may be a novel therapeutic target for T1D, identified as a result of larger sample sizes in this genetic analysis. Finally, a novel association on chromosome 6p25.3 is near the interferon regulatory factor 4 (IRF4) gene. It has been shown that reduced IRF4 in mice protects against T1D [1] and the genetic association from the present analysis increases candidacy to target this molecule to prevent or treat T1D.

In addition to a large variant-discovery analysis, the other major focus of this analysis was to fine map each T1D-associated region to identify the most likely causal variants driving the association. There was good concordance between the current fine mapping results and the previous iteration of fine mapping, performed on a subset of individuals included in this analysis [31]. I highlighted a number of regions that were of particular interest for a number of reasons. Firstly, two regions that contain a 'tag' SNP, which is a problem that is now highlighted in the literature [202, 8], where the index variant from the univariable analysis is unlikley to be causal but is in LD with two disease-associated variants. Secondly, a region where the resolution has been improved due to increase sample size (near *BACH2*). Finally, I highlighted two regions of particular complexity. The most complex region of association with T1D is the region near *IL2RA*. The four previously identified disease signals [202] were identified here, including group A, which contains a variant that has been shown to be causal for autoimmunity [155], but two additional signals were also identified, indicating any functional analyses in this region have many other genetic factors to control for. The *CTLA4* and *IFIH1* regions are similarly complex, with a mixture of common and rare variants associated with disease risk.

A novel aspect to my analyses is the inclusion of African ancestry cases and controls. The association statistics from the African ancestry group were used to prioritise the most likely causal variants in combination with the European association summary statistics. There were three regions where the benefit of this approach was striking, where the European association was part of a large haplotype and the African association clearly prioritised a subset of variants within the haplotype. In the *CENPW* and *RBPJ* regions,

the variants prioritised in the trans-ethnic fine mapping analysis also lay in likely functional regions in immune cell types. This could happen by chance, but also might be the case that the trans-ethnic fine mapping analysis has prioritised the causal variants, which are intuitively more likely to be in open chromatin in immune cell types. The association with T1D in the *CENPW* region has been shown to colocalise with type 2 diabetes risk in the region [11] and is not associated with any other autoimmune disease. *CENPW* is expressed in many cell types (`https://dice-database.org/genes/CENPW`) and tissues (`https://gtexportal.org/home/`) and the expression of *CENPW* in islets is positively correlated with levels of transcription factor Nuclear Factor Of Activated T Cells 2 (NFATC2) in mice [101]. It is also associated with T2D relevant processes [65], indicating there might be an underlying process that increases the risk of T1D and T2D, for example beta-cell fragility [55].

The enrichment of T1D variants prioritised from fine-mapping analyses in open chromatin of immune cell types is unsurprising and consistent with previous observations [63, 144]. However, the results should be interpreted with caution for two reasons. The first is that the ImmunoChip is biased towards regions that have shown association to one or more autoimmune disease and therefore it would be unsurprising if the associated regions preferentially lay in regions active in immune cell types. Secondly, the enrichment analyses in this chapter are performed by examining where the GUESSFM prioritised variants are located in relation to open chromatin in each cell type. Only regions of dense genotyping were fine mapped and therefore the enrichment analyses were specifically examining the enrichment of GUESSFM prioritised variants that lie in densely genotyped ImmunoChip regions, so it is not an unbiased examination of ImmunoChip variants that are examined. The lack of enrichment of GUESSFM prioritised variants in open chromatin in islets may be as a result of this design, since it is expected that islets are not simply innocent bystanders in the aetiology of T1D. It remains to be seen whether enrichment would be observed if a GWAS platform were used in a comparably large genetic association study. The trend toward T1D-associated variants being more strongly associated with early-diagnosed T1D was also observed in the novel regions in this analysis, where 17 of the

22 variants were examined using the same collection of individuals as were studied in chapter 3, with eight having the largest effect size in the < 7 group, intermediate effect in the $7 - 13$ group and the largest effect in the ≥ 13 group, compared to $\approx$ three expected by chance. This result gives further explanation for the increased heritability observed amongst those diagnosed at a young age, as even the rarer variants or variants with smaller effect sizes have bigger effects in early-diagnosed T1D compared to later diagnosed [159, 158].

In conclusion, the largest genetic analysis of T1D to date has identified up to 146 regions of association with T1D. Despite the fact that the largest genetic effects for T1D are in the HLA region, this analysis provides evidence that there are many more regions outside the HLA region that contribute towards T1D risk. This reflects the genetic architecture of many complex traits, for example height, BMI [213] and T2D [123].

Chapter 5

Conclusions

In this thesis, I have examined genetic determinants of AAD of T1D, which lead to the identification of a novel T1D risk region that acts most strongly in individuals diagnosed at < 5 years. This region on chromosome 6 contains the PTPRK and THEMIS genes, both of which are expressed in the thymus, an organ that is most active early in life and functions not only to produce T cells but also to delete reactive-to-self effector T cells. In addition to examining the genetic determinants of AAD in the genome-wide fashion, I also examined each robustly T1D-associated chromosome region in the genome, to determine if any acted particularly strongly in those diagnosed at < 7 years compared to ≥ 13 years, guided by the observation that there are different profiles of immune cells infiltrating the pancreatic islets between T1D individuals diagnosed in these age categories. I found that the six classical alleles or haplotypes in the HLA region and six non-HLA loci showed larger effect sizes in the < 7 group compared to the ≥ 13 group, as well as identifying evidence that many other loci could be consistent with this trend if there were greater statistical power in our analysis to detect such differences. Finally, I performed the largest genetic analysis of T1D to date, with over $60,000$ individuals, identifying 74 chromosome regions associated with T1D at a p value of $< 5 \times 10^{-8}$, but suggestive evidence of 146 regions with a relaxed association threshold of FDR < 0.01. I fine mapped each robustly-associated region that was in a densely-genotyped region and then integrated publicly available ATAC-seq data with these prioritised variants to identify whether open chromatin in any cell type under consideration was enriched for

variants prioritised in the fine mapping. I found an enrichment in many immune cell types when examining all ATAC-seq peaks, and a particular enrichment in $CD4^+$ T cells when examining stimulation-specific peaks. Overall, I have expanded knowledge of the genetics of T1D, highlighted the regions and pathways that predispose to the most aggressive form of the disease, the early diagnosed, and provided some mechanistic insight into how some chromosome regions might effect T1D risk through *in silico* analyses.

Since the most aggressive form of T1D is observed in individuals diagnosed at a young age [121, 14, 54], and it is known that loss of C-peptide is associated with increased T1D complications [106], this population represents those in most urgent need of therapeutic intervention and with the largest potential impact on T1D-related health outcomes. If this is to be done in a genetics-driven fashion, for which there is an estimated two-fold increased success rate for downstream drug approval over non-genetics driven trials [136], my research points towards the genetic regions of susceptibility that should be studied and understood with regards to disease mechanism. The primary genetic driver of early-diagnosed T1D is the HLA region. It is known that some of the susceptibility in the HLA region for T1D risk comes from the structure that the HLA molecules form as a result of the genotype, for example both HLA DRB1*04:01 and DRB1*03:02 result in binding groove structures that associate with disease risk [161]. However, it has also been shown that the HLA alleles can alter the T cell receptor repertoire of an individual [171] and the development of intestinal microbiome [146], indicating that the association of this region with T1D could be acting on many levels in multiple tissues. Understanding fully the mechanism behind increased T1D risk in this region is likely to be the most effective way to decrease T1D incidence, since this is the region with the largest effect on disease, which is largest in the early-diagnosed individuals but remains large even in later-diagnosed individuals.

Outside the HLA region, amongst common variants showing association with T1D, six regions of association, near *PTPRK*, *GLIS3*, *CTSH*, *IKZF3*, *IL2RA* and *IL10* should be examined functionally to understand the mechanism driving these associations. If the hypothesised but controversial [210] 'omnigenic' model of genetic association is true,

whereby each complex trait is driven by a small number of core genes, then additional associations will be revealed that lie in the pathways of the core genes [29], then fully understanding the associations highlighted here, as well as other regions with particularly strong association with T1D (for example near *INS* and *PTPN22*), may inform on the drug targets likely to be effective on the largest number of people. Even if the 'omnigenic' model is incorrect, and each disease association across the genome is biologically independent, the effect sizes may become so small that therapeutic intervention in these pathways could be ineffective. However, pharmacological intervention will have both on-target and off-target effects, making them unpredictable, and therefore it is possible that targets identified from a variant with a small effect size or a rare variant could lead to an effective medicine. To truly understand T1D aetiology, though, would require mechanistic understanding of each genetic association, which is a sensible next step in T1D research.

The analyses in chapter 4 have expanded the knowledge of the genetic architecture of T1D, and points towards T1D being a complex trait with many more associated genetic regions than had previously been identified. There is additional future work that can and should be carried out using this dataset. For example in [61] the authors use 3-dimensional chromatin formation data, eQTL data and protein-protein interactions to identify potential novel therapeutic targets. This kind of analysis is an important next step for the MEGAT1D analysis and will be considered in due course. Perhaps of more benefit, however, would be to examine a dataset of this size, but using genome-wide as opposed to ImmunoChip data, where novel targets are more likely to be identified, given the genome will be examined in an unbiased way. To date, there are approximately five T1D-associated loci that are also associated with T2D (near *GLIS3*, *CENPW*, *CTRB1* *RASGRP1* and *INS*), but with increasing sample sizes and by looking genome-wide, the possibility of identifying more loci that alter disease risk for both diseases seems likely. On a related note, the MEGAT1D analysis did not identify any enrichment of T1D-associated variants in open chromatin of islets; this result may be misleading as the platform used to assess the enrichment (the ImmunoChip) was biased towards regions associated with

immune-related conditions. For this reason, performing a large T1D association study with an unbiased genotyping platform is an important analysis to perform, since the downstream enrichment analyses will potentially point towards novel biological pathways and drug targets. Another question that can be examined using the MEGAT1D database is to more fully resolve the associations in the HLA region. Previous analyses have identified many disease-associated class I classical alleles and class II haplotypes, but it is likely that even after conditioning on these covariates, additional signal remains in the region. For example, a recent analysis implicated a gene near the HLA region, the compliment 4 (C4) gene, in the pathogenesis of SLE and as an underlying cause of the disease prevalence sex bias of the disease [97]. They did this by imputing $C4$ alleles from ImmunoChip data. This analysis and others can be performed for T1D using the MEGAT1D dataset to try to understand the genetic architecture in this most strongly associated region. Such an analysis is likely to take a number of years and is thus beyond the scope of this thesis.

Aside from synthetic insulin injections, there has been limited success in identifying efficacious treatments for T1D, with many drug trials failing to show efficacy in the primary endpoint [74, 203, 102, 174]. Recent mechanistic trials have shown some promise with intervention shortly after diagnosis with, for example, low-dose IL-2 [189, 168], administration of peripheral blood- derived regulatory T cell populations [148] or peptide immunotherapy [2], where an immunodominant proinsulin peptide for the HLA DRB1*04:01 molecule has been administered and shown to be safe. With focus on translation of genetic associations into therapeutic targets, such as in cardiovascular disease, which has had drugs approved to lower low-lipoprotein cholesterol levels based on genetic association analyses [59], the next therapeutic target or approach to slow, prevent or reverse T1D is on the horizon.

Bibliography

[1] S Akazawa et al. "Haploinsufficiency of interferon regulatory factor 4 strongly protects against autoimmune diabetes in NOD mice". In: *Diabetologia* 58.11 (2015), pp. 2606–2614.

[2] M Alhadj Ali et al. "Metabolic and immune effects of immunotherapy with proinsulin peptide in human new-onset type 1 diabetes". In: *Science translational medicine* 9.402 (2017), eaaf7779.

[3] R Alonso-Ramirez et al. "Rationale for targeting CD6 as a treatment for autoimmune diseases". In: *Arthritis* 2010 (2010), p. 130646.

[4] N Amin, CM Van Duijn, and YS Aulchenko. "A genomic background based method for association analysis in related individuals". In: *PloS one* 2.12 (2007), e1274.

[5] CA Anderson et al. "Data quality control in genetic case-control association studies". In: *Nature protocols* 5.9 (2010), pp. 1564–1573.

[6] AF Arechiga et al. "Cutting edge: the PTPN22 allelic variant associated with autoimmunity impairs B cell signaling". In: *The Journal of Immunology* 182.6 (2009), pp. 3343–3347.

[7] A Asano et al. "A deletion mutation of the protein tyrosine phosphatase kappa (Ptprk) gene is responsible for T-helper immunodeficiency (thid) in the LEC rat". In: *Mammalian Genome* 18.11 (2007), pp. 779–786.

[8] JL Asimit et al. "Stochastic search and joint fine-mapping increases accuracy and identifies previously unreported associations in immune-mediated diseases". In: *Nature communications* 10.1 (2019), p. 3216.

[9] W Astle and DJ Balding. "Population structure and cryptic relatedness in genetic association studies". In: *Statistical Science* 24.4 (2009), pp. 451–471.

[10] YS Aulchenko, D De Koning, and C Haley. "Genomewide rapid association using mixed model and regression: a fast and simple method for genomewide pedigree-based quantitative trait loci association analysis". In: *Genetics* 177.1 (2007), pp. 577–585.

[11] AJ Aylward et al. "Shared genetic risk contributes to type 1 and type 2 diabetes etiology". In: *Human Molecular Genetics* (2018), 10.1093/hmg/ddy314.

[12] Soskic B et al. "Chromatin activity at GWAS loci identifies T cell states driving complex immune diseases". In: *Nature genetics* (2019).

[13] SC Bain, JA Todd, and AH Barnett. "The British Diabetic Association-Warren Repository". In: *Autoimmunity* 7.2-3 (1990), pp. 83–85.

[14] A Barker et al. "Age-dependent decline of β-cell function in type 1 diabetes after diagnosis: a multi-centre longitudinal study". In: *Diabetes, Obesity and Metabolism* 16.3 (2014), pp. 262–267.

[15] JC Barrett et al. "Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes". In: *Nature genetics* 41.6 (2009), pp. 703–707.

[16] I Barroso and MI McCarthy. "The genetic basis of metabolic disease". In: *Cell* 177.1 (2019), pp. 146–161.

[17] T Bayes. "LII. An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, FRS communicated by Mr. Price, in a letter to John Canton, AMFR S". In: *Philosophical transactions of the Royal Society of London* 53 (1763), pp. 370–418.

[18] AH Beecham et al. "Analysis of immune-related loci identifies 48 new susceptibility variants for multiple sclerosis". In: *Nature genetics* 45.11 (2013), pp. 1353–1360.

[19] Y Benjamini and Y Hochberg. "Controlling the false discovery rate: a practical and powerful approach to multiple testing". In: *Journal of the Royal statistical society: series B (Methodological)* 57.1 (1995), pp. 289–300.

[20] Y Benjamini and D Yekutieli. "The control of the false discovery rate in multiple testing under dependency". In: *The annals of statistics* 29.4 (2001), pp. 1165–1188.

[21] C Benner et al. "FINEMAP: efficient variable selection using summary data from genome-wide association studies". In: *Bioinformatics* 32.10 (2016), pp. 1493–1501.

[22] C Bondar et al. "THEMIS and PTPRK in celiac intestinal mucosa: coexpression in disease and after in vitro gliadin challenge". In: *European Journal of Human Genetics* 22.3 (2014), pp. 358–362.

[23] D Boomsma, A Busjahn, and L Peltonen. "Classical twin studies and beyond". In: *Nature reviews genetics* 3.11 (2002), pp. 872–882.

[24] N Bottini et al. "A functional variant of lymphoid tyrosine phosphatase is associated with type I diabetes". In: *Nature genetics* 36.4 (2004), pp. 337–338.

[25] N Bottini et al. "Role of PTPN22 in type 1 diabetes and other autoimmune diseases". In: *Seminars in immunology* 18.4 (2006), pp. 207–213.

[26] L Bottolo and S Richardson. "Evolutionary stochastic search for Bayesian model exploration". In: *Bayesian Analysis* 5.3 (2010), pp. 583–618.

[27] TJ Bouchard. "The Wilson effect: the increase in heritability of IQ with age". In: *Twin Research and Human Genetics* 16.5 (2013), pp. 923–930.

[28] E Bouzigon et al. "Effect of 17q21 variants and smoking exposure in early-onset asthma". In: *New England journal of medicine* 359.19 (2008), pp. 1985–1994.

[29] EA Boyle, YI Li, and JK Pritchard. "An expanded view of complex traits: from polygenic to omnigenic". In: *Cell* 169.7 (2017), pp. 1177–1186.

[30] JP Bradfield et al. "A genome-wide meta-analysis of six type 1 diabetes cohorts identifies multiple associated loci". In: *PLoS genetics* 7.9 (2011), e1002293.

[31] OS Burren et al. “Chromosome contacts in activated T cells identify autoimmune disease candidate genes”. In: *Genome biology* 18.1 (2017), p. 165.

[32] D Calderon et al. “Landscape of stimulation-responsive chromatin across diverse human immune cells”. In: *Nature genetics* (2019).

[33] A Carreras-Sureda et al. “ORMDL3 modulates store-operated calcium entry and lymphocyte activation”. In: *Human molecular genetics* 22.3 (2013), pp. 519–530.

[34] A Casu et al. “Type 1 diabetes among Sardinian children is increasing: the Sardinian diabetes register for children aged 0–14 years (1989–1999)”. In: *Diabetes care* 27.7 (2004), pp. 1623–1629.

[35] CR Catacchio et al. “Inversion variants in human and primate genomes”. In: *Genome research* 28.6 (2018), pp. 910–920.

[36] F Cerutti et al. “Younger age at onset and sex predict celiac disease in children and adolescents with type 1 diabetes: an Italian multicenter study”. In: *Diabetes care* 27.6 (2004), pp. 1294–1298.

[37] M Chabod et al. “A spontaneous mutation of the rat Themis gene leads to impaired function of regulatory T cells linked to inflammatory bowel disease”. In: *PLoS genetics* 8.1 (2012), e1002461.

[38] MJP Chaisson et al. “Multi-platform discovery of haplotype-resolved structural variation in human genomes”. In: *Nature communications* 10.1 (2019), p. 1784.

[39] H Chen et al. “Control for population structure and relatedness for binary traits in genetic association studies via logistic mixed models”. In: *The American Journal of Human Genetics* 98.4 (2016), pp. 653–666.

[40] J Chen et al. “Insulin-dependent diabetes induced by pancreatic beta cell expression of IL-15 and IL-15Rα”. In: *Proceedings of the National Academy of Sciences* 110.33 (2013), pp. 13534–13539.

[41] W Chen et al. “Fine mapping causal variants with an approximate Bayesian method using marginal test statistics”. In: *Genetics* 200.3 (2015), pp. 719–736.

[42] H Chipman, EI George, and RE McCulloch. "The practical implementation of Bayesian model selection". In: *IMS Lecture Notes-Monograph Series* 38 (2001), pp. 65–134.

[43] SB Cohen et al. "Long-term safety of tofacitinib for the treatment of rheumatoid arthritis up to 8.5 years: integrated analysis of data from the global clinical trials". In: *Annals of the rheumatic diseases* 76 (2017), pp. 1253–1262.

[44] MP Conomos, MB Miller, and TA Thornton. "Robust inference of population structure for ancestry prediction and correction of stratification in the presence of relatedness". In: *Genetic epidemiology* 39.4 (2015), pp. 276–293.

[45] 1000 Genomes Project Consortium et al. "A global reference for human genetic variation". In: *Nature* 526.7571 (2015), pp. 68–74.

[46] ENCODE Project Consortium et al. "An integrated encyclopedia of DNA elements in the human genome". In: *Nature* 489.7414 (2012), pp. 57–74.

[47] GTEx Consortium et al. "Genetic effects on gene expression across human tissues". In: *Nature* 550.7675 (2017), pp. 204–213.

[48] Wellcome Trust Case Control Consortium et al. "Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls". In: *Nature* 447.7145 (2007), pp. 661–678.

[49] NJ Cooper et al. "Type 1 diabetes genome-wide association analysis with imputation identifies five new risk regions". In: *bioRxiv* (2017), p. 120022.

[50] A Cortes and MA Brown. "Promise and pitfalls of the Immunochip". In: *Arthritis research & therapy* 13.1 (2011), p. 101.

[51] ME Craig et al. "Prevalence of celiac disease in 52,721 youth with type 1 diabetes: international comparison across three continents". In: *Diabetes Care* 40.8 (2017), pp. 1034–1040.

[52] D Dabelea et al. "Clinical evolution of beta cell function in youth with diabetes: the SEARCH for Diabetes in Youth study". In: *Diabetologia* 55.12 (2012), pp. 3359–3368.

[53] S Das et al. "Next-generation genotype imputation service and methods". In: *Nature genetics* 48.10 (2016), pp. 1284–1287.

[54] AK Davis et al. "Prevalence of detectable C-peptide according to age at diagnosis and duration of type 1 diabetes". In: *Diabetes care* 38.3 (2015), pp. 476–481.

[55] J Dooley et al. "Genetic predisposition for beta cell fragility underlies type 1 and type 2 diabetes". In: *Nature genetics* 48.5 (2016), pp. 519–527.

[56] MJ Dufort et al. "Cell type–specific immune phenotypes predict loss of insulin secretion in new-onset type 1 diabetes". In: *JCI insight* 4.4 (2019), p. 125556.

[57] D Ellinghaus et al. "Analysis of five chronic inflammatory diseases identifies 27 new associations and highlights disease-specific patterns at shared loci". In: *Nature genetics* 48.5 (2016), pp. 510–518.

[58] WJ Ewens and RS Spielman. "The transmission/disequilibrium test: history, subdivision, and admixture." In: *American journal of human genetics* 57.2 (1995), pp. 455–464.

[59] L Fala. "Repatha (Evolocumab): second PCSK9 inhibitor approved by the FDA for patients with familial hypercholesterolemia". In: *American health & drug benefits* 9.Spec Feature (2016), pp. 136–139.

[60] DS Falconer. "The inheritance of liability to certain diseases, estimated from the incidence among relatives". In: *Annals of human genetics* 29.1 (1965), pp. 51–76.

[61] H Fang et al. "A genetics-led approach defines the drug target landscape of 30 immune-related traits". In: *Nature genetics* 51.7 (2019), pp. 1082–1091.

[62] W Fang et al. "Association of the CTLA4 gene CT60/rs3087243 single-nucleotide polymorphisms with Graves' disease". In: *Biomedical reports* 3.5 (2015), pp. 691–696.

[63] KK Farh et al. "Genetic and epigenetic fine mapping of causal autoimmune disease variants". In: *Nature* 518.7539 (2015), pp. 337–343.

[64] A Fasano. "Clinical presentation of celiac disease in the pediatric population". In: *Gastroenterology* 128.4 (2005), S68–S73.

[65] J Fernández-Tajes et al. "Developing a network view of type 2 diabetes risk pathways through integration of genetic, genomic and functional data". In: *Genome medicine* 11.1 (2019), p. 19.

[66] MD Fortune et al. "Statistical colocalization of genetic risk variants for related autoimmune diseases in the context of common controls". In: *Nature genetics* 47.7 (2015), pp. 839–846.

[67] A Frankish et al. "GENCODE reference annotation for the human and mouse genomes". In: *Nucleic acids research* 47.D1 (2018), pp. D766–D773.

[68] G Fu et al. "Themis controls thymocyte selection through regulation of T cell antigen receptor–mediated signaling". In: *Nature immunology* 10.8 (2009), pp. 848–856.

[69] Y Furumoto et al. "Tofacitinib ameliorates murine lupus and its associated vascular dysfunction". In: *Arthritis & Rheumatology* 69.1 (2017), pp. 148–160.

[70] JM Gaspar. "Improved peak-calling with MACS2". In: *BioRxiv* (2018), p. 496521.

[71] S Genovese et al. "Distinct cytoplasmic islet cell antibodies with different risks for type 1 (insulin-dependent) diabetes mellitus". In: *Diabetologia* 35.4 (1992), pp. 385–388.

[72] C Giambartolomei et al. "Bayesian test for colocalisation between pairs of genetic association studies using summary statistics". In: *PLoS genetics* 10.5 (2014), e1004383.

[73] WR Gilks, S Richardson, and DJ Spiegelhalter. *Introducing markov chain monte carlo*. 1996. ISBN: 0412055511.

[74] CJ Greenbaum et al. "Through the fog: recent clinical trials to preserve β-cell function in type 1 diabetes". In: *Diabetes* 61.6 (2012), pp. 1323–1330.

[75] R Hannonen et al. "Academic skills in children with early-onset type 1 diabetes: the effects of diabetes-related risk factors". In: *Developmental Medicine & Child Neurology* 54.5 (2012), pp. 457–463.

[76] V Harjutsalo, A Reunanen, and J Tuomilehto. "Differential transmission of type 1 diabetes from diabetic fathers and mothers to their offspring". In: *Diabetes* 55.5 (2006), pp. 1517–1524.

[77] V Harjutsalo et al. "Incidence of type 1 diabetes in Finland". In: *The Journal of the American Medical Association* 310.4 (2013), pp. 427–428.

[78] RD Hernandez et al. "Ultra-rare variants drive substantial cis-heritability of human gene expression". In: *Nature genetics* 51.9 (2019), pp. 1349–1355.

[79] F Hormozdiari et al. "Colocalization of GWAS and eQTL signals detects target genes". In: *The American Journal of Human Genetics* 99.6 (2016), pp. 1245–1260.

[80] F Hormozdiari et al. "Identifying causal variants at loci with multiple signals of association". In: *Genetics* 198.2 (2014), pp. 497–508.

[81] B Howie, J Marchini, and M Stephens. "Genotype imputation with thousands of genomes". In: *G3: Genes, Genomes, Genetics* 1.6 (2011), pp. 457–470.

[82] BN Howie, P Donnelly, and J Marchini. "A flexible and accurate genotype imputation method for the next generation of genome-wide association studies". In: *PLoS genetics* 5.6 (2009), e1000529.

[83] JMM Howson et al. "Confirmation of HLA class II independent type 1 diabetes associations in the major histocompatibility complex including HLA-B and HLA-A". In: *Diabetes, Obesity and Metabolism* 11.Suppl 1 (2009), pp. 31–45.

[84] JMM Howson et al. "Evidence of gene-gene interaction and age-at-diagnosis effects in type 1 diabetes". In: *Diabetes* 61.11 (2012), pp. 3012–3017.

[85] JMM Howson et al. "Genetic analysis of adult-onset autoimmune diabetes". In: *Diabetes* 60.10 (2011), pp. 2645–2653.

[86] H Huang et al. "Fine-mapping inflammatory bowel disease loci to single-variant resolution". In: *Nature* 547.7662 (2017), pp. 173–178.

[87] V Hyttinen et al. "Genetic liability of type 1 diabetes and the onset age among 22,650 young Finnish twin pairs: a nationwide follow-up study". In: *Diabetes* 52.4 (2003), pp. 1052–1055.

[88] Jamie RJ Inshaw et al. "Genetic variants predisposing most strongly to type 1 diabetes diagnosed under age 7 years lie near candidate genes that function in the immune system and in pancreatic β-cells". In: *Diabetes Care* 43.1 (2020), pp. 169–177.

[89] JRJ Inshaw et al. "The chromosome 6q22.33 region is associated with age at diagnosis of type 1 diabetes and disease risk in those diagnosed under 5 years of age". In: *Diabetologia* 61.1 (2018), pp. 147–157.

[90] B Iovane et al. "Diabetic ketoacidosis at the onset of Type 1 diabetes in young children Is it time to launch a tailored campaign for DKA prevention in children¡ 5 years?" In: *Acta biomedica* 89.1 (2018), pp. 67–71.

[91] R Iwata, N Sasaki, and T Agui. "Contiguous gene deletion of Ptprk and Themis causes T-helper immunodeficiency (thid) in the LEC rat". In: *Biomedical Research* 31.1 (2010), pp. 83–87.

[92] MK Iyer et al. "The landscape of long noncoding RNAs in the human transcriptome". In: *Nature genetics* 47.3 (2015), pp. 199–208.

[93] AL Johnson et al. "Themis is a member of a new metazoan gene family and is required for the completion of thymocyte positive selection". In: *Nature immunology* 10.8 (2009), pp. 831–839.

[94] MKB Jonsson et al. "A transcriptomic and epigenomic comparison of fetal and adult human cardiac fibroblasts reveals novel key transcription factors in adult cardiac fibroblasts". In: *JACC: Basic to Translational Science* 1.7 (2016), pp. 590–602.

[95] ME Jørgensen, TP Almdal, and B Carstensen. "Time trends in mortality rates in type 1 diabetes from 2002 to 2011". In: *Diabetologia* 56.11 (2013), pp. 2401–2404.

[96] BD Juran et al. "Primary biliary cirrhosis is associated with a genetic variant in the 3' flanking region of the CTLA4 gene". In: *Gastroenterology* 135.4 (2008), pp. 1200–1206.

[97] N Kamitaki et al. "Complement component 4 genes contribute sex-specific vulnerability in diverse illnesses." In: *bioRxiv* (2019), p. 761718.

[98] HM Kang et al. "Efficient control of population structure in model organism association mapping". In: *Genetics* 178.3 (2008), pp. 1709–1723.

[99] HM Kang et al. "Variance component model to account for sample structure in genome-wide association studies". In: *Nature genetics* 42.4 (2010), pp. 348–354.

[100] GR Kazeem and M Farrall. "Integrating case-control and TDT studies". In: *Annals of human genetics* 69.3 (2005), pp. 329–335.

[101] MP Keller et al. "The transcription factor Nfatc2 regulates β-cell proliferation and genes associated with type 2 diabetes in mouse and human islets". In: *PLoS genetics* 12.12 (2016), e1006466.

[102] B Keymeulen et al. "Insulin needs after CD3-antibody therapy in new-onset type 1 diabetes". In: *New England Journal of Medicine* 352.25 (2005), pp. 2598–2608.

[103] G Kichaev and B Pasaniuc. "Leveraging functional-annotation data in trans-ethnic fine-mapping studies". In: *The American Journal of Human Genetics* 97.2 (2015), pp. 260–271.

[104] G Kichaev et al. "Integrating functional data to prioritize causal variants in statistical fine-mapping studies". In: *PLoS genetics* 10.10 (2014), e1004722.

[105] H Kolb et al. "Prospective analysis of islet cell antibodies in children with type 1 (insulin-dependent) diabetes". In: *Diabetologia* 31.4 (1988), pp. 189–194.

[106] WM Kuhtreiber et al. "Low levels of C-peptide have clinical significance for established Type 1 diabetes". In: *Diabetic Medicine* 32.10 (2015), pp. 1346–1353.

[107] D La Torre et al. "Decreased cord-blood phospholipids in young age–at–onset type 1 diabetes". In: *Diabetes* 62.11 (2013), pp. 3951–3956.

[108] SH Lee et al. "Estimating missing heritability for disease from genome-wide association studies". In: *The American Journal of Human Genetics* 88.3 (2011), pp. 294–305.

[109] P Leete et al. "Differential insulitic profiles determine the extent of β-cell destruction and the age at onset of type 1 diabetes". In: *Diabetes* 65.5 (2016), pp. 1362–1369.

[110] H Li. "Minimap2: pairwise alignment for nucleotide sequences". In: *Bioinformatics* 34.18 (2018), pp. 3094–3100.

[111] Q Li et al. "Measuring reproducibility of high-throughput experiments". In: *The annals of applied statistics* 5.3 (2011), pp. 1752–1779.

[112] YR Li et al. "Genetic sharing and heritability of paediatric age of onset autoimmune diseases". In: *Nature communications* 6 (2015), p. 8442.

[113] Y Li et al. "CD6 as a potential target for treating multiple sclerosis". In: *Proceedings of the National Academy of Sciences* 114.10 (2017), pp. 2687–2692.

[114] Y Liao, GK Smyth, and W Shi. "featureCounts: an efficient general purpose program for assigning sequence reads to genomic features". In: *Bioinformatics* 30.7 (2014), pp. 923–930.

[115] S Lindley et al. "Defective suppressor function in CD4+ CD25+ T-cells from patients with type 1 diabetes". In: *Diabetes* 54.1 (2005), pp. 92–99.

[116] PS Linsley et al. "B lymphocyte alterations accompany abatacept resistance in new-onset type 1 diabetes". In: *JCI insight* 4.4 (2019), p. 126136.

[117] JZ Liu et al. "Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations". In: *Nature genetics* 47.9 (2015), pp. 979–986.

[118] J Liu et al. "Dense fine-mapping study identifies new susceptibility loci for primary biliary cirrhosis". In: *Nature genetics* 44.10 (2012), pp. 1137–1141.

[119] PR Loh et al. "Reference-based phasing using the Haplotype Reference Consortium panel". In: *Nature genetics* 48.11 (2016), pp. 1443–1448.

[120] MI Love, W Huber, and S Anders. "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2". In: *Genome biology* 15.12 (2014), p. 550.

[121] J Ludvigsson et al. "Decline of C-peptide during the first year after diagnosis of Type 1 diabetes in children and adolescents". In: *Diabetes research and clinical practice* 100.2 (2013), pp. 203–209.

[122] M van Lummel et al. "Type 1 diabetes-associated HLA-DQ8 transdimer accommodates a unique peptide repertoire". In: *Journal of Biological Chemistry* 287.12 (2012), pp. 9514–9524.

[123] A Mahajan et al. "Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps". In: *Nature genetics* 50.11 (2018), pp. 1505–1513.

[124] JB Maller et al. "Bayesian refinement of association signals for 14 loci in 3 common diseases". In: *Nature genetics* 44.12 (2012), pp. 1294–1301.

[125] A Manichaikul et al. "Robust relationship inference in genome-wide association studies". In: *Bioinformatics* 26.22 (2010), pp. 2867–2873.

[126] T Mantere, S Kersten, and A Hoischen. "Long-read sequencing emerging in medical genetics". In: *Frontiers in genetics* 10 (2019), p. 426.

[127] J Marchini and B Howie. "Genotype imputation for genome-wide association studies". In: *Nature Reviews Genetics* 11.7 (2010), pp. 499–511.

[128] A Márquez et al. "Meta-analysis of Immunochip data of four autoimmune diseases reveals novel single-disease and cross-phenotype associations". In: *Genome medicine* 10.1 (2018), p. 97.

[129] R McGinnis, S Shifman, and A Darvasi. "Power and efficiency of the TDT and case-control design for association scans". In: *Behavior genetics* 32.2 (2002), pp. 135–144.

[130] PM McKeigue et al. "Persistent C-peptide secretion in Type 1 diabetes and its relationship to the genetic architecture of diabetes". In: *BMC medicine* 17.1 (2019), p. 165.

[131] ML Metzker. "Sequencing technologies - the next generation". In: *Nature reviews genetics* 11.1 (2010), pp. 31–46.

[132] AP Morris. "Transethnic meta-analysis of genomewide association studies". In: *Genetic epidemiology* 35.8 (2011), pp. 809–822.

[133] AP Morris et al. "A powerful approach to sub-phenotype analysis in population-based genetic association studies". In: *Genetic epidemiology* 34.4 (2010), pp. 335–343.

[134] S Myers et al. "A fine-scale map of recombination rates and hotspots across the human genome". In: *Science* 310.5746 (2005), pp. 321–324.

[135] S Nejentsev et al. "Localization of type 1 diabetes susceptibility to the MHC class I genes HLA-B and HLA-A". In: *Nature* 450.7171 (2007), pp. 887–892.

[136] MR Nelson et al. "The support of human genetic evidence for approved drug indications". In: *Nature genetics* 47.8 (2015), pp. 856–860.

[137] C Nguyen et al. "Definition of high-risk type 1 diabetes HLA-DR and HLA-DQ types using only three single nucleotide polymorphisms". In: *Diabetes* 62.6 (2013), pp. 2135–2140.

[138] J Ni et al. "CTLA-4 CT60 (rs3087243) polymorphism and autoimmune thyroid diseases susceptibility: a comprehensive meta-analysis". In: *Endocrine research* 39.4 (2014), pp. 180–188.

[139] KK Nicodemus, A Luna, and YY Shugart. "An evaluation of power and type I error of single-nucleotide polymorphism transmission/disequilibrium–based statistical methods under different family structures, missing parental data, and population stratification". In: *The American Journal of Human Genetics* 80.1 (2007), pp. 178–185.

[140] L Nisticò et al. "The CTLA-4 gene region of chromosome 2q33 is linked to, and associated with, type 1 diabetes". In: *Human molecular genetics* 5.7 (1996), pp. 1075–1080.

[141] JA Noble and AM Valdes. "Genetics of the HLA region in the prediction of type 1 diabetes". In: *Current diabetes reports* 11.6 (2011), pp. 533–542.

[142] TC Nogueira et al. "GLIS3, a susceptibility gene for type 1 and type 2 diabetes, modulates pancreatic beta cell apoptosis via regulation of a splice variant of the BH3-only protein Bim". In: *PLoS genetics* 9.5 (2013), e1003532.

[143] Y Okada et al. "Genetics of rheumatoid arthritis contributes to biology and drug discovery". In: *Nature* 506.7488 (2014), pp. 376–381.

[144] S Onengut-Gumuscu et al. "Fine mapping of type 1 diabetes susceptibility loci and evidence for colocalization of causal variants with lymphoid gene enhancers". In: *Nature genetics* 47.4 (2015), pp. 381–386.

[145] RA Panganiban et al. "A functional splice variant associated with decreased asthma risk abolishes the ability of gasdermin B to induce epithelial cell pyroptosis". In: *Journal of Allergy and Clinical Immunology* 142.5 (2018), pp. 1469–1478.

[146] A Paun et al. "Association of HLA-dependent islet autoimmunity with systemic antibody responses to intestinal commensal bacteria in children". In: *Science immunology* 4.32 (2019), eaau8125.

[147] A Pham-Short et al. "Coeliac disease in type 1 diabetes from 1990 to 2009: higher incidence in young children after longer diabetes duration". In: *Diabetic Medicine* 29.9 (2012), e286–e289.

[148] BE Phillips et al. "Tolerogenic Dendritic Cells and T-Regulatory Cells at the Clinical Trials Crossroad for the Treatment of Autoimmune Disease; Emphasis on Type 1 Diabetes Therapy". In: *Frontiers in immunology* 10 (2019), p. 148.

[149] AL Price et al. "Long-range LD can confound genome scans in admixed populations". In: *The American Journal of Human Genetics* 83.1 (2008), pp. 132–135.

[150] S Purcell et al. "PLINK: a tool set for whole-genome association and population-based linkage analyses". In: *The American Journal of Human Genetics* 81.3 (2007), pp. 559–575.

[151] R Qi, D Singh, and CC Kao. "Proteolytic processing regulates Toll-like receptor 3 stability and endosomal localization". In: *Journal of Biological Chemistry* 287.39 (2012), pp. 32617–32629.

[152] Q Qiao et al. "A genome-wide scan for type 1 diabetes susceptibility genes in nuclear families with multiple affected siblings in Finland". In: *BMC genetics* 8 (2007), p. 84.

[153] FJ Quintana et al. "Aiolos promotes TH17 differentiation by directly silencing Il2 expression". In: *Nature immunology* 13.8 (2012), pp. 770–777.

[154] MJ Rahman et al. "Restoration of the type I IFN–IL-1 balance through targeted blockade of PTGER4 inhibits autoimmunity in NOD mice". In: *JCI insight* 3.3 (2018), p. 97843.

[155] DB Rainbow et al. "A rare IL2RA haplotype identifies SNP rs61839660 as causal for autoimmunity". In: *BioRxiv* (2017), p. 108126.

[156] J Rasschaert et al. "Toll-like receptor 3 and STAT-1 contribute to double-stranded RNA+ interferon-γ-induced apoptosis in primary pancreatic β-cells". In: *Journal of Biological Chemistry* 280.40 (2005), pp. 33984–33991.

[157] A Rawshani et al. "Excess mortality and cardiovascular disease in young adults with type 1 diabetes in relation to age at onset: a nationwide, register-based cohort study". In: *The Lancet* 392.10146 (2018), pp. 477–486.

[158] MJ Redondo et al. "Concordance for islet autoimmunity among monozygotic twins". In: *New England Journal of Medicine* 359.26 (2008), pp. 2849–2850.

[159] MJ Redondo et al. "Heterogeneity of type I diabetes: analysis of monozygotic twins in Great Britain and the United States". In: *Diabetologia* 44.3 (2001), pp. 354–362.

[160] SS Rich et al. "Overview of the type I diabetes genetics consortium". In: *Genes and immunity* 10.Suppl 1 (2009), S1–S4.

[161] CL Roark et al. "Multiple HLA epitopes contribute to type 1 diabetes susceptibility". In: *Diabetes* 63.1 (2014), pp. 323–331.

[162] MD Robinson and GK Smyth. "Moderated statistical tests for assessing differences in tag abundance". In: *Bioinformatics* 23.21 (2007), pp. 2881–2887.

[163] D Roshandel et al. "Meta-genome-wide association studies identify a locus on chromosome 1 and multiple variants in the MHC region for serum C-peptide in type 1 diabetes". In: *Diabetologia* 61.5 (2018), pp. 1098–1111.

[164] WJ Sandborn et al. "Tofacitinib as induction and maintenance therapy for ulcerative colitis". In: *New England Journal of Medicine* 376.18 (2017), pp. 1723–1736.

[165] F Sanger, S Nicklen, and AR Coulson. "DNA sequencing with chain-terminating inhibitors". In: *Proceedings of the National Academy of Sciences* 74.12 (1977), pp. 5463–5467.

[166] S Sawcer et al. "Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis". In: *Nature* 476.7359 (2011), pp. 214–219.

[167] BJ Schmiedel et al. "17q21 asthma-risk variants switch CTCF binding and regulate IL-2 production by T cells". In: *Nature communications* 7 (2016), p. 13426.

[168] E Seelig et al. "The DILfrequency study is an adaptive trial to identify optimal IL-2 dosing in patients with type 1 diabetes". In: *JCI insight* 3.19 (2018), p. 99306.

[169] DP Shanley et al. "An evolutionary perspective on the mechanisms of immunosenescence". In: *Trends in immunology* 30.7 (2009), pp. 374–381.

[170] SS Shapiro and MB Wilk. "An analysis of variance test for normality". In: *Biometrika* 52.3/4 (1965), pp. 591–611.

[171] E Sharon et al. "Genetic variation in MHC proteins is associated with T cell receptor expression biases". In: *Nature genetics* 48.9 (2016), pp. 995–1002.

[172] SA Sharp et al. "Development and standardization of an improved type 1 diabetes genetic risk score for use in newborn screening and incident diagnosis". In: *Diabetes care* 42.2 (2019), pp. 200–207.

[173] MJ Sheehy et al. "A diabetes-susceptible HLA haplotype is best defined by a combination of HLA-DR and-DQ alleles." In: *The Journal of clinical investigation* 83.3 (1989), pp. 830–835.

[174] N Sherry et al. "Teplizumab for treatment of type 1 diabetes (Protege study): 1-year results from a randomised, placebo-controlled trial". In: *The Lancet* 378.9790 (2011), pp. 487–497.

[175] BM Shields et al. "C-peptide decline in type 1 diabetes has two phases: an initial exponential fall and a subsequent stable phase". In: *Diabetes care* 41.7 (2018), pp. 1486–1492.

[176] N Siva. "1000 Genomes project". In: *Nature biotechnology* 26.3 (2008), p. 256.

[177] A Skowera et al. "CTLs are targeted to kill β cells in patients with type 1 diabetes through recognition of a glucose-regulated preproinsulin epitope". In: *The Journal of clinical investigation* 118.10 (2008), pp. 3390–3402.

[178] D Smyth et al. "Replication of an association between the lymphoid tyrosine phosphatase locus (LYP/PTPN22) with type 1 diabetes, and evidence for its role as a general autoimmunity locus". In: *Diabetes* 53.11 (2004), pp. 3020–3023.

[179] H Solow, R Hidalgo, and DP Singal. "Juvenile-onset Diabetes: HLA-A,-B,-C, and-DR Alloantigens". In: *Diabetes* 28.1 (1979), pp. 1–4.

[180] D Speed et al. "Reevaluation of SNP heritability in complex human traits". In: *Nature genetics* 49.7 (2017), pp. 986–992.

[181] RS Spielman, RE McGinnis, and WJ Ewens. "Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM)." In: *American journal of human genetics* 52.3 (1993), p. 506.

[182] M Stephens and DJ Balding. "Bayesian statistical methods for genetic association studies". In: *Nature Reviews Genetics* 10.10 (2009), pp. 681–690.

[183] N Suciu-Foca et al. "Juvenile diabetes mellitus and the HLA system". In: *Transplantation proceedings* 11.2 (1979), pp. 1309–1313.

[184] M Svensson, JW Eriksson, and G Dahlquist. "Early glycemic control, age at onset, and development of microvascular complications in childhood-onset type 1 diabetes: a population-based study in northern Sweden". In: *Diabetes care* 27.4 (2004), pp. 955–962.

[185] GR Svishcheva et al. "Rapid variance components–based method for whole-genome association analysis". In: *Nature genetics* 44.10 (2012), pp. 1166–1170.

[186] D Taliun et al. "Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program". In: *BioRxiv* (2019), p. 563866.

[187] MA Taub et al. "Incorporating genotype uncertainties into the genotypic TDT for main effects and gene-environment interactions". In: *Genetic epidemiology* 36.3 (2012), pp. 225–234.

[188] Nicholas J Thomas et al. "Frequency and phenotype of type 1 diabetes in the first six decades of life: a cross-sectional, genetically stratified survival analysis from UK Biobank". In: *The lancet Diabetes & endocrinology* 6.2 (2018), pp. 122–129.

[189] JA Todd et al. "Regulatory T cell responses in participants with type 1 diabetes after a single dose of interleukin-2: a non-randomised, open label, adaptive dose-finding trial". In: *PLoS medicine* 13.10 (2016), e1002139.

[190] JA Todd et al. "Robust associations of four new chromosome regions from genome-wide analyses of type 1 diabetes". In: *Nature genetics* 39.7 (2007), pp. 857–864.

[191] B Torres et al. "Association of the CT60 marker of the CTLA4 gene with systemic lupus erythematosus". In: *Arthritis & Rheumatism* 50.7 (2004), pp. 2211–2215.

[192] Gosia Trynka et al. "Disentangling the effects of colocalizing genomic annotations to functionally prioritize non-coding variants within complex-trait loci". In: *The American Journal of Human Genetics* 97.1 (2015), pp. 139–152.

[193] G Trynka et al. "Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease". In: *Nature genetics* 43.12 (2011), pp. 1193–1201.

[194] P Vafiadis et al. "Insulin expression in human thymus is modulated by INS VNTR alleles at the IDDM2 locus". In: *Nature genetics* 15.3 (1997), pp. 289–292.

[195] AM Valdes et al. "Use of class I and class II HLA loci for predicting age at onset of type 1 diabetes in multiple populations". In: *Diabetologia* 55.9 (2012), pp. 2394–2401.

[196] A Varshney et al. "Genetic regulatory signatures underlying islet gene expression and type 2 diabetes". In: *Proceedings of the National Academy of Sciences* 114.9 (2017), pp. 2301–2306.

[197] A Vella et al. "Localization of a type 1 diabetes locus in the IL2RA/CD25 region by use of tag single-nucleotide polymorphisms". In: *The American Journal of Human Genetics* 76.5 (2005), pp. 773–779.

[198] E Vercammen, J Staal, and R Beyaert. "Sensing of viral infection and activation of innate immunity by toll-like receptor 3". In: *Clinical microbiology reviews* 21.1 (2008), pp. 13–25.

[199] U Võsa et al. "Unraveling the polygenic architecture of complex traits using blood eQTL meta-analysis". In: *bioRxiv* (2018), p. 447367.

[200] P Wainschtein et al. "Recovery of trait heritability from whole genome sequence data". In: *bioRxiv* (2019), p. 588020.

[201] J Wakefield. "Bayes factors for genome-wide association studies: comparison with P-values". In: *Genetic epidemiology* 33.1 (2009), pp. 79–86.

[202] C Wallace et al. "Dissection of a complex disease susceptibility region using a Bayesian stochastic search approach to fine mapping". In: *PLoS genetics* 11.6 (2015), e1005272.

[203] M Walter et al. "No effect of the altered peptide ligand NBI-6024 on β-cell residual function and insulin needs in new-onset type 1 diabetes". In: *Diabetes care* 32.11 (2009), pp. 2036–2040.

[204] JH Wang et al. "Aiolos regulates B cell activation and maturation to effector state". In: *Immunity* 9.4 (1998), pp. 543–553.

[205] J Weng et al. "Incidence of type 1 diabetes in China, 2010-13: population based study". In: *British medical journal* 360 (2018), j5295.

[206] JM Wenzlau et al. "A common nonsynonymous single nucleotide polymorphism in the SLC30A8 gene determines ZnT8 autoantibody specificity in type 1 diabetes". In: *Diabetes* 57.10 (2008), pp. 2693–2697.

[207] HJ Westra et al. "Fine-mapping and functional studies highlight potential causal variants for rheumatoid arthritis and type 1 diabetes". In: *Nature genetics* 50.10 (2018), pp. 1366–1374.

[208] CJ Willer, Y Li, and GR Abecasis. "METAL: fast and efficient meta-analysis of genomewide association scans". In: *Bioinformatics* 26.17 (2010), pp. 2190–2191.

[209] J Wollenhaupt et al. "Safety and efficacy of tofacitinib for up to 9.5 years in the treatment of rheumatoid arthritis: final results of a global, open-label, long-term extension study". In: *Arthritis research & therapy* 21.1 (2019), p. 89.

[210] NR Wray et al. "Common disease is more complex than implied by the core gene omnigenic model". In: *Cell* 173.7 (2018), pp. 1573–1580.

[211] J Yang et al. "Concepts, estimation and interpretation of SNP-based heritability". In: *Nature genetics* 49.9 (2017), pp. 1304–1310.

[212] J Yang et al. "Genome-wide complex trait analysis (GCTA): methods, data analyses, and interpretations". In: *Methods in molecular biology* 1019 (2013), pp. 215–236.

[213] L Yengo et al. “Meta-analysis of genome-wide association studies for height and body mass index in ̃700000 individuals of European ancestry”. In: *Human molecular genetics* 27.20 (2018), pp. 3641–3649.

[214] S Yokoyama et al. “Tofacitinib, a janus kinase inhibitor demonstrates efficacy in an IL-15 transgenic mouse model that recapitulates pathologic manifestations of celiac disease”. In: *Journal of clinical immunology* 33.3 (2013), pp. 586–594.

[215] TJ Ypma. “Historical development of the Newton–Raphson method”. In: *SIAM review* 37.4 (1995), pp. 531–551.

[216] X Zheng et al. “HIBAG HLA genotype imputation with attribute bagging”. In: *The pharmacogenomics journal* 14.2 (2014), pp. 192–200.

[217] X Zhou and M Stephens. “Genome-wide efficient mixed-model analysis for association studies”. In: *Nature genetics* 44.7 (2012), pp. 821–824.